IWIN2018

International Workshop on Informatics

Proceedings of
International Workshop on Informatics

September 9-12, 2018

Salzburg, Austria

Informatics Society

Sponsored by Informatics Society

IWIN2018

International Workshop on Informatics

Proceedings of
International Workshop on Informatics

September 9-12, 2018
Salzburg, Austria

Informatics Society

Sponsored by Informatics Society

Publication office:
Informatics Laboratory
3-41, Tsujimachi, Kitaku, Nagoya 462-0032, Japan
Publisher:
Tadanori Mizuno, President of Informatics Society
ISBN:
978-4-902523-43-0
Printed in Japan

Table of Contents

Session 1: Intelligent Transportation Systems

(Chair: Yoshia Saito) (9:10 - 10:30, Sep. 10)

Session 2: IoT Systems

(Chair: Ryozo Kiyohara) (10:40 - 12:00, Sep. 10)

Keynote Speech 1

(13:30 - 14:10, Sep. 10)

Session 3: Multimedia Systems

(Chair: Katsuhiko Kaji) (14:20 - 15:20, Sep. 10)

Session 4: Network and Security

(Chair: Kozo Okano) (15:30 - 16:50, Sep. 10)

Session 5: Data Models

(Chair: Yoshia Saito) (9:00 - 10:20, Sep. 11)

Session 6: Systems and Applications

(Chair: Ohki Tetsushi) (10:30 – 12:10, Sep. 11)

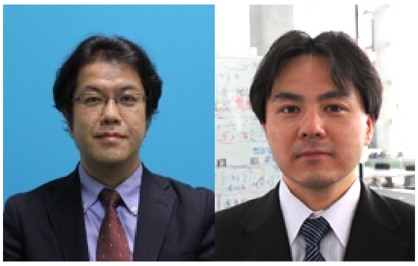
Keynote Speech 2

(13:30 - 14:10, Sep. 11)

Session 7: Trust and Risk

(Chair: Tomoki Yoshihisa) (14:20 - 15:20, Sep. 11)

Message from the General Chairs



It is our great pleasure to welcome all of you to Salzburg, Austria, for the Twelves International Workshop on Informatics (IWIN 2018). This workshop has been held annually by the Informatics Society. Since 2007, the workshops were held in Naples in Italy, Wien in Austria, Hawaii in the USA, Edinburgh in Scotland, Venice in Italy, Chamonix in France, Stockholm in Sweden, Prague in Czech Republic, Amsterdam in Netherlands, Riga in Latvia, and Zagreb in Croatia, respectively.

In IWIN 2018, 26 papers were accepted after peer reviewing by the program committee. Based on the papers, seven technical sessions were organized in a single track format, which highlighted the latest research results in the areas such as Intelligent Transport System (ITS), Internet of Things (IoT), Trust and Risks, Mobile Computing, Networking, Data Models and Multimedia Systems. IWIN 2017 will also welcome two keynote speakers: Mr. Kazuhiko Ohkubo of Vice President, Head of NTT Secure Platform Laboratories and Dr. Shigetoshi Sameshima of Deputy General Manager, Center for Technology Innovation and General Manager, Yokohama Research Laboratory, Hitachi, Ltd., Research & Development Group. We really appreciate their participation in the workshop.

We would like to thank all the participants and contributors who made the workshop possible. It is indeed an honor to work with a large group of professionals around the world for making the workshop a great success. We are looking forward to seeing you all in the workshop. We hope you enjoy IWIN 2018.

September 2018

Hiroshi Inamura
Yoh Shiraishi

Organizing Committee

General Co-Chairs

Hiroshi Inamura (Future University Hakodate, Japan)

Yoh Shiraishi (Future University Hakodate, Japan)

Steering Committee

Hitoshi Aida (The University of Tokyo, Japan)

Toru Hasegawa (Osaka University, Japan)

Teruo Higashino (Osaka University, Japan)

Tadanori Mizuno (Aichi Institute of Technology, Japan)

Jun Munemori (Wakayama University, Japan)

Yuko Murayama (Tsuda University, Japan)

Ken-ichi Okada (Keio University, Japan)

Norio Shiratori (Chuo University / Tohoku University, Japan)

Osamu Takahashi (Future University Hakodate, Japan)

Program Chair

Yoshia Saito (Iwate Prefectural University, Japan)

Financial Chair

Tomoya Kitani (Shizuoka University, Japan)

Publicity Chair

Yoshitaka Nakamura (Future University Hakodate, Japan)

Program Committee

Keiichi Abe (Kanagawa Institute of Technology, Japan)
Naoya Chujo (Aichi Institute of Technology, Japan)
Chiaki Doi (NTT DOCOMO, Inc., Japan)
Yu Enokibori (Nagoya University, Japan)
Hisao Fukuoka (Tokyo Denki University, Japan)
Yusuke Gotoh (Okayama University, Japan)
Takaaki Hishida (Aichi Institute of Technology, Japan)
Hiroshi Horikawa (Mitsubishi Electric Information Network Corporation, Japan)
Toru Hoshi (Tokyo University of Technology, Japan)
Yusuke Ichikawa (NTT Corporation, Japan)
Makoto Imamura (Tokai University, Japan)
Tomoo Inoue (University of Tsukuba, Japan)
Hiroshi Ishikawa (Tokyo Denki University, Japan)
Masahiko Ishino (Bunkyo University, Japan)
Katsuhiko Kaji (Aichi Institute of Technology, Japan)
Tomoko Kaneko (Information-technology Promotion Agency, Japan)
Masaji Katagiri (NTT DOCOMO, Inc., Japan)
Yoshinobu Kawabe (Aichi Institute of Technology, Japan)
Tomoya Kitani (Shizuoka University, Japan)
Ryozo Kiyohara (Kanagawa Institute of Technology, Japan)
Minoru Kobayashi (Meiji University, Japan)
Tsukasa Kudo (Shizuoka Institute of Science and Technology, Japan)
Hiroshi Mineno (Shizuoka University, Japan)
Shinichiro Mori (Chiba Institute of Technology, Japan)
Hiroaki Morino (Shibaura Institute of Technology, Japan)

Katsuhiro Naito (Aichi Institute of Technology, Japan)
Yoshitaka Nakamura (Future University-Hakodate, Japan)
Masakatsu Nishigaki (Shizuoka University, Japan)
Kozo Okano (Shinshu University, Japan)
Tetsushi Ohki (Shizuoka University, Japan)
Hideyuki Takahashi (Tohoku University, Japan)
Yusuke Takatori (Kanagawa Institute of Technology, Japan)
Yoshiaki Terashima (Soka University, Japan)
Hirosato Tsuji
(Mitsubishi Electric Information Systems Corporation, Japan)
Hiroshi Yoshiura
(The University of Electro-Communications, Japan)
Takaya Yuizono
(Japan Advanced Institute of Science and Technology)

Ken Ohta (NTT DOCOMO, Inc., Japan)
Masashi Saito (Mitsubishi Electric Corporation, Japan)
Fumiaki Sato (Toho University, Japan)
Masaaki Shirase (Future University Hakodate, Japan)

Hirozumi Yamaguchi (Osaka University, Japan)
Tomoyuki Yashiro (Chiba Institute of Technology, Japan)
Takuya Yoshihiro (Wakayama University, Japan)
Tomoki Yoshihisa (Osaka University, Japan)
Takashi Yoshino (Wakayama University, Japan)

Session 1:
Intelligent Transportation Systems

(Chair: Yoshia Saito)

Rightfulness Evaluation of Obtained Data From Traffic Simulation for Improving Traffic Flow

Kazuki Someya*, Masashi Saito** , and Ryzo Kiyohara*

*Kanagawa Institute of Technology, Japan
**Kanazawa, Institute of Technology, Japan
({s1521138@cco, kiyohara@ic}.kanagawa-it, msaito@neptune.kanazawa-it).ac.jp

Abstract – As traffic jams in urban areas are becoming worse, some cities are taking measures, such as increasing the number of roads through the effective utilization of underground roadways and detecting increasing traffic volume. Moreover, many cities can avoid traffic jams by encouraging the use of public transportation. However, in provincial cities, public transportation services are poor, and many people use cars because of the small number of trains, even if there are railroads. Therefore, traffic jams continue to be a problem. One method of grasping the flow of traffic is to use the data of a road traffic census to investigate the traffic volume and the data of mobile space statistics expressed in the mesh region based on the position information of users of mobile phones. A simulation was performed to obtain the same data at the same time, and the accuracy of the data was demonstrated.

Keywords: ITS, Traffic jam, Simulation, Open data .

1 INTRODUCTION

Recently, traffic jams have become increasingly serious. In a developed city, traffic jams can be avoided when commuting by using public transportation. Moreover, because of the increasing use of automated vehicles in the future, people have greater concern about traffic jams becoming even more severe.

In a city with financial strength, such as Tokyo, it is possible to respond to this problem by using underground roads. However, not only in Japan but also in many provincial cities, public transportation has not developed as much, and movement using automobiles is most common. Therefore, reduction of traffic jams in provincial cities is important. As a result of improving roads, for cases in which the frequency of use is small, no effect can be obtained at all. In provincial cities with little financial power, there is no room to use low-impact road maintenance, thus it is necessary to implement it reliably and effectively. Therefore, it is necessary to carry out a simulation beforehand to determine whether an effect can be obtained.

Especially in the provincial cities of Japan developed mainly around rivers, there are many cities having the characteristics of sea, mountains, and rivers. As a result, because roads over bridges and through mountain areas are limited, a bottleneck exists that causes traffic jams because of the small number of roads that can pass, even though the traffic volume is high.

Many cities have introduced park and ride (P&R) systems [1][2] and road pricing [3] as measures to alleviate traffic jams in these provincial cities. The P&R system has a parking lot near public transportation, and a system that makes public transport available for long distances. However, many people are not using P&R owing to low parking lot capacity and high usage fees [4].

Road pricing is a system that collects fees when traveling on a specific road, such as an expressway. As it is limited to large cities that already have many public transportation facilities, it will be ineffective in the provincial cities with limited alternative transportation methods.

The above considerations show that it is necessary to improve the accuracy of the simulation and improve the accuracy of the effect obtained when the road is improved. For that purpose, it is necessary to know the origin–destination traffic volume [5] of the driver as a precondition.

Using a personal trip survey [3] is the conventional method. However, the cost of implementing a personal slip investigation is large, and it lacks accuracy because it requires data to be provided that depend on a person's memory. Thus, the frequency of such surveys is not high, and is often limited to weekdays in a city every few years because of costs, so this method is not recommended.

However, in Japan, a road traffic census is conducted every five years and it discloses traffic surveys as open data. This is data obtained by manually counting the number of vehicles at an intersection or the like.

Moreover, mobile spatial statistics are available, which disclose information on terminals existing in base stations of mobile phones as open data.

In this study, we propose traffic flow simulation using the road traffic census and mobile space statistics data. As shown in Figure 1, the traffic flow movement is predicted from past data. As shown in Figure 2, simulation is performed so that the same data as the road traffic census and the mobile space statistics can be acquired, and the validity of the data is obtained for the area around Kanazawa station in the Ishikawa prefecture. Because the exact number of cars and people cannot be known from the open data, the data cannot be used as is. Therefore, the same acquisition method as for open data was implemented using simulation.

2 RELATED WORKS

Several methods of forecasting the traffic volume have been proposed. Particularly, there is the method of using a personal trip survey. However, problems of implementation cost and the accuracy of the data have been pointed out for personal trip surveys. However, by putting an application on a smartphone, one can extract it as human behavior information and use it as data for a personal trip survey [6]. However, it is related to the data acquisition method of personal trip surveys, and it cannot be used immediately because of the necessity of installing applications.

Also, a method of estimating the origin–destination traffic volume by means of transportation using retention population data has also been proposed and implemented [7].

A "demand bus" operates in response to a demand. It is like a taxi, and it is effective for rural areas with few bus stops, because it does not run on a fixed route. As a result, the fee is expensive compared with the usual bus, and the traveling time depends on other users. However, if the number of users increases, there is a problem that route calculation becomes difficult to design in a timely manner. Therefore, relaxing the calculation time required for path planning using a hierarchical cooperative transport system can eliminate the problem [8]. It also makes it possible to use this option in provincial cities with high traffic demand, such as regional core cities

There is also a system that combines the advantages of demand-type taxis and merged-type buses, the smart access vehicle. We address the inverse estimation problem of OD traffic volume by a transportation method using mobile residence population data from mobile zone statistics, the number of people getting on and off on specific bus route, and individual traveling locus data of a small sample from the viewpoint of easily obtained observation data. The calculated traffic volume between 200 zones of a 500-m mesh in Hakodate in the Hokkaido prefecture was used, and it shows the usefulness of the method by verifying its accuracy [5]. We verified the separation between migrants and residents, but the estimation accuracy is not high.

We estimated the number of users by mode of transportation, but, because it is not based on precise techniques or automobile traffic survey data, the actual traffic volume cannot be predicted and is a future topic for study.

NTT DOCOMO has been experimenting with the "near future people forecast" [9], predicting movement several hours ahead by using a real-time version of mobile space statistics and spatiotemporal variable online prediction technology. NTT DOCOMO is currently developing an artificial-intelligence taxi, which predicts the number of taxis that will be used in 30 min. Forecasting the number of people in the near future is a technique for predicting the number of people in an area several hours ahead.

We propose a vehicle direction estimation method using the Road Traffic Census and mobile space statistics for Kanazawa and Nonoichi. If we assume that an error of ±20% is tolerated, a regression equation at approximately a 63% mesh is derived. Because the determination coefficient R2 of the regression equation is 0.58 and the accuracy is low, improvement of the accuracy is a problem.

3 PROPOSED METHOD

The following problems exist when using either road traffic census or mobile space statistics data alone.

① The road traffic census cannot grasp the traffic volume of all roads, because it is implemented only on specific roads.

② Because mobile space statistics acquire the position information of the mobile terminal on a mesh unit basis, it is impossible to determine whether the object is a pedestrian or a car.

③ Additionally, because the data do not have IDs for recognizing an item, it is not possible to find the movement data of a specific car.

Figure 1. Traffic flow prediction

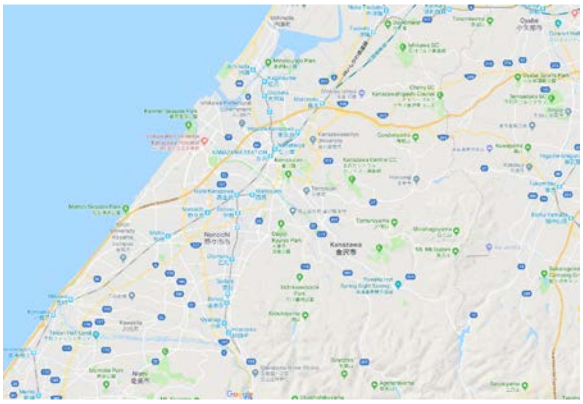


Figure 2. Around Kanazawa in Ishikawa prefecture（Source: Google）

Table 1. Example of Road Traffic Census (Partially Extracted)

id	date	up・down	type	Hourly traffic volume (count/hour)						24 hour traffic volume
				7hour	8hour	9hour		5hour	6hour	
17300080010	20151008	1	1	550	337	327	～	66	274	5968
17300080010	20151008	1	2	129	120	179		116	130	2274
17300080010	20151008	2	1	360	318	331		38	168	5777
17300080010	20151008	2	2	73	111	169		107	109	2376

Table 2. Example of mobile space statistics

	date	day	hour	area	residence	age	gender	population
Kanazawa	2015/11/19	木	7:00	543665645	17201	-1	-1	7896
Kanazawa	2015/11/19	木	8:00	543665645	17201	-1	-1	8012
Nonoichi	2016/04/06	水	18:00	543654865	17203	-1	-1	1779
Nonoichi	2016/04/06	水	19:00	543654865	17203	-1	-1	1542
Hakusan	2016/02/10	火	9:00	543614465	17210	-1	-1	823
Hakusan	016/02/10	火	10:00	543614465	17210	-1	-1	902

Thus, accuracy is low when trying to understand the movement of a car with only one source of open data. Therefore, in this research, we evaluate the validity of the simulation data, which is the preliminary stage of deriving the regression equation from the road traffic census and mobile space statistics data.

Instead of using actual data directly, we installed the observation spots/mesh ranges where the road traffic census and the mobile space statistics were executed in the simulator, acquired the data, and carried out the experiment.

In converting the information of the road traffic census based on the mesh used in the mobile space statistics, we borrowed data from the research group and used it as shown in Figure 3.

The ratio of automobiles and pedestrians was changed, and data were acquired. The ratios used were 10:0 and 5:5..

3.1 Road Traffic Census

The road traffic census [11] is conducted to understand the condition of roads and road traffic throughout Japan every five years by the Ministry of Land, Infrastructure and Transport Road Bureau. It investigates nationwide road conditions, traffic volume, travel speed, departure place/destination of automobile operation, purpose of operation, etc. The contents are 12 h of traffic between roads, traffic volume of 24 h, average traffic speed during congestion, crowding at the peak traffic times in the morning or evening, and vehicle-type classification. Details of the data are shown in Table 1.

The road traffic census data used in this study was for 2017, which is the most recent.

3.2 Mobile space statistics

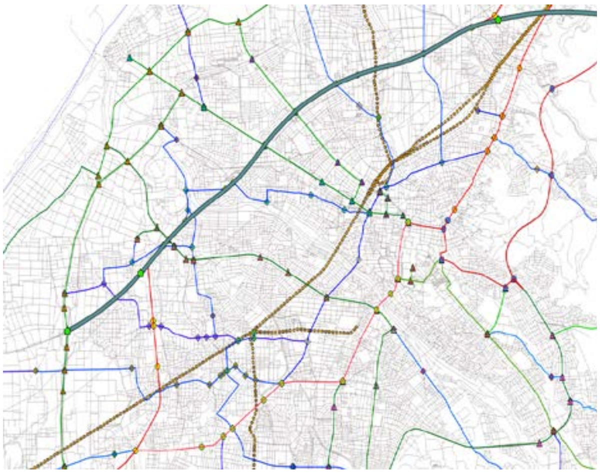


Figure 3. Road traffic census survey area around Kanazawa station

Table 3. Simulation settings

Human Agent	3,176
Vehicle Agent	3,176
Maximum number of passengers	1
Vehicle Speed	60km/h
Pedestrian Speed	3~5km/h
Experiment Time	2,160sec

Mobile space statistics [12] are open data provided by NTT DOCOMO. They periodically detect the mobile phones that exist in each area of each base station. The population during daytime and night can be roughly predicted, and, depending on the time, it can be known how many people commute to and drive in the central city. Details of the data are shown in Table 2

Because the information based on the mobile space statistics data includes date, time, area, residence, age, gender, and population, and because the information is based on cell phone contracts, a cell phone that a parent bought for a child in that case is nominally a parent’s phone, so the age is considered the parent's age, but although obtaining precise information is not possible, the data are sufficient as information for knowing overall trends.

The population is the number of mobile phones. The data of the area is cut in a mesh, and the area is confirmed by side 200 m - 2 km. Because the number of base stations is large in urban areas such as Tokyo, the distance of one side becomes short. Because age is based on information that can be contracted, such data are collected as the written age shown in Table 4, and they become rough data. The determination of low age group and high age group assumes that the contract does not necessarily belong to the person himself/herself. It is based on the premise that family registration is not necessarily done. With respect to the problem of personal information, processing is done without specifying an individual, and because the contract data of the corporate name is removed, it is safe.

4 EXPERIMENT

4.1 Environment of Simulation

An experiment was conducted using "Scenarige" [13], a multiagent traffic simulator of the Space–Time Engineering Company. The simulator was set as shown in Table 3. The map data published in OpenStreetMap were used [14].

4.2 Model

The human agent represents a person and treats the person as a pedestrian when not on a vehicle agent. For that reason, there are two types of means of transport: walking and cars. The maximum number of passengers represents the number of people who can ride in one car. In other words, in this research, only the driver is in the car. The number of human and vehicle agents is based on the hourly traffic volume table of the Road Traffic Census, based on the number 317,597.6, which is the average of the number of units of 7:00 to 12:00 a.m. that are needed to carry out the simulation if the number of moving vehicles is large. Because the time is enormous, it is rounded off to 1/100 and is set to 3176. Also, 34 patterns of human and vehicle agents are prepared so that they head to their destinations.

The experiment time was based on 6 h = 21,600 s from 7:00 to 12:00 a.m. At the start of the simulation, there were 0 cars, and they appeared exponentially. As a result, the number of vehicles increased with the passage of time, causing serious traffic jams and a situation in which mobility could not be obtained. When falling into a state where this mobility could not be acquired, it was a behavior that could not be realized, and the result of the simulation became unexpected. Thus, the experiment time was set to 21,600 s to 1/10, 2,160 s.

Examining the data of mobile space statistics, the mesh around Kanazawa Station is a 2-km mesh, so, to represent the whole area, as shown in Figure 4, a 5 × 5 mesh with 2 km per side was created.

As shown in Table 5, the data acquired were based on a judgement whether the moving object is a car or a pedestrian, on the name of the current road, and on position information on the x-axis and the y-axis. These data were acquired at intervals of 360 s. Because the name of the road was based on the data of the road traffic census, in the case of walking, it was blank. The x- and y-axes were used for conversion to a mesh. From the data of the road traffic census, shown in Table 5, only the row in which the moving object was a car and the road name were extracted. Originally, data on road traffic census did not describe the latitude and longitude of the observation point that replaces the x-axis and the y-axis. Therefore, in this

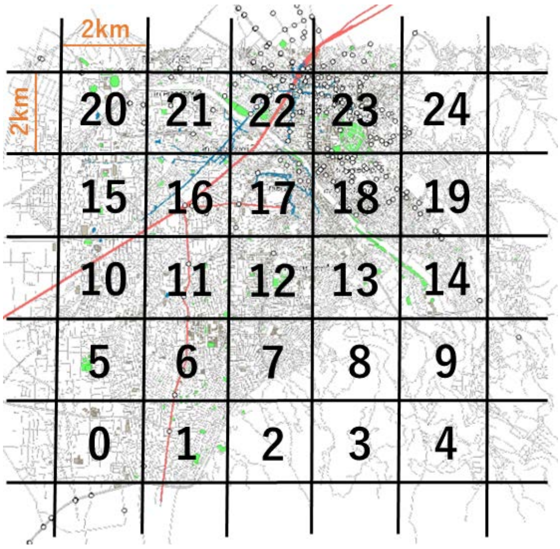


Figure 4. Allocation of mesh number

Table 4. Age classification

Actual age	15~19	20~29	30~39	40~49	50~59	60~69	70~79
Recorded age	15	20	30	40	50	60	70

Table 5. Example of acquired data

Time(sec)	How to move	Road name	X axis	Y axis
720	Car	Ishikawa Prefecture Route 146 Kanazawa Stop Station South Line	179.102	2815.36
720	Car	road777	41.13	1982.31
720	Walk		-324.44	3949.05

study, we obtained the position information of the car as the x-axis and y-axis.

Experimental results were for automobile-to-pedestrian ratios of 10:0 and 5:5, and the experiment was conducted 10 times. The average is shown as the result..

4.3 Result

We divided by mesh per hour and compared the road traffic census data obtained from simulation with the data of mobile space statistics. When a correlation was obtained with actual data. because it was convex, the correlation was derived by a quadratic polynomial as well.. For an automobile, when the ratio was 10:0, R2 = 0.9504; when the correlation coefficient R2 = 0.9688, the ratio was 5:5. It was confirmed that both have considerably high accuracy, indicating that there is no problem with the data. The respective graphs are shown in Figures 5 and 6. Moreover, if the road traffic census data and the data of the mobile space statistics were both 0 with the mesh at the same time, they were clearly unnecessary data for an accurate evaluation and were excluded.

The reason why 10: 0 is not R2 = 1 is considered to be that road traffic census does not acquire data on all roads, and excludes cars on narrow streets such as back streets and residential areas.

Th reason why 5: 5 is not very different from the correlation coefficient of 10: 0 is that the road traveled by car and the path walked on by foot are the same setting. It is also considered that it moved to an area not covered in the north.

5 CONCLUSION

In this research, as a preliminary step for finding countermeasures to the increasing traffic congestion in regional core cities, a simulation was performed using items that could acquire data of the road traffic census and mobile space statistics as a method for determining the flow of people The validity of the data was evaluated. Automobile-to-pedestrian ratios of 10:0 and 5:5 showed that the correlation coefficient exceeds 0.95 and that the data are reliable.

Because pedestrian movement is the same as that of the automobiles at the present stage, it was no problem to simulate at a ratio of 5: 5. However, in reality, because

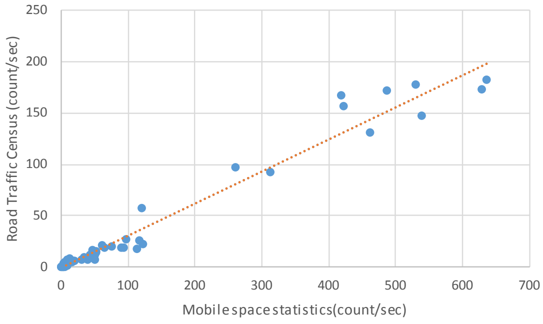


Figure 5. Correlation of ratio 10: 0

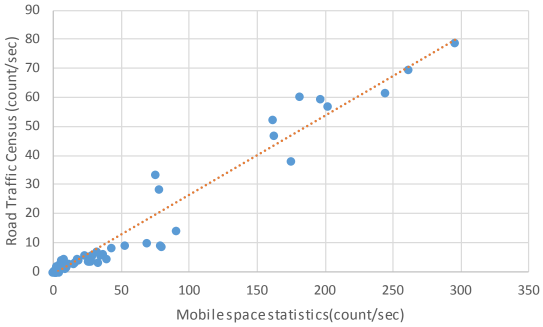


Figure 6. Correlation of ratio 5: 5

people who do not work outside the home do not move from the residential area and students go to school a more realistic evaluation environment is needed. Furthermore, various means of transport other than cars, such as buses.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant number 16K00433

REFERENCES

[1] K.Asai, “Dictionary for road and way,” NIPPON JITSUGYO PUBLISHING, 2001 (Japanese)

[2] S. Aoshima, S Suda, et al. “Analysis on Cost-Time Characteristics of Park & Ride Trip and Necessary Conditions of the Parking in Provincial Regions,” INFRASTRUCTURE PLANNING REVIEW, Vol. 16 (1999) P 863-868 (Japanese)

[3] The Ministry of Land, Infrastructure, Transport and Tourism, “Report of the Fifth Person Trip Survey ” 2009 (Japanese)

[4] K. Enomoto, M. Saito, R. Kiyohara,” Prediction of Traffic Flow in Provincial Cities by Open Data,” IPSJ 2016-MBL-80, 2016 (Japanese)

[5] R.Kanamori, K.Mizuno, I. Noda, H. Nakashima,” Estimation of Multi-Mode Origin-Destination Flows by using Staying Population Data at Base-Stations,” IPSJ 2015-ICS-178 (Japanese)

[6] H.Nakamura, K. Miyashita, E. Hato, T. Kishii,” The Estimation Method of Transportation mode for Person Trip Survey using Acceleration Sensor with Random Forest,” JSTE Journal of Traffic Engineering Vol. 1 (2015) No. 5 p. 10-18 (Japanese)

[7] H. Kobayashi, S. Hirata,” A Proposal of Evaluation Method for People with Mobility Difficulties using the Person Trip Survey Results,” Journal of the City Planning Institute of Japan, Vol. 48 (2013) No. 3 p. 159-164 (Japanese)

[8] K.Uehara, Y.Akamine, N.Toma, M.Nerome, S.Endo, “A Propose of Hierarchical Cooperative Transport System Using Demand Responsive Vehicles for Mid-size Metropolitan Area”, IPSJ Journal, Vol.57, No.1

[9] NTT DOCOMO "Near Future People Forecast" https://www.nttdocomo.co.jp/info/news_release/2017/09/20_00.html

[10] M.Saito, D.Hatano, R.Kiyohara : A Proposal of Vehicle Driving Volume Estimation Method using Spatial Statistics Data, IPSJ Journal, (Japanese)2018-ITS-073

[11] The Ministry of Land, Infrastructure, Transport and Tourism, “Census of Traffic Volume” http://www.mlit.go.jp/road/census/h27/

[12] T. Seike, H. Mimaki, S. Morita, ” RESEARCH ON THE EVALUATION OF REGIONAL PECULIARITIES IN KASHIWA AND YOKOHAMA BY “MOBILE SPATIAL STATISTICS”, ” AIJ Journal of Technology and Design,Vol. 21 (2015) No. 48 p. 821-826(Japanese)

[13] Scenargie, https://www.spacetime-eng.com/jp

[14] Open Street Map Japan, https://openstreetmap.jp

Study on State Transition Table Based on Context and Risk Degrees

Hirotaka Sakai* and Ryozo Kiyohara*

* Kanagawa Institute of Technology, Japan
{s1521077@cco, kiyohara@ic}.kanagawa-it.ac.jp

Abstract -There are many researches for autonomous vehicles in Intelligent Transport System (ITS) fields in the world. These technologies are defined from driver support system to driverless systems. However, the lifetime of vehicles are very long. Therefore, there will be mixed environment of autonomous and non-autonomous vehicles which have to require the driver support system. In this paper, we introduce the state transition table based on contexts which are driver status, vehicle status, other vehicles status, road status and etc. Then, we defined the risk degrees by the experiment using driving simulator.

Keywords: Context, vehicle, ITS, Risk degree, State Transition table.

1 INTRODUCTION

Recently, autonomous vehicle technologies are researched and developed widely. National Highway Traffic Safety Administration (NHSTA) defined the autonomous vehicle technology level as shown in Table 1 [1].

Current level of vehicles in the market is from level 0 to level 2. The autonomous vehicle of level 3 is demonstrated by many manufacturers. In the near future, the autonomous vehicle of level4 and level5 will be developed and shipped. However, human operated vehicle will be driven more than 20 years. Therefore, many vehicles will be driven in the mixed environment of the autonomous vehicles and human operated vehicles. It means the further preventive safety technologies for the driver are required.

Table 1: Definition of the autonomous vehicle technology level

Level	substance	operator
level 0 Human operated	Driver operates all tasks	human
Level 1 supported	System operates a sub-task of controlling a vehicle	human
Level 2 Partially automated	System operates some sub-tasks of controlling a vehicle	human
level 3 Conditionally automated	System operates all tasks for controlling a vehicle except the emergency	System and human
Level 4 Automated high level	System operates all tasks for controlling a vehicle when the conditions are satisfied	system
Level 5 Fully automated	System operates all tasks in anytime	system

There are many preventive safety technologies. Many of these technologies are event driven technologies. Most of these technologies are supported to avoid the accident or keep safety. However, these support system are sometimes too much or too early, because the timing to support and level to support is different for each drivers. Therefore, we focus on the contexts which the driver, the vehicle, vehicles around the vehicle, roads, weather and etc. Suitable information or action which support to driver and reduce the accidents for each driver depends on the contexts. There are many accidents which reasons are shown in Table 2.These reason is depend on the many contexts.

Moreover, distracted driving is the reason from other aspects. Therefore, we have to understand how dangerous of distracted driving, specially using smartphones.

In this paper, we defined these contexts and discuss about driver's contexts with risk degrees. Then we experiment how dangerous the distracted driving with driving simulator. Moreover, we discuss about the contexts from this result.

2 WHAT IS CONTEXT

We defined that the contexts are status of the driver, the vehicle who the driver operates, other vehicles around the vehicle, road surface, and weather. Moreover, the status is changed by some events (e.g. sudden braking of front vehicle, traffic signal, called by hands-free phone and etc.)

Knowledge of location is also context. Because, the driver's distraction is depend on the knowledge of locations or roads.

We think we can define the state transition table. The state is changed by many events. This table might be too large to translate for the program. However we can cover the all status for guarantee of safety in any situation.

3 RELATED STUDIES

There are many studies for estimating the human status. Sleepiness is one of the most important factors as a driver's context. In many case, the camera located on dashboard, on review mirror or on the information devices records the

T able 2: Number of accidents in Japan in 2013

Reason	Number of accident
Mistaking of steering	5,940
Mistaking of pedals	6,402
Mistaking of braking	20,117
Others	9,346

status of eye [4]. Therefore, we can estimate easily to sleepiness. Moreover, the drivers fatigue can be estimated by Bayesian Network technologies [5].

The degree of driver's vigilance is estimated by camera [6]. These technologies can estimate the driver's status by camera which has to be installed on the vehicle. There are other studies for estimating the drive's status by smartphones [7]. We think we can estimate the important status of driver by these technologies.

The knowledge of location is constructed in the cognitive maps. This cognitive map depends on many landmarks [8]. This status can be estimated from logs in navigation devices.

There are many studies for vehicle to vehicle or roadside equipment [9] which can get many kinds of information around the vehicle (e.g. other vehicles velocity, status of road surfaces, traffic accident information and etc.)

There are no studies which gather many kinds of information and feedback to the driver to the dangerous information for safety. Therefore, in this paper, we defined status transition table which can be decided from many these technologies.

4 PROPOSED METHOD

4.1 State Transition of Context

We have studied by gathering many kinds of information from many drivers by interviewing what events give them the impact of their driving. We have classified the answers to three types which are driver, the vehicle state and prediction. Then, we defined five context which we classified more detail of three types as follows:

- Vehicles state
  Velocity, braking and etc. which can be gotten from vehicle network (on board diagnostic port)
- Driver's state
  Estimating from the data which can be gotten from sensors or camera on the vehicle.
- Driver's behavior
  Getting from the operational logs from camera or on-vehicles devices
- State around the vehicle
  Getting from vehicle to vehicle (V2V) or vehicle to roadside equipment (V2R) communication
- State of destination or on the route getting from V2V or V2R communication with multi-hopping technologies

Driver's sate can be refined to two types. One is location context where driver know well or not. Location context is related to look the road carefully or not. Another is driver's mentalities which are fatigue, sleepiness, and etc.

Table 3 shows a part of state transition table. First, we collect many contexts and categorize them. Left side of table is wide categories. Three columns of right side are example of events. In this table, these events are caused by human behavior. The events defined in this table are only caused during driving.

There are many other types of events which are watching navigation, operating smartphone, front vehicle's braking, jumping out in front of the vehicle, car horn and etc.

There are too many states and too many events. Therefore, we classified and categorized to degree the state transition.

4.2 What is risk degree

We defined the risk degree corresponded for each state in Table 3. There are three kinds of risks which the driver has, other vehicles have and the driver gives other vehicles. We defined the degree five levels temporarily. After experiment or evaluation, we have to redefine these levels. Because, there are no bases of risk levels.

5 EXPERIMENT

We experimented to evaluate for definition of risk levels by driving simulator.

5.1 Environment of experiment

The driving simulator which used is simple type of simulator as shown in Figure 1. The events are raised by scenario which can be edited to many types of events. We prepared two types of scenario. One is sudden braking scenario which cause the rear-end accidents. Another scenario is the entanglement accident for bikes. Each operation and events are logged by 10ms intervals.

5.2 Scenarios

We prepared three scenarios for the two types of scenario. Moreover, we prepared the scenario for preliminary experiment. These are as follows:

(1) Preliminary scenario
After 4 minutes driving, go to the mountain road and bikes move suddenly

(2) Sudden braking scenario

- Route1:After driving 300 m, front vehicles suddenly brakes.
- Route2: front vehicles suddenly brakes on the straight road
- Route3: after waiting the traffic signal and start,front vehicles suddenly brakes.

Figure:1 driving simulator

Table 3: Example of state transition table

events status					Press the Acceleration pedal	Release acceleration pedal	Brinker for right turn
Status of vehicle	Status in motion	Going ahead	High speed	Go straight		Decreasing speed, go straight	Decreasing speed, turn right, brinker
				Turn right		Decreasing speed, turn right	Decreasing speed, turn right, brinker
				Turn left		Decreasing speed, turn left	Decreasing speed, turn right, brinker
			normal speed	Go straight	High speed,go straight	Decreasing speed, go straight	Decreasing speed, turn right, brinker
				Turn right	High speed,turn right	Decreasing speed, turn right	Decreasing speed, turn right, brinker
				Turn left	High speed,turn left	Decreasing speed, turn left	Decreasing speed, turn right, brinker
			Low speed	Go straight	Normal speed,go straight	Decreasing speed, go straight	Decreasing speed, turn right, brinker
				Turn right	Normal speed, turn right	Decreasing speed, turn right	Decreasing speed, turn right, brinker
				Turn left	Normal speed, turn left	Decreasing speed, turn left	Decreasing speed, turn right, brinker
			Slow speed	Go straight	Slow speed,go straight	Slow speed, go straight	Slow speed, turn right, brinker
				Turn right	Slow speed, turn right	Slow speed, turn right	Slow speed, turn right, brinker
				Turn left	Slow speed, turn left	Slow speed, turn left	Decreasing speed, turn right, brinker
			Decreasing speed	Go straight	Normal speed, go straight	Decreasing speed, go straight	Decreasing speed, turn right, brinker

5.3 Experiments

We experimented for Six men and two women. 4 men drive frequently and one of them has a driver’s license but no experience. After training of operation for driving simulator, we experimented. We gathered various information as follows:

- Simulation time
- Velocity of the vehicle
- Steering angle
- Acceleration
- Braking
- Location
- Distance of stop after braking

6 EVALUATION AND DISCUSSION

6.1 How to evaluate

We evaluate by comparing two experiments. one is the data of driving normally. Another is the data of driving with operating the smartphone. We focus the data of the distance of stop after braking. If the decision of the braking is delayed, there is sudden braking. The distance of stop after sudden braking is shorter than normal braking.

Figure 2 shows the response time of three routes in sudden braking scenario. Table 3 shows the distance from the front vehicle and the distance for stopping after the sudden braking. These data is average data of all subjects. Figure 3 shows the deference of the driver who is beginner or not.

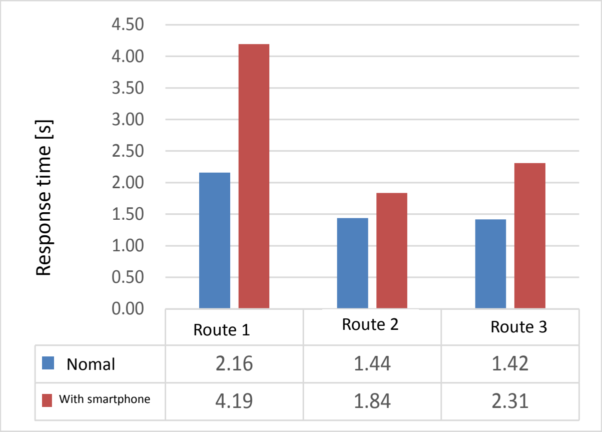


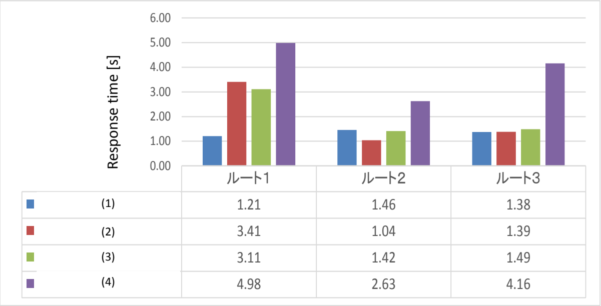
Figure 2. Response time of sudden braking

Table 3 distance of stopping and distance from front vehcile

	Route1	
	Normal	With smartphone
Distance for stopping [m]	42.2	61.23
Distance from front vehicle	15.71	35.92

	Route2	
	Normal	With smartphone
Distance for stopping [m]	28.05	33.65
Distance from front vehicle	1.89	7.33

	Route3	
	Normal	With smartphone
Distance for stopping [m]	29.05	43.62
Distance from front of vehicle	9.55	22.87



(1) Beginner drives normally
(2) Beginner drives with operating the smartphone
(3) Driver who has experience drives normally
(4) Driver who has experience drives with operating the sma

6.2 Discussion

We should evaluate the dangerous degree. Therefore, we have to define the dangerous degree. Dangerous degree should be the percentage of accidents in the situation. In many cases, the percentage is very small. If the all vehicle keep the rule of traffic, there are a few accidents. Moreover, accidents are raised by which both vehicles do not keep the rule.

Then we define the dangerous degree with the percentage of the accidents of the area and driver's status which we try to define by this experiment.

7 CONCLUSION

We introduce the driver's contexts and we experiment the dangerous degree is depend on the contexts. However, we should experiment many cases and we get the true degree of dangerous. So, we are planning to evaluate it by sensors which measured the blood flow in the head of each part.

REFERENCES

[1] NHTSA, https://www.nhtsa.gov/technology-innovation /automated-vehicles-safety

[2] Institute for Traffic Accident Research and Data Analisys (ITARDA), http://www.itarda.or.jp/english/

[3] Yuhei Ikeda, Ryota Horie, Midori Sugaya, "Estimating Emotion with Biological Information for Robot Interaction," Procedia Computer Science, Procedia Computer Science (2017)

[4] Han, W., Yang, Y., Huang, G.-B., Sourina, O., Klanner, F. and Denk, C. .Driver Drowsiness Detection Based on Novel Eye Openness Recognition Method and Unsupervised Feature Learning, Systems, Man, and Cybernetics(SMC), 2015 IEEE International Conference on, pp.1470-1475 (2015)

[5] Yang, G., Lin, Y. and Bhattacharya, P. . A Driver Fatigue Recognition Model Based on Information Fusion and Dynamic Bayesian Network, Special Issue on Intelligent Distributed Information Systems, Vol. 180, p.19421954 (2010).

[6] Ji, Q. and Yang, X. . Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance, Real-Time Imagin, Vol.8, pp. 357-377 (2002).

[7] Lee, B. –G. and Chung, W.-Y. . A Smartphone-Based Driver Safety Monitoring System Using Data Fusion, Sensors, Vol, 12, pp. 17536-17552 (2012).

[8] Seiji Matsuyama, Yuichi Tokunaga and Ryozo Kiyohara, "Recognizing User Location Context for Car Navigation Devices," IEEE International Symposium on Consumer Electronics (ISCE) (2016)

[9] Hiroto Furukawa, Ryozo Kiyohara, Yuich Tokunaga, Masashi Saito, "Vehicle Control Method at T-Junctions for Mixed Environments Containing Autonomous and Non-Autonomous, " IEEE International Conference on Advanced Information and Network Application (AINA) (2017)

A method to synchronize movie and acceleration sensor data using directions of vectors determined by corresponding points between video frames

Yosuke Ishiwatari[*][**], Takahiro Otsuka[*], Masahiro Abukawa[*], and Hiroshi Mineno[**]

[*] Information Technology R&D Center, Mitsubishi Electric Corporation, Japan
[**] Graduate School of Science and Technology, Shizuoka University, Japan
{Ishiwatari.Yosuke@dr, Otsuka.Takahiro@dw, Abukawa.Masahiro@bx}.MitsubishiElectric.co.jp,
mineno@inf.shizuoka.ac.jp

Abstract - Data acquisition and analysis systems using plural sensors are gaining popularity owing to the diversity, ongoing miniaturization, and inexpensiveness of sensors. These systems use two or more sensors to acquire sensor data. It is therefore important to synchronize different sensor data to analyze. In this paper, we propose a method for synchronizing video data and acceleration data from a moving car. We evaluate the performance of our method by using video data and acceleration data acquired using a smartphone and by extracting the intervals when a car turns right or left as synchronization points. The error found is 1.12 frames using this approach. We intend to expand and further optimize our methodology by extracting data from different scenarios.

Keywords: multimodal, data synchronization, motion estimation of a vehicle

1 INTRODUCTION

Data acquisition and analysis systems using plural sensors is gaining popularity owing to the diversity, ongoing miniaturization, and inexpensiveness of sensors. Autonomous driving is one of the applications using this approach. In an autonomous driving car, different sensors are used [1][2]. Even within common household vehicles, some utilize dashboard cameras with a GPS receiver. Vehicles that do not possess any sensors are therefore rare. However, sensors used are not commonly connected into one system. In this situation, multiple systems are therefore used for acquiring data.

Synchronization between the acquisition times of different sensor data is very important for analyzing the relationship between the data to find correlation values. If sensors are not set in one systems, a synchronization method is needed for analyzing the relationship between the data. In general, a car does not have such a synchronization method for a user, so an extra system must be added.

A typical correction method is to synchronize system times, but as each system clock is different, system times become incorrect after multiple synchronizations over a long period of time. Furthermore, an extra cost for using external time for synchronization is added, for example, a cost for GPS receivers for using GPS time.

In this paper, we propose a method for synchronizing sensor data by extracting the data ranges of different car motions through analysis of the characteristics of sensor data while not recording time. In [3], a method used a correlation value between sensor data without consideration of time, as described below; however, our target is not suitable for employing that method.

2 DETERMINING SYNCHRONIZAION POINTS USING IMAGE FEATURES AND CHARACTERISTICS OF ACCELERATION DATA

2.1 Our Target

In this paper, we aim to synchronize camera video and acceleration data, both which are recorded by sensors on a car. Each sensor is connected to different systems. We will then perform synchronization after all data are acquired (Fig. 1). Each system’s time is approximately similar, and the difference in time between the systems is not known.

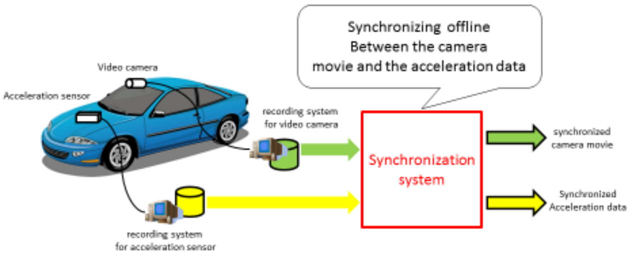


Figure 1: Our target

In this situation, the difference in time when sensor data were acquired cannot be obtained. In addition, each system clock is different; thus the differences in time between camera and acceleration data are not always constant. Therefore, if the synchronization is performed at one data point, it does not necessarily mean that all data points can be synchronized in a similar way.

Camera data and acceleration data have different characteristics, therefore same reference points related to both characteristics of the two data types are important for synchronization. Therefore, we propose a synchronization method via the detection of car motion behavior using sensor data and by matching the behavior ranges of these data.

2.2 Target Car Motion Events

As described previously, our method uses plural synchronization points. Thus, car motion events that determine the synchronization points must be easy to detect and appear frequently while driving. In this paper, we wish detect events wherein a car is turning right or turning left. We will then determine the points at the beginning or end of the events as synchronization points.

2.3 Detecting "Turning Right" and "Turning Left" Events Using Image Features

We used the optical flows of image features for detecting the behaviors of the car as "turning right" or "turning left" ("turning right/left") from camera images because the optical flows of image features that are on stationary objects show vectors that are opposite to the vector of a moving car. We can calculate the tendencies of the vectors from the optical flows of image features using whole frames. These image features are not solely on stationary objects; however, there are not many objects that move around the car, so the tendencies of the vectors that can be regarded as a vector are the same as those of a vector that shows the movement of the car. When a car is turning right/left, the optical flows from the image features on stationary objects are opposite to the direction in which the car is moving. If a camera is recording in front of a car and when a car is turning right, optical flows turn left. However, when a car is turning left, optical flows turn right. Meanwhile, the vertical direction of the optical flow varies with the position of each image feature within the camera image. As described in Fig. 2 as an example of "turning left," each vertical direction of the optical flow varies from each other.

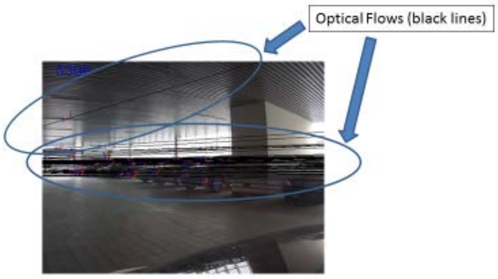


Figure 2: Optical flows of a "turning left" event

2.4 Detecting "Turning Right" or "turning Left" Using Acceleration Data

In this paper, we hypothesize that we can acquire acceleration data in the crosswise direction. When a car turns right/left, a driver operates the steering wheel to move in the crosswise direction. Thus, this operation is equal to accelerating the car in the crosswise direction, so the start of this operation causes a significant change in acceleration. This means that this operation can be detected using the difference in acceleration. We then used the difference in acceleration to detect the start or the end of the operation by detecting a peak or an inflection of the acceleration.

2.5 Detecting synchronization points

Following the ideas for detecting the "turning right/left" events from camera images or acceleration data, we propose a method to synchronize camera images and acceleration data.

2.5.1 Overview

Our method comprises two functions. The first is a function that detects ranges of the frame that indicates the car is turning right/left. The other is a function that calculates the difference between the ranges and the range of acceleration data by searching points that match the desired data. To reduce the search range, as described in section 2.1, the times of the systems are approximately similar and the difference in time is not known; however, the start of the searching point can be determined (the point that matches the point of the other sensor if the difference is zero).

2.5.2 Detecting the range of frames

As described in section 2.3, in camera videos, a "turning right/left" event has characteristics such that the optical flows tend to turn left/right. As such, we use these characteristics for detecting the ranges of the frames. We detect the ranges based on the functions as follows (Fig. 3):

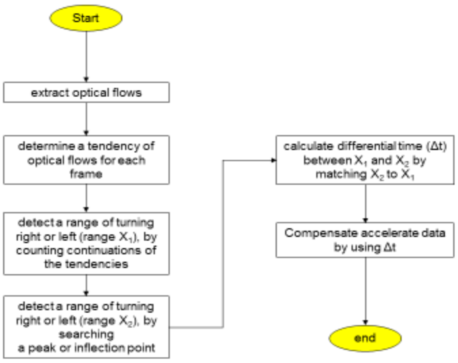

Start
extract optical flows
determine a tendency of optical flows for each frame
detect a range of turning right or left (range X_1), by counting continuations of the tendencies
detect a range of turning right or left (range X_2), by searching a peak or inflection point
calculate differential time (Δt) between X_1 and X_2 by matching X_2 to X_1
Compensate accelerate data by using Δt
end

Figure 3: Flow of matching time between acceleration data and movie data

1) Obtain the optical flows of the image features between successive frames
2) Calculate the tendency of the vectors of the optical flows by classifing the vectors into 16 bins based on the direction of the vector, and select the bin that includes the majority of the vectors
3) Calculate the range by counting the frames in which the bin of the start frame is the left binor the right bin. If the bin is near (within 2 bins) the former frame in a continuous fashion

2.5.3 Correcting separated situations

In [4], we discussed how an incorrect detection of a turning right/left event that was divided for more than one situation. This was caused by another action occurring simultaneously when a car is turning right/left, e.g., riding on a curb and heavy braking. If the number of frames between two of the "turning right/left" events is very small (a few frames) and if these situations are the same (i.e., these situations are "turning right" and "turning right" and vice versa), these situations should be regarded as one situation (Fig. 4).

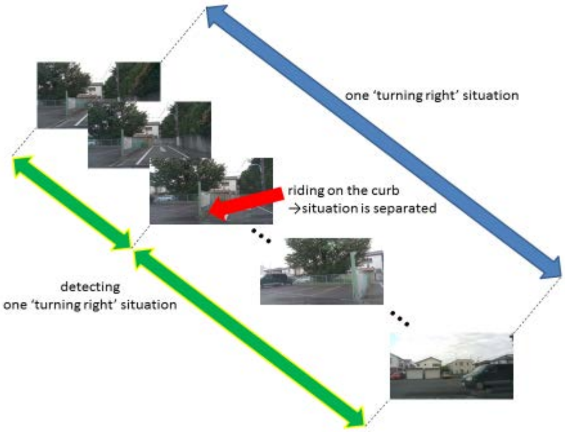


Figure 4: An example of separating one situation "turning right" into two situations

2.5.4 Matching the range of frames into ranges of the acceleration data

After calculating the range of the frames, as described above, the matching points between the start/end frame of the desired range and the acceleration data are calculated. As described in section 2.4, a peak or an inflection point of the acceleration data is searched. A peak or an inflection point of the acceleration data closest to the start point is regarded as the corresponding point of the start/end frame.

After calculating the corresponding point of the start frame (point C_1) in the acceleration data (point A_1) and the corresponding point of the end frame (point C_2) in the same data (point A_2), A_1 and A_2 are corrected to have the same range of time between the range from C_1 to C_2 and the range from A_1 to A_2. In detail, point A_1 is moving to A_1' and point A_2 is moving to A_2'. Therefore, "(time of A_2') − (time of A_1') = (time of C_2) – (time of C_1)" and "(time of A_1) − (time of A_1') = −[(time of A_2) − (time of A_2')]." This means that A_1' and A_2' are calculated as follows:

$$\Delta \text{tc} = (\text{time of } C_2) - (\text{time of } C_1)$$
$$\Delta \text{ta} = (\text{time of } A_2) - (\text{time of } A_1)$$
$$\text{diff} = (\Delta \text{tc} - \Delta \text{ta})/2$$
$$\text{time of } A_1' = (\text{time of } A_1) - \text{diff}$$
$$\text{time of } A_2' = (\text{time of } A_2) + \text{diff}$$

Hence, (time of C_1) − (time of A_1'), i.e., D_1, means the difference in time between the camera video and the acceleration data at the start frame of the range and (time of C_2) − (time of A_2'), i.e., (D_2), means the difference in time between the camera video and the acceleration data at the end frame of the range. D_1 and D_2 are not always identical. Therefore, the difference value ΔE at time E (E is the time between A_1' and A_2') is calculated as follows:

$$\Delta \mathrm{E} = \frac{D_1 * \left((time\ of\ A_2') - (time\ of\ E)\right) + D_2 * ((time\ of\ E) - (time\ of\ A_1'))}{(time\ of\ A_2') - (time\ of\ A_1')}.$$

2.5.5 Synchronizing the range other than "turning right/left"

For synchronizing the point (point X) in the range other than a "turning right/left" event, the difference (Δd1) between the point X and point X1, that point is the end frame of the range of the "turning right/left" that occurs immediately before the point X, and the difference (Δd2) between the point X and point X2, that point is the start frame of the range of the "turning right/left" event that occurs immediately after than the correcting point X, are used. The time difference ΔX at point X is calculated as follows:

$$\Delta \mathrm{X} = \frac{\Delta d_2 * \left((time\ of\ X) - (time\ of\ X_1)\right) + \Delta d_1 * ((timeof\ X_2) - (time\ of\ X))}{(time\ of\ X_2) - (time\ of\ X_1)}.$$

3 FUNDAMENTAL EVALUATION OF OUR METHOD

We evaluated the fundamental accuracy of our method. The data and our evaluation method is described as follows.

3.1 Data for Evaluation

We used the camera video and the acceleration data acquired by a smartphone in a car. The camera recorded the front view of the car. The acceleration sensor recorded the accelerations in three dimensions: the direction in which the car moves, the crosswise direction of the car, and the vertical direction of the car.

For evaluation, the recorded time was also acquired for these data (camera frames and acceleration data). The difference after employing our method for these data is equal to the accuracy of our method.

3.2 Data Preprocessing

To reduce noise in the acceleration data, a smoothing process was employed. In addition, an interpolation process was used to adjust the sampling rate of the acceleration data to the interval between the frames of the camera video (30 fps).

3.3 Range of Searching the Matching Point

As described in section 2.5.4, a peak or inflection point in the acceleration point is searched from the start. During evaluation, the range of search was defined for 30 samples (equal to 1 s).

3.4 Evaluation Target Range

Using the method that is described in sections 2.5.2 and 2.5.3, we obtained the ranges from the camera image. The ranges included the "turning right/left" events, but some events that were not regarded as the target event were included. The desired target events were also not included. For example, the event was a car avoiding an electric pole. In this situation, the car was moving in the right/left direction opposite to the pole after passing through the pole.

Events that were not the desired targets were manually removed, and 20 events were used for evaluation.

3.5 Evaluation Results

We evaluated the 20 situations, as described above, by calculating the difference values between the start time camera frames and the synchronized time of the acceleration data during “turning right/left” events (after employing our method).

The evaluation results are summarized in Table 1. In Table 1, a unit of error value is “frame” (1/30 s). The average of error value is 1.12 frames, and the standard deviation is 0.81 frames.

Table 1 Evaluation results

No.	Difference of frames	Absolute value of difference
1	1	1
2	−1	1
3	1	1
4	−1	1
5	0	0
6	−0.5	0.5
7	0.5	0.5
8	−0.5	0.5
9	−0.5	0.5
10	2	2
11	1	1
12	−1.5	1.5
13	2	2
14	1	1
15	−2	2
16	−0.5	0.5
17	−0.5	0.5
18	0	0
19	−1	1
20	0	0

3.6 Discussion

In the 20 situations described above, the average of the error frames is short, so it can be concluded that our methodology accounts for the data adequately. However, we have to consider some issues:

1) Handling the difference between Δta and Δtc
In our method, the difference between Δta and Δtc is divided equally and employed for A_1 and A_2.This is not always true as the difference between A_1' and A_1 is not always the same as that between A_2' and A_2. We can potentially resolve this issue by matching the correlation value.
2) Problems associated with insufficient number of image features
Some frames do not possess multiple image features. For example, the acquisition time is at night, and therefore, the frame has dark features. A frame is also occupied by the sky or the ground with no lines, signs, and other objects.
3) Tradeoff between accuracy of our method and frame rate
We calculated optical flows between successive frames (interval time between those frames is 0.33sec). The interval time is one of key parameters that determine accuracy of our method, but we think the interval time and accuracy is not necessarily in proportion, because if the interval time is large, optical flows are mainly determined from the behavior of car, so some noise, for example, very small motion of car or other objects, are less influenced for optical flows. So, it is important to determine appropriate interval time.
4) Expanding our method for other situations
Our method uses all the optical flows in the frame to calculate the tendency of their direction. The “turning right/left” event is an appropriate situation to be detected using the method described above. However, our method is not suitable for some situations, e.g., “moving straight.” In this situation, all the optical flows do not tend to turn in the same direction. The direction is determined according to the point of the image feature within the frame.
To expand our methodology, we can split a frame into subframes and calculate the tendencies within the subframes, followed by detection of the range based on the characteristics of the tendencies.

4 CONCLUSION

In this paper, we propose a method to synchronize a camera video and acceleration data onboard a moving car that are acquired from different systems. In our method, we calculated the synchronization points by determining a “turning right/left” event from the camera image and the acceleration data. From the camera image, we use the tendency of optical flows of the camera frame to detect the range of the event by detecting the specific tendency continuously. From the acceleration data, we detected the situation by identifying the peak or inflection point of the acceleration.

We evaluated the fundamental performance of our method using the camera image and the acceleration data acquired from a smartphone in a car, and the average of error frame was 1.12 frames. However, we have some issues to be resolved.

REFERENCES

[1] “Autonomous long-distance drive”, Mercedes-Benz, https://www.mercedes-benz.com/en/mercedes-benz/innovation/autonomous-long-distance-drive/
[2] Google Self-Driving Car Project, https://www.google.com/selfdrivingcar/
[3] S. Tanaka, *et al.*, “Study of Time Synchronization Method between Multiple Wearable Sensors”, the 13th of the *SOFT Kyushu Chapter Annual Conference*. (2011).
[4] Y. Ishiwatari, *et al.*, “A Synchronization Method Between Movie and Sensor Data Using Directions of Vectors Determined by Corresponding Points Between Video Frames”, IPSJ SIG Technical Reports, 2018-ITS-73. (2018).

A Method for Estimating Degradation Level of Road Markings by Participatory Sensing with In-Vehicle Camera

Yoh Shiraishi* and Issei Suga*

*Graduate School of Systems Information Science, Future University Hakodate, Japan
siraisi@fun.ac.jp

Abstract - The total road length in Japan is increasing year by year. On the contrary, the budget for road maintenance is decreasing in recent years. Efficient road maintenance and management are required to maintain roads with less financial resources. Due to the direct influence by vehicles, road markings are extremely degraded. Currently, the authorities concerned conduct visual examination or investigation by dedicated vehicles to assess the degradation level of road markings. However, such methods are expensive and inefficient. Accordingly, it is essential to estimate the degradation level of road markings at low cost and efficiently. Existing studies estimate the degradation level of road markings from the road surface image captured by the in-vehicle camera. However, these methods have high installation cost because the cameras used in these studies are not common.

Therefore, this paper proposes a method to estimate the degradation level of road markings using a driving recorder or a smartphone as an in-vehicle camera. Also, this study adopts participatory sensing to collect road surface images efficiently and widely. The proposed system scans the captured image, and extracts the part including a road marking based on the painted ratio. The method estimates the degradation level of the extracted marking by comparing the template image without degradation. We conducted the experiments in order to examine the basic performance of the proposed method, and the experimental results suggested the potential and availability of the proposed method.

Keywords: Road surface marking, degradation level, in-vehicle camera, participatory sensing, Intelligent Transport Systems.

1 INTRODUCTION

Roads in Japan continue to grow year by year, and the total length of these roads exceeds 1,270,000 km as of April 1st, 2016. On the other hand, the road project cost is a peak around 1998 and has been decreasing in recent years. It is requested to maintain and manage more efficient roads in order to maintain and manage many roads with limited financial resources.

Roads have various kinds of markers such as road surface makings and road signs, and various kinds of structure such as bridges and tunnels. There are short-term deterioration such as the paint peeling of markers on roads, and long-term deterioration such as the collapse of the tunnel and the rust of the bridge. Especially, road markings are directly affected by the tires running vehicles, and the deterioration is intense due to the road damage in a short period. Also, in the snowy region, road markings remarkably peels in winter due to traveling using the tires chains such as buses and trucks and the repeated snow removal work by the dedicated vehicles.

Currently, road management operator visually inspects such damage and peeling of road markings or inspects them by using the dedicated vehicle. However, such inspection has high cost and inefficient in information collection from the view point of road maintenance. From the view point of drivers who use road, it is necessary to grasp road markings or road signs in order to operate their vehicle safety and comfortably. Information such as the position of the stop line and the traffic classification by travelling direction is important on unfamiliar roads. However, it is difficult to confirm the road marking when the road marking is peeled off.

In this study, we adopt participatory sensing [1] as an approach for collecting the degradation level of road surface markings effectively, and share the information about road markings provided lots of drivers. In this paper, we propose a method to estimate the degradation level of road surface markings as an indicator of how degree the road marking deteriorates, by using the shot images from an in-vehicle camera. We assume that many drivers can use the proposed method by using a driving recorder or a smartphone which becomes popular as an in-vehicle camera.

2 RELATED WORK

As researches using in-vehicle cameras, the related works for recognizing road markings [2, 3] and for estimating the degradation level of road markings [4, 5, 6].

The works [2, 3] recognize road markings by transforming the shot image to the bird’s eye view. When shooting the image including multiple road surfaces by using the in-vehicle camera, the size of the road marking is different depending on the distance from the shooting vehicle to the targeted road sign. By this transformation, the size of the road markings in the image becomes constant irrespective of the distance from the shooting vehicle. Consequently, it is easy to apply various kinds of image processing for recognizing road markings. We think we can apply this transformation [2, 3] into our study.

The works [4, 5, 6] estimate the degradation level of road markings with simple shape. These works use specific camera such as a single lens reflex camera or a spherical camera. The work [4] shoots the image ahead of the vehicle, and estimates the degradation level of line for road partition by adapting the template matching mechanism to the shot image. The line for road partition indicates center line in road and roadway outside line. The work [5] installs a 4K camera on the passenger side of the front bumper of the vehicle, and estimates the degradation level of road markings. The work [6] installs a spherical camera at the bottom of the vehicle. These works need to estimate complete state of the targeted road marking from the shot image in order to calculate the degradation level of the road marking. It is easy to estimate only road markings with simple shape such as road partitioned line. However, it is difficult to apply the estimation of the degradation level of various kinds of road markings that this study will tackle. Also, the installation cost is high because it is necessary to install the camera device not to fall on roads while vehicle traveling.

3 PROPOSED METHOD

3.1 Research purpose

The purpose of this study is to estimate the degradation level of road markings by using the images from in-vehicle camera. We use driving recorders and smartphones as in-vehicle cameras because these devices are becoming popular. We tackle the estimation of the degradation level of various kinds of road markings: line for lane partition, speed limitation, lane classification by traveling direction and approach of pedestrian crossing. In this paper, we focus on lane classification by traveling direction that is composed of several arrows such as straight, right and left.

3.2 Problems and approaches

In order to archive our purpose described in the previous section, we need to solve the following problems:

Problem I: Reduction of the installation cost of devices.
Problem II: Recognition of road signs
Problem III: Identification of road signs
Problem IV: Calculation of the degradation level of road signs

As an approach for solving the Problem I, we use a driving recorder and a smartphone with the high penetration rate as an in-vehicle camera. We need to collect the image including the targeted road signs in wide area. If we use lots of vehicles with in-vehicle cameras, we can collect the image about roads in wide-area effectively. Recently, general drivers often have smartphones and are setting driving recorders in their vehicles. Specific cameras used in the reference [6, 8] are not easy to set the vehicle because these cameras are not designed for in-vehicle use. By adopting smartphones and driving recorders as in-vehicles cameras, lots of general drivers will participate with data collection. Accordingly, the introduction of commercially available smartphones and driving recorders reduce the installation cost for the devices for collecting the images with road markings. Participatory sensing is a powerful approach for collecting the road images in city scale effectively rather than using the dedicated vehicle for road maintenance.

For solving the Problem II, we focus on the rate of white pixels in the extracted road area. We need to recognize where the target road signs are on road, and extract the road area for the next image processing: identification of road markings. Therefore, we slide the particular window along the road, and calculate the rate for each window. If the window includes many white pixels (the rate is high), there is high probability that the window contain a road marking.

For solving the Problem III, we use SIFT (Scale-Invariant Feature Transform) feature quantities and BoVW (Bag of Visual Words) model for identifying road markings. As the image processing after recognizing road markings, we need identify which road markings the extracted markings are. SIFT is one of representative feature quantities for similarity measurement among images, and can classify the target images with high accuracy even if there are the difference in size and angle between the targeted image and template images. We express each road marking as distribution representation based on the calculated SIFT feature quantities. The representation is a BoVW model that consists of vectors (visual words). The BoVW model is adaptation of BoW (Bag of Words) for image domain. We can express a sentence as vectors in the BoW model. In the same way, we can express an image as vectors in the BoVW model.

As an approach for solving the Problem IV, we compare the rate of the painted occupancy for each road markings. The related works [7, 8] can estimate the degradation level of only simple road segmented line due to the characteristics of the processing algorithm. In this study, we tackle on estimation of the degradation level of various kinds of road markings, and cannot adopt the previous methods [7, 8] for our targets. Therefore, we prepare the template for each road marking without damage (namely, whole painted), and estimate the degradation level of the targeted road sign by comparing the extracted road marking in the camera image with the template.

3.3 The details of the proposed method

3.3.1 An overview of the proposed system

The proposed system consists of the part for acquiring camera images and the part for estimating the degradation level of road markings. The first part dedicates to shoot images about road markings by using a driving recorder and a smartphone set on dash-board in each vehicle. The second part estimates the degradation level of road markings by recognition and identification of the targeted road markings for the shot image. The image processing part consists of the following procedure:

(1) Pre-processing before recognizing road markings
(2) Recognition and identification of road markings
(3) Estimation of the degradation level of road markings

If lots of drivers use the proposed system, we can estimate the degradation level of road markings on many roads in wide-area. The estimated results are recorded in a database incorporating with location information. Consequently, we can grasp the degradation level of road markings on a map, and will be also useful for road maintenance.

3.3.2 Pre-processing before recognizing road markings

There are two kinds of pre-processing before recognizing road markings. One is image binarization and another is bird's eye view transformation. By image binarization, we can expect to extract the area including road markings from the in-vehicle camera image because most of road markings are painted with white color. Figure 2 shows the result of the binarization of an example image shown in Figure 1.

We transform the binarized image to bird's eye view. The bird's eye view transformation is to transform like an image taken from directly above by expanding the pixels far from the shooting vehicle. If the original image includes multiple road markings with different distance from the shooting vehicle, the size of the markings will be different and the shape will be distorted. By using the bird's eye transformation, the size of road markings in the shot image can keep the same. Also, it is easy to recognize road markings because the transformation adjusts the distortion in the position and angle of the in-vehicle camera. Figure 3 shows the result of the bird's eye view transformation for the image shown in Figure 2.

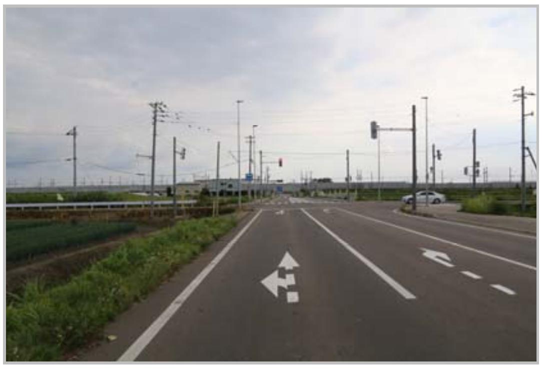


Figure 1: An example of images from in-vehicle camera

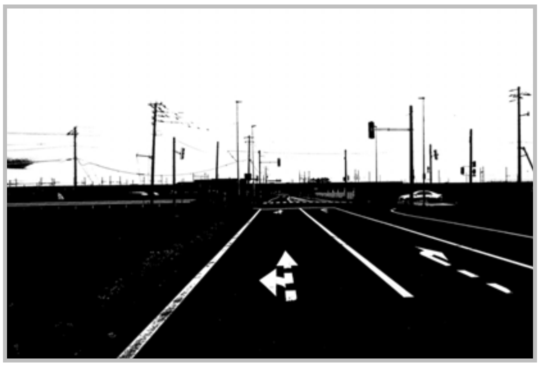


Figure 2: The result of binarization for the image of Figure 1

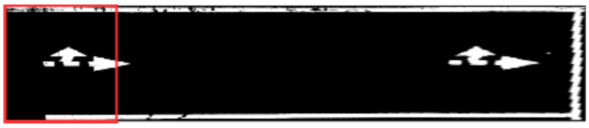


Figure 3: The result of bird's eye view transformation (This image is the transformation result from the image of Figure 2)

3.3.3 Recognition and identification of road markings

Our method recognizes road markings while performing the sliding window processing over the image transformed to bird's eye view. Concretely, we extract the area that is highly possible that it contains a road marking.

We extract the partial image from the transformed view by using the sliding window, and search the position having the maximum number of white pixels while sliding the window. The position having the maximum number of white pixels is the position where it is high possibility to exist road markings.

Next, we calculate SIFT feature quantities from the extracted area including a road marking, generates BoVW vectors and identify road markings based on the similarity among the BoVW representations.

3.3.4 Estimation of the degradation level of road markings

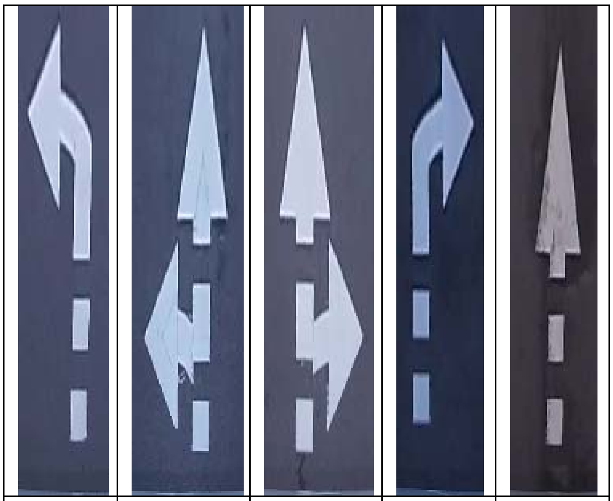
We compare the ratio of the painted occupancy in order to estimate the degradation level of road markings. The ratio of the painted occupancy is an index indicating how degree painting of the road surface marking occupies the targeted area.

First, we prepare the template image for all targeted road markings, and calculate the ratio of the painted occupancy p_{tmp} for the template image in advance. The template image is the image including complete road marking without degradation. Next, it calculates the total number of pixels px_{all} in the area where the road marking exists, and the white pixel number px_{white} of the road marking. We calculate the absolute ratio of the painted occupancy by using the total number of pixels px_{all} and the number of white pixels px_{white}. Finally, we calculate the relative ratio of the painted occupancy p_{rel} by comparing the absolute ratio of the painted occupancy p_{abs} with the ratio of the painted occupancy for the template image p_{tmp}. In this study, we regard the relative ratio p_{rel} as the estimation results of the degradation level of the road marking.

4 EXPERIMENTAL RESULTS

4.1 The experiment about identification of road markings

In this section, we show the result of the experiment of identification of road markings. In this paper, we focus on traffic classification by traveling direction as the representative road markings. As shown in Figure 4, there are 5 types of arrows indicating the traffic classification. The indication for left turn is Arrow L (Figure 4 (a)), the indication for straight and left turn is Arrow SL (Figure 4(b)), the indication for straight and right turn is Arrow SR (Figure 4(c)), the indication for right turn is Arrow R (Figure 4(d)) and the indication for straight is Arrow S (Figure 4(e)).



(a) Arrow L | (b) Arrow SL | (c) Arrow SR | (d) Arrow R | (e) Arrow S

Figure 4: Examples of road markings (arrows for traffic classification) that we used in the
.

The SIFT feature quantities for each image were calculated and these images were classified. Table 1 shows the classification result. We used the leave-one-out method as cross validation.

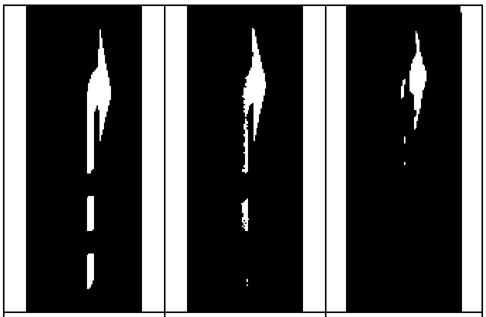
Table 1: The result of classification of arrows

	The Estimated value					
label	L	SL	SR	R	S	Recall
L	10	1	2	7	0	0.500
SL	0	15	15	0	0	0.500
SR	0	14	18	3	0	0.514
R	3	1	12	7	7	0.233
S	0	0	0	1	29	0.967
Precision	0.769	0.484	0.383	0.389	0.806	

The F-measure is 0.541. The estimation result is not good except the Arrow S. There are many mis-estimation between the Arrow SL and SR.

4.2 The experiment about estimation of the degradation level of road markings

In this section, we show the result of the experiment about estimation of the degradation level of road markings. We prepare road markings (Arrow R) with different degradation level as shown in Figure 5. Figure 5(a) shows the arrow with slightly degradation, Figure 5(b) shows the arrow with largely degradation and Figure 5(c) shows the arrow with intensely degradation.



(a) Slightly degradation | (b) Largely degradation | (c) Intensely degradation

Figure 5: Examples of road markings with different degree in degradation

In this experiment, we use Figure (a) as the template image, and calculated the degradation level of the targeted road marking (Arrow R) described in Section 3.3.3. Table 2 shows the estimation result for the images (b) and (c) in Figure 5. This result suggests the proposed method can adequately estimate the degradation level for road markings.

Table 2: The result of estimation of the degradation level

Image type	(a)	(b)	(c)
Absolute ratio of the painted occupancy	5.85	4.15	2.41
Relative ratio of the painted occupancy	(100)	70.9	41.2
The degradation level of the road marking	(0)	29.1	58.8

(unit: percentage)

5 CONCLUSION AND FUTURE WORKS

This paper proposed a method for estimating the degradation level of road surface markings by using the images from in-vehicle cameras. Our methods transform the image to the bird’s eye view, and extract the area including the targeted road markings by sliding window processing. The degradation level of road markings is calculated on based on the ratio of the painted occupancy. We conducted the preliminary experiment by using representative (but simple) road markings and discuss the potential of the proposed method.

In the future works, we should collect many images including road markings and estimate other kinds of road markings. The effective extraction method for collecting the targeted road markings is required. Currently, we calculate the degradation level of road markings based on the ratio of white pixel. In the real environment, there are several road markings with colors except white (for example, orange). We need extend the proposed method to apply such road markings.

REFERENCES

[1] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy and B.M. Srivastava, Participatory Sensing, Proc. of World Sensor Web Workshop (WSW'06) (2006).

[2] M. Noda, T. Takahashi, I. Ide, Y. Mekada and H. Murase, Recognition of Road Markings from In-Vehicle Camera Images using Generative Learning Method, Technical report of IEICE, Vol.108, No.263 (PRMU2008-93), pp.31-36 (2008).*(in Japanese)*

[3] Y. Kemuriyama and K. Onoguchi, Lane Marking Recognition for Constructing a Multi-information Map, Transactions of SICE, Vol.49, No.1, pp.25-32 (2013). *(in Japanese)*

[4] T. Asada, S. Honda and S. Kameyama, The Development of a Method for Estimating the Peeling Ratio of Road Partitioned Line by Using Image Feature Quantities, Journal of Japan Society of Civil Engineers E1 (Pavement Engineering), Vol.67, No.1, pp.10-21 (2011). *(in Japanese)*

[5] S. Nishino, K. Wakayama and H. Kawanaka and K. Oguri, Estimation of the Stripping Ratio from Projection Transformed Images of Lane Markings Using an In-Vehicle Camera, Technical report of IEICE, Vol.114, No.508, pp.17-22 (2015). *(in Japanese)*

[6] T. Kawazaki, T. Uchikoshi, T. Iwamoto, M.Matsumoto, T.Yonezawa, J. Nakazawa and H. Tokuda, A Method for Detecting Friction on Road Markings by Using General-purpose Cameras and General Vehicles, IPSJ SIG Technical Report, Vol.2015-ITS-60, No.2, pp.1-8 (2015). *(in Japanese)*

Session 2:
IoT Systems

(Chair: Ryozo Kiyohara)

A Proposal of PC Facilities Utilization State Management System in Company Office or Educational Institution by using the IoT Technology

Keiichi Abe, Kurika Kobayashi, Kouki Iizuka, Masao Isshiki

Kanagawa Institute of Technology, Atsugi City, Kanagawa, Japan
abe@he.kanagawa-it.ac.jp

Abstract *–*

At present time, information management systems that can survey and monitor an office, movement of people in an office, situation of utilization of facilities and so on in real time have been developed by companies and others using the wireless sensor network and IoT (Internet of Things) technology.

The reason is because information systems aimed at finding waste (electricity usage, unnecessary equipment, work etc.) and improving work environments are being promoted by companies and others. Therefore, in this study, we propose a system that adds a function that can grasp detailed usage of various applications software used within PC work time to the PC management system we developed in the past［1］. We evaluated the effectiveness of our proposed system by comparison with a piece of commercial application analysis software.

Keywords: Personal Computer Management System, Smart Tap , Mat Sensor, IoT Technology

1 INTRODUCTION

At present time, information management systems that can survey and monitor an office, movement of people in an office, state of utilization of facilities and so on in real time have been developed by companies and others using the wireless sensor network and IoT (Internet of Things) technology.

The reason is because information systems aimed at finding waste (electricity usage, unnecessary equipment, work etc.) and improving work environments are being promoted by companies and others.

Our research is to realize an information system that enables us to grasp the operation status of PC facilities in offices and find waste of unused software by accurately monitoring the work situation of PC workers inside offices etc. In our past research [1], we developed a system in which a smart tap and a mat sensor node were installed per PC user, and these two pieces of sensor information were acquired by wireless communication.

We showed that we can accurately calculate the four conditions of PC work time, work time other than PC, wasted electric use time (time when PC is running with no user present), and absence time.

However, in the system we proposed in the past [1], we could not grasp the operating time and frequency of use of each application software used within PC work time. Therefore, unused applications software which have been installed on the PC but are unused cannot be found. In addition, there remains a problem that it is impossible to accurately grasp the usage situation of PC facilities including software.

Currently, there are many dedicated software products that monitor and analyze the operation status of PC applications software, but since those products acquire and manage information retrieved from the log information managed by the OS, they are occasionally prevented from logging due to a communication interrupt or the like. Another problem is that the period of time where the computer is in the ON state with no user present is incorrectly counted as the PC work time. Therefore, with those commercially available application analysis software products, it is impossible to accurately grasp detailed usage of the applications during the PC working time while accurately managing the work state of the PC user.

Therefore, in this research, we propose a system that adds a function that can grasp detailed usage of various applications software used within PC work time to the PC management system we developed in the past. We also evaluated the effectiveness of our proposed system by comparing it with a commercially available application analysis software product (Manic Time).

This paper is organized as follows. Chapter 2 describes related technologies. Chapter 3 gives an overview of the proposed system, and Chapter 4 details the prototype development. In Chapter 5, we will describe the evaluation results of the prototype. In Chapter 6, this research is summarized.

2 RELATED WORKS

Examples of the conventional techniques for checking the presence and attendance of users in a PC practice room include an attendance system which checks the entry and leaving times by IC cards, and an attendance management system which determines attendance of users in a lecture based on a database which holds PC use histories [2] [3] [4] [5].

There are also studies on automated systems for checking the entry and leaving times in a room, such as a system which employs the iBeacon technology[6], a hands-free system for checking the entry, presence and leaving of users[7], and an attendance checking system which uses a camera [8] [9] [10] [11].

These management systems can manage the total number of hours of the presence in the office room, but it is not possible to manage the work time etc of the people actually using PCs.

Moreover, it requires a large-scale construction at the time of introduction, and there is a problem that the introduction cost increases.

There is also an application that monitors the operating status of software using log data of a personal computer [12] [13] [14]. This application can acquire both information such as usage time of various software and information on the ON / OFF information of the power supply of the PC etc. However, this commercially available analysis software cannot record data when an interruptive event such as an update program occurs during the PC operation; in such a situation, the data-recording stops and it is impossible to grasp the use situation of the PC. Also, because it is not possible to acquire the presence information and it does not always mean that a person is working on the PC even if the power supply of the PC is in the on state, it is considered difficult to measure an accurate PC work time.

Therefore, in this paper, we propose a system that can distinguish between “PC work” and “work other than PC” and can grasp and monitor the usage status of various software in detail during the PC working time. We also propose a method that can reduce the construction cost at the time of introduction.

3 SUMMARY OF PC Facilities Utilization State Management System

3.1 Design concept

I. Installation Location

The system is supposed to be installed in a computer room such as company office or educational institution.

II. Personal Computer to be Monitored

The use of desktop personal computers is assumed. Mobile computers are excluded.

III. Person to be monitored

PC users in the office and PC training room.

IV. Scale of data management

On the assumption that this system will be used in a PC practice room etc. in corporate offices and educational institutions, the system should manage the maximum number of about 100 people.

V. Method of data collection

In this system, sensor data is collected from various sensor nodes by wireless communication, such as Wi-Fi standards. One data collection personal computer (host computer) also uses the same wireless communication standard as the sensor node, and collects data from various sensor nodes with a dedicated application.

VI. Number of sensor nodes installed

The number of various sensor nodes may be about 2 nodes for each PC user.

VII. System installation method

This system uses sensor nodes equipped with a wireless communication function (Wi-Fi, IEEE 802.15.4, etc.). This allows the system to be installed with a free layout. This makes it easier for individuals with no technical skills to install the system and, as a result, reduces its introduction cost .

Ⅷ. Contents of information on PC users to be managed

For the management of information on PC users in offices or the like, this system should have the following functions:

(1) Grasp the number of PC users

(2) Manage the presence status of each PC user.

(3) Distinguish between the PC work time and work time other than PC by the user, and find the actual use time of the personal computer.

(4) Count and manage the time when the personal computer is running with no user present and wasting electricity.

(5) Record the usage time and frequency of use of each software application used within the actual use time of the personal computer

In our previous system [1], we realized the above functions (1) to (4), but we did not fully deal with (5). In this paper, we describe a management method for (5).

3.2 Summary of this system

Section3.2 explains the outline of the proposed system for managing the PC work situation by users in an office or the like. Figure1 shows configuration of smart tap and mat sensor node installed per PC Worker. This time we developed a dedicated application that detects the software application being used on the client PC. Thereby, The our system can be transmitted information on the application currently used by the user.

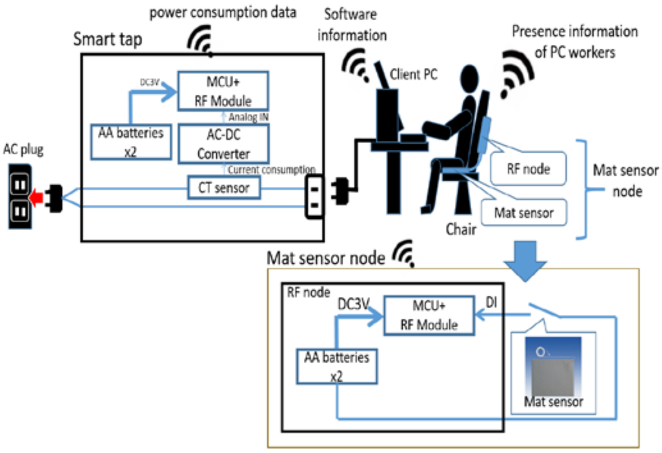


Figure 1: Configuration of smart tap and mat sensor node installed per PC.

In order to monitor the operation status of each desktop personal computer installed in an office or the like, the smart tap constantly measures the power of the desktop personal computer being used and sends the result to the host computer for data collection. Also, in each personal computer, a commercially available mat sensor detects the presence of the PC user, and the detection result is transmitted to the host computer by wireless communication. A commercially available mat sensor equipped with a wireless communication function is called a “mat sensor node” in this paper. We have originally developed the smart tap and mat sensor node for this system. In our proposed system, wireless sensor nodes such as smart taps and mat sensor nodes are used. A general-purpose wireless communication standard (Wi-Fi, IEEE 802.15.4, etc.) is used for communication between these wireless sensor nodes and the host computer for data collection. Therefore, when introducing this system, its components can be freely laid out, so no wiring work is required.

Figure 2 shows an situation used in a PC classroom such as an office and an educational institution. The host computer collects and manages presence information and PC power consumption information from each client. In addition, this time we also collect data on information on the application currently used by the user.

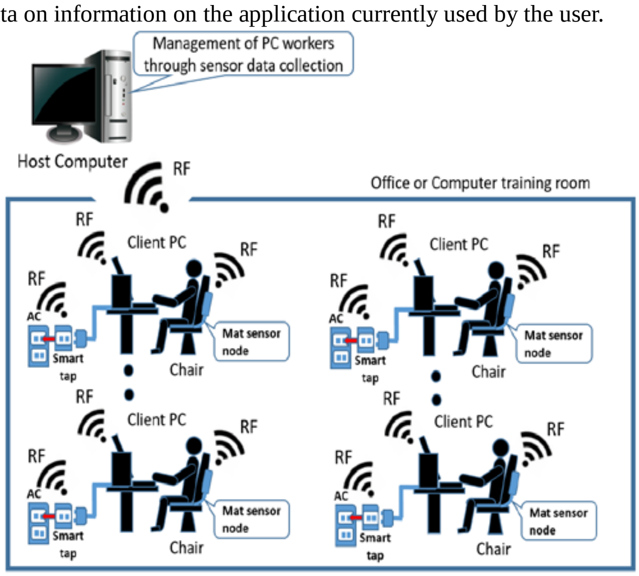


Figure2:Overview of our system.

In the proposed system, the number of PC users is determined by counting the user-detection signals from the mat sensor nodes. The actual use time of each PC (hereinafter referred to as the PC work time) and other related information are accurately calculated from the power-use signal from the smart tap and the user-detection signal from the mat sensor node. This calculation method will be described in detail in Section 3.3. This algorithm was implemented in a dedicated application of the host computer.

In addition, a dedicated application for acquiring information on the application currently used by the user and sending it to the host computer is implemented on each client PC. This application enables us to analyze details of the use of the various applications on each PC within the PC usage time.

Figure 3 shows outline of the utilization analysis image graph of PC facilities in the our proposed system By using this system, it is possible to make detailed breakdown of usage time etc of various applications used within PC work time in addition to the four major divisions, i.e. the PC work time, work time other than PC, wasteful electricity use time, and absence time for each PC user. By doing this, our system can monitor and manage the utilization situation of PC facilities installed in a PC room etc of offices and educational institutions.

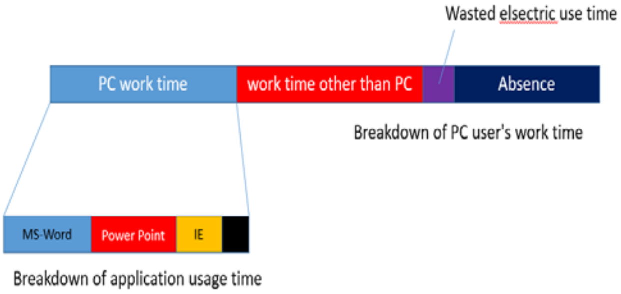

Wasted elsectric use time
PC work time
work time other than PC
Absence
Breakdown of PC user's work time
MS-Word
Power Point
IE
Breakdown of application usage time

Figure 3:Outline of the utilization analysis image graph of PC facilities in the our proposed system.

3.3 Calculation method of person's presence and PC working time [1]

In our past research, as a method to grasp details of the work situation of users' PCs etc., we used the sleep function of the PC in addition to the two pieces of sensor information acquired from the smart tap and mat sensor node. By the method, we could accurately calculate the four states of PC work time, work time other than PC, From the two kinds of sensor data acquired with the smart tap and mat sensor, we can estimate the activity of the user in four ways as shown in Table 1.

Table 1: Each sensor information and action estimate of the user.

A Smart tap (PC-ON/OFF)	B Mat sensor (Sitting/Not sitting)	Action estimate of the person
0	0	Absence (PC-OFF/Not sitting)
0	1	Work time other than PC (PC-OFF/Sitting)
1	0	Wasted electric use time (PC-ON/Not sitting)
1	1	PC work time (PC-ON/Sitting)

Firstly, we understand the ON/OFF state of the PC from the power consumption information of the smart tap. An output of "1" from the smart tap shows that the PC is ON, and "0" shows that the PC is OFF. In the system, the power consumption threshold is 15 W. Therefore, if it exceeds 15 W, the output is "1" when the power consumption exceeds 15 W and "0" when less than 15 W. On the other hand, an output of "1" from the mat sensor indicates a state where someone is in the chair, and "0" indicates no one is in the chair. It is possible to manage one PC by the two pieces of sensor information.

When A = 0 and B = 0, the PC is in OFF state and no user is present, hence "Absence". A = 0 and B = 1 indicates that the PC is in the OFF state, but because someone is present, it indicates "Work time other than PC". When A = 1 and B = 0, the PC is in the ON state, but because no one is present, it indicates "Wasted electric use time". When A = 1 and B = 1, since the user is present while the PC is ON, this state is judged to be "PC work time" in this study.

When the output of the smart tap is A and the output of the mat sensor is B, the output W in the absence state, the output X in the working state other than the PC, the output Y in the waste electric utilization state, the output Z in the PC working state, the following formulae (3.1) to (3.4)are derived from Table 1.

$$W = \overline{A} \cdot \overline{B} \qquad (3.1)$$

$$X = \overline{A} \cdot B \qquad (3.2)$$

$$Y = A \cdot \overline{B} \qquad (3.3)$$

$$Z = A \cdot B \qquad (3.4)$$

Our system defines the output Z of equation (3.4) as PC working, but a person sitting in the chair is not necessarily doing PC work but may actually be doing other work than PC. Therefore, in this research, in order to distinguish between PC work and non-PC work, the sleep function of the PC is utilized in addition to the two kinds of sensor data acquired from the smart tap and mat sensor node. More specifically, while the PC is ON, if the user sitting on the chair does not perform any PC operation for a certain time, the PC automatically goes to the sleep mode and its power is turned off. This indicates the beginning of the work other than PC.

According to our prototype system experiment result [1], if we shorten the sleep time of the PC, the difference from the correct answer data recorded with the camera became small. The result showed that the four conditions of PC work time, different work time other than PC, useless electric usage time, and absence time can be measured with high accuracy. However, setting a short sleep time may be inconvenient for the user in PC work.

4 PROTOTYPE SYSTEM IMPLEMENTATION

Section 4 describes the prototype development of the system proposed in this paper. In the prototype developed this time, in order to monitor the situation of PC work by users, one smart tap and one mat sensor node were arranged for one PC as shown in Figure 1. One host computer was installed to collect sensor information from each sensor node. The communication between each sensor node and the host computer was performed through Wi-Fi wireless connections via a Wi-Fi wireless communication router. In Chapter 4, the prototype development of application on client PC, application on host computer, etc. will be described in detail.

4.1 Application on client PC

This time, we developed a dedicated application that detects the software application being used on the client PC.

Figure 4 shows the operation flow of the dedicated application on the client PC. In this application, software information about which application is displayed on the forefront window is acquired from the OS (Windows) and sent to the data collection PC.

Here, we will describe a more detailed method of obtaining software information. Some software programs are executed as background processes on the PC. We only need information on the software currently operated by the user. For this reason, we decided to acquire software information only from the window displayed at the forefront of the computer screen in the application developed this time. because the unused application may be got when multiple windows are displayed on the computer screen.

Therefore, there is a problem that application information which is not actually used is also acquired, in this Application of client PC, only the application information of the window used at the forefront is acquired and sent to the host computer.

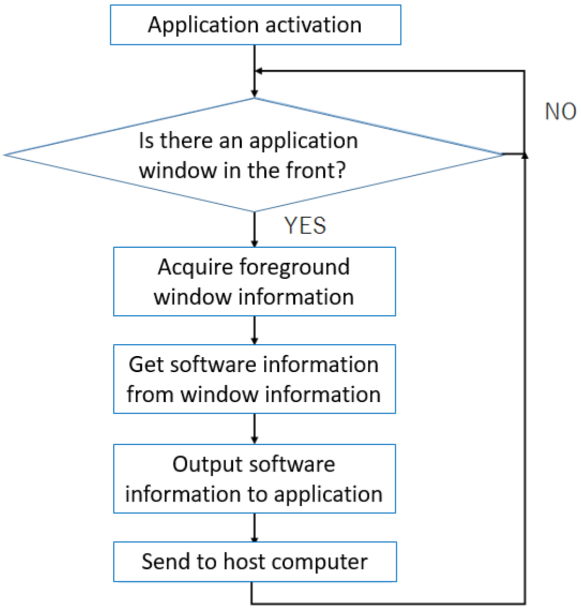
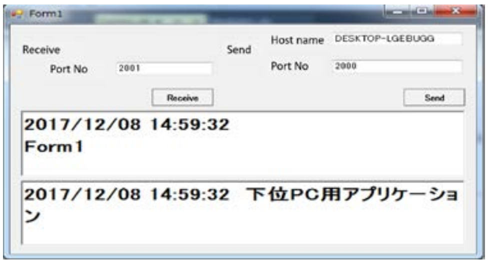

Application activation
Is there an application window in the front?
NO
YES
Acquire foreground window information
Get software information from window information
Output software information to application
Send to host computer

Figure 4: Application operation flow of client PC.

Figure 5 shows the application screen of the client PC actually developed. Information on the application in the forefront window is acquired from Windows OS. As shown in Figure 5, the acquired information, including the date, time and application name, is displayed on the text box. This information is also transmitted to the host PC by wireless communication (Wi-Fi).



Host name DESKTOP-LGEBUGG

Port No 2001

Port No 2000

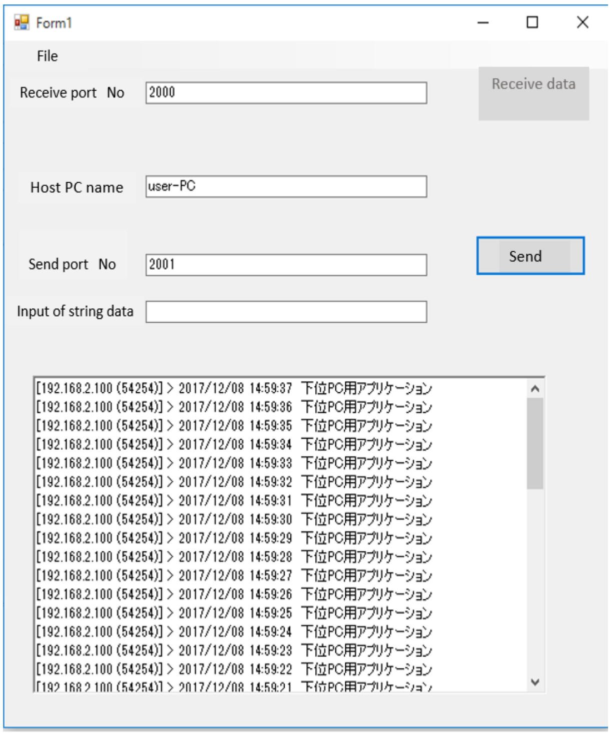
```
2017/12/08 14:59:32
Form1
```

```
2017/12/08 14:59:32 下位PC用アプリケーショ
ン
```

Figure 5: The application screen on the client PC developed this time.

4.2 Application of host computer

The screen of the application on the host computer developed this time is shown in Figure 6. The application on the host computer receives the information on the used software transmitted from the client-PC application, the information on the PC operation state from the smart tap, and the presence information from the mat sensor. For the management of the presence information and the operation status of the personal computer, the operation time and the like of the personal computer are calculated by the same method as the conventional system, and the operation state is displayed and recorded on the application screen on the host computer side. The software information used in each client PC was developed so that the information on the used software transmitted from the client PC can be saved in a file in CSV format. The applications on the host computer and client PC were developed with Visual Basic 2015 language.



Receive port No 2000

Host PC name user-PC

Send port No 2001

Input of string data

```
[192.168.2.100 (54254)] > 2017/12/08 14:59:37  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:36  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:35  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:34  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:33  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:32  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:31  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:30  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:29  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:28  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:27  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:26  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:25  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:24  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:23  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:22  下位PC用アプリケーション
[192.168.2.100 (54254)] > 2017/12/08 14:59:21  下位PC用アプリケーション
```

Figure 6: Data collection application screen on host computer.

5. EVALUATION EXPERIMENT OF PROTOTYPE SYSTEM

5.1 Experimental method

A prototype of the system proposed in this paper was developed and experimental evaluation was carried out. The experiment site was the E602 room of Kanagawa Institute of Technology C2, 6th floor. The experiment period was about 1 hour from 16:31:21 to 17:41:12 on December 18, 2017. In this example, each node device was arranged as shown in Figure 10 so as to be able to measure the operating situation of PC for one person. In addition, one application which can analyze software usage was installed on one server PC and one client PC. On the client PC, the sleep function of the OS of the personal computer (sleep timeout set to 3 minutes) was used to measure the total PC use time and total PC sleep time, and it was evaluated whether various software information can be acquired accurately within PC use time.

In the experimental evaluation of this time, the software operation time was calculated from the video taken by the camera as the correct answer data. We evaluated the proposed system by comparing the correct answer data with the result obtained by a commercially available software analysis application (Manic Time / Finkit doo) and the system we proposed this time. In order to evaluate whether the data acquired in our system were accurate, two cameras were prepared as shown in Figure 7, one was used to check the operation status of the software and installed in front of the screen. The second was installed at the position shown in Figure 7 in order to confirm the presence of the user and PC operation status. Regarding the PC operation status, it was judged to be ON when the screen of the display was lit, and it was judged to be OFF when the screen was dark. We evaluated this proposed system by comparing the videos taken with these two cameras (as correct answer data) with the data measured by this prototype system

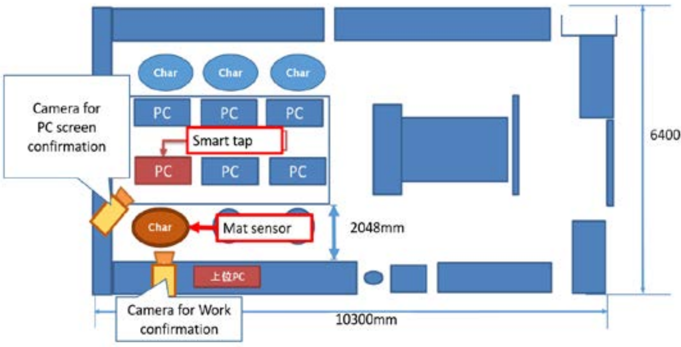


Figure 7:Experiment layout (Camera position etc.)

5. 2 Results of the experiment

Table 2 shows the total time of PC work and the total time of work other than PC, comparing the proposed system and camera (correct answer data). In the proposed system, both the PC work time and the work time other than PC had an error of about ± 3 minutes compared to the camera. The sum of the PC work time and the work time other than PC was the same as with the camera. Therefore, it is likely that the error of breakdown of PC work time and work time other than PC changes depending on the PC sleep time.

Table 2：Measurement result of total PC work time and total PC sleep time

	Camera (Video)	Our prototype system	Error	
Total of PC work time	0:48:30	0:51:31	−	0:03:01
Total of work time other than PC	0:21:17	0:18:16	+	0:03:01

Table 3：Measurement result of operation time of various applications by comparing camera and Manic Time

Various applications	Camera (Video)	Manic Time	Error	
Client application (Our application)	0:01:21	0:01:14	+	0:00:07
explorer	0:03:56	0:03:57	−	0:00:01
chrome	0:12:45	0:13:03	−	0:00:18
ManicTimeClient	0:02:18	0:02:20	−	0:00:02
notepad	0:04:06	0:04:15	−	0:00:09
WINWORD	0:22:46	0:23:15	−	0:00:29
dllhost	0:00:05	0:00:06	−	0:00:01

Table 4：Measurement result of various application usage times by comparing camera and prototype system

Various applications	Camera (Video)	Our prototype system	Error	
Client application (Our application)	0:01:21	0:01:21	±	0:00:00
explorer	0:03:56	0:03:56	±	0:00:00
chrome	0:12:45	0:12:50	−	0:00:05
ManicTimeClient	0:02:18	0:02:18	±	0:00:00
notepad	0:04:06	0:04:13	−	0:00:07
WINWORD	0:22:46	0:22:45	+	0:00:01
dllhost	0:00:05	0:00:05	±	0:00:00

Table 3 shows the results of running time of various applications, comparing the commercial application analysis software (Manic Time) with the camera, and Table 4 shows the comparison between our prototype system and the camera. Compared to the camera, the commercial software (Manic Time) had errors of up to 29 seconds in operation time of various software. By comparison, in the proposed system, the error was smaller and within 7 seconds.

Next, a graph of various software analysis results used for PC working time is shown in Figure 8. As shown in Figure 8, in the proposed system, there is no loss of software information and the error is small. The results of graphing the analysis results of various software of the proposed system and rearranging them in order of frequency of use are shown in Figure 9. The graphical representation as shown in Figure 9 enables easy distinction between frequently used ones and less used ones.

From the above results, we confirmed that the proposed system can analyze the usage time of various software applications used during the PC work in detail. Moreover, it was confirmed that the operation time of each software application can be measured with higher accuracy than the commercially available application analysis software Manic Time.

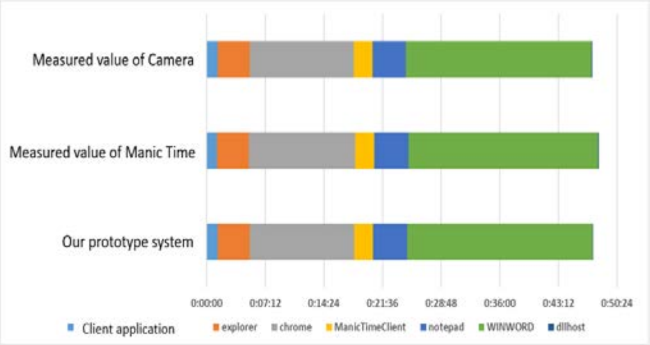


Figure 8: Analysis results of various software by comparing camera, Manic Time and proposed system.

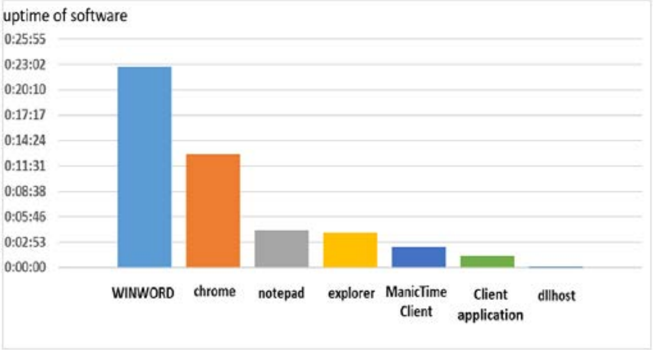


Figure 9: Analysis result of frequently used ones and low ones of various software usage.

6.CONCLUSION

In this paper, we described a system with which we can grasp various applications used during the PC working time in detail for the PC management system we developed in the past.

In the experimental evaluation of the prototype of the proposed system of this paper, the error was less than 7 seconds as compared with the correct data recorded with the camera. The error was 29 seconds at maximum in the case of Manic Time, which is one of the commercially available application analysis software products. It was confirmed that this error occurred greatly when the PC goes to sleep and returns from sleep. A likely reason for the large error which occurs in the case of the commercially available application analysis software is because the application is given a low processing priority on the PC and is forced to wait for the log data to be saved in the PC before being available for readout by the application.

Since our proposed system adopts the method of directly acquiring data from the window information on the screen, unlike the commercially available software, the processing time is not required and it is possible to obtain accurate information on the usage situation of each software application.

Therefore, we found that the system we propose can accurately grasp the usage time of each software application used during the PC work time than the commercial application analysis software (Manic Time) which relies solely on the log data. In addition to this, PC work hours, work hours other than PC, absence time etc can be accurately grasped. It is expected that our system will be practically applicable in the future as a system to monitor the utilization situation of PC facilities in office and educational institution PC practice rooms in real time.

REFERENCES

[1] Y.Takabayashi, Y.Kagami, H.Kawauchi, M.Isshiki, K.Abe, “Development of PC management system for the office and the PC practice room using IoT technology(In Japanese)", *International Workshop on Informatics(IWIN2017)*, pp.3-8 (2017).

[2] N. Morita, “Development and Result of Attendance Registration System Using IC cards(In Japanese)”, *The Institute of Electronics, Information and Communication Engineers (IEICE)/Forum on Information Technology (FIT)*, 11(3), pp.65-68 (2012).

[3] H. Suda, S. Nakamura, M. Ogawa, H. Kumazawa, N. Komdou, “Attendance Management System using Password Distribution of NFC Tag with Electronic Paper(In Japanese)”, *The 76TH National Convention of IPSJ*, No.4, pp.853-854 (2014).

[4] S. Kubota, S. Furukawa, Y. Soejima, R. Kawamura, and K. Sugitani,”Presence-type Attendance Management System in Educational PC Classrooms(In Japanese)”, *The Institute of Electronics, Information and Communication Engineers (IEICE)/ Forum on Information Technology (FIT)*, 8(4), pp.175-177 (2009).

[5] H. Matsumoto, S. Wada, S. Hara, N. Morita, “Development and Evaluation of Attendance Registration System using Suica(In Japanese)”, *Proceedings of the School of Information and Telecommunication Engineering Tokai University*, Vol.1, No.2, pp.16-21 (2011).

[6] K. Tanaka, K. Shuwa, “Application of iBeacon in Laboratory Room Management System Automation(In Japanese)”, *Tokyo City University Yokohama Campus Journal of Information Studies*, Vol.4, No.16, pp.31–37 (2015).

[7] N.Takayama, M.Kitamura, “Hand Free Entry/ Attendance Management System(In Japanese)”, *NEC Technical Report*, Vol.63, No.3, pp.60–63 (2010).

[8] S. Ono, H. Yasuda, “Proposal of Setting Management System”, *Proceedings of the 2014 IEICE General Conference*, p.184 (2014).

[9] K. Onozawa, K. Watanabe, K Suwa, “Room Management System and Learning Situation Management System Using Smartphone(In Japanese)”, *Tokyo City University Yokohama Campus Journal of Information Studie*s, Vol.4, No.13, pp.6-15 (2012).

[10] H. Suda, S. Nakamura, M. Ogawa, H. Kumazawa, N. Komdou, “Attendance Management System using Password Distribution of NFC Tag with Electronic Paper(In Japanese)”, *The 76TH National Convention of IPSJ*, No.4, pp.853-854 (2014).

[11] N. Nishimura, “New・Suddenly Attend(In Japanese)”, *Journal of Yasuda Women’s University*, Vol.41, pp.371-379 (2013).

[12] ManicTime https://www.manictime.com/

[13] Microsoft “Software metering in System Center Configuration Manager**(In Japanese)**” 〈https://technet.microsoft.com/ja-jp/library/gg682005.aspx〉 (2018).

[14] Hitachi,“JP1 Version 8 NETM / DM Installation and Design Guide (for Windows (R)) Monitor the operation status of software(In Japanese)”, 〈http://itdoc.hitachi.co.jp/manuals/3020/30203L36A0/DMDS0011.HTM〉 (2017).

[15] CQ publishing companies, “Web brain connection Wi-Fi×3G/LTE of IoT(In Japanese)”, *Transistor technology*, Vol.9, pp.35-93 (2016).

Proposal of IoT Edge Optimization Model for Balancing Local System and Total System and its Simulation Evaluation

Shinji Kitagami [*], Yuichi Tokunaga[**] , and Norio Shiratori[***]

[**] Department of Management and Information Sciences, Fukui University of Technology, Japan
[**] Information Technology R&D Center, Mitsubishi Electric Corporation, Japan
[***] Research and Development Initiative, Chuo University, Japan

Abstract - In IoT-based system, there is trade-off relationship between the optimization for total system in the cloud computing and the optimization for local system in the edge computing. That is, the result of the cloud optimization may become a constraint condition for the edge optimization and vice versa. In this paper, we propose an IoT edge optimization model for balancing the total system and local system. In this optimization model, the balance of both cloud optimization and edge optimization is treated as Pareto optimization problem. Further, the cloud optimization process and the edge optimization process negotiate each other by an intelligent protocol to find the optimal balancing condition.

Keywords: IoT System, Cloud Computing, Edge Computing, Optimization Problem

1 INTRODUCTION

IoT systems that connect many sensors and devices directly to the Internet and provide various services without human intervention are expanding in the industrial sector, the home sector, and the social sector [1]. The conventional IoT system is cloud-centric IoT computing model (CC-IoT model) based on the cloud computing [2]. The CC-IoT model realizes optimization of the system using various data collected from sensors installed widely. On the other hand, edge-centric IoT computing model (EC-IoT model) has been proposed to solve the problem of the CC-IoT model such as increase of network load and delay of feedback response [2][3]. The EC-IoT model performs data processing at the edge server and the IoT gateway close to the data source and the device to be controlled for local optimization within the edge area. However, in many IoT systems, there is trade-off relationship between cloud optimization and edge optimization. That is, the result of the cloud optimization may become a constraint condition for edge optimization and vice versa.

In this paper, we propose an IoT edge optimization model for balancing the cloud optimization and edge optimization. In this optimization model, the balance of both optimizations is treated as Pareto optimization problem. Also, the cloud optimization process and the edge optimization process negotiate each other by an intelligent protocol to find the optimal balancing condition.

2 OPTIMIZATION IN CLOUD AND EDGE

The IoT architecture has changed from the vertical integration type to the horizontal integration type and the distributed type [1] [2]. In this paper, we refer the horizontally integrated IoT architecture to the cloud-centric IoT computing model (CC-IoT model). Also, the distributed IoT architecture is referred to as the edge-centered IoT computing model (EC-IoT model) in this paper. The CC-IoT model optimizes the system by utilizing various data collected from many sensors. In this paper, optimization by CC-IoT model is referred to as the cloud optimization. For example, the CC-IoT model is suitable for optimizing the energy supply-demand balance in the energy management system (EMS) and optimizing the relaxation of traffic congestion by ITS [4] [5].

However, recently, some problems such as an increase in network load and a delay in feedback control have been pointed out in the CC-IoT model [2] [3]. The EC-IoT model proposed to solve these problems optimizes the local system in the edge area such as buildings and vehicles close to the data source. In this paper, the optimization by EC-IoT model is referred as to the edge optimization. For example, the EC-IoT model is suitable for optimization for the energy saving and comfort in a building by EMS and automatic drive control by V/C for comfort and safety of drivers [5].

In general, there is trade-off relationship between the CC-IoT model and the EC-IoT model. In other words, the result of the cloud optimization may be a constraint condition for edge optimization. Also, the result of edge optimization may be a constraint condition of cloud optimization.

3 IOT EDGE OPTIMAZATION MODEL

3.1 Basic Concept

In this paper, we propose an IoT edge optimization model to realize effective next-generation mobility services. Figure 1 shows the basic concept of the IoT edge optimization model. In the figure, the vertical axis represents the cost by ITS cloud optimization, and the horizontal axis represents the cost by V/C edge optimization. Both of which are higher optimization level as they are closer to zero. Also, each cost has an acceptable range. For example, there is an allowable cost range of the degree of congestion in the cloud optimization by ITS, and there is an allowable cost range of the driver's comfort in the edge optimization by V/C. In addition, the curve in figure 1 is a Pareto optimal curve. In the case where priority is given to the cloud optimization by ITS, the cost of the edge optimization does not fall within the allowable range. Conversely, when priority is given to the edge optimization by V/C, the cost of the cloud optimization cannot be within the allowable range.

In the IoT edge optimization model proposed in this paper, the cloud optimization process and the edge optimization process negotiate by an intelligent protocol and balance the cloud optimization and the edge optimization.

3.2 Formulations

The formulations of the IoT edge optimization model is shown as below.

- Cloud Optimization
  $$MIN_{v_s,v_c}(cost_c(\boldsymbol{v_s}, \boldsymbol{v_c}))$$
  $$under\ constraints_c(\boldsymbol{v_s}, \boldsymbol{v_c})$$

- Edge Optimization
  $$MIN_{v_s,v_e}(cost_e(\boldsymbol{v_s}, \boldsymbol{v_e}))$$
  $$under\ constraints_e(\boldsymbol{v_s}, \boldsymbol{v_e})$$

- Total Optimization
  $$MIN_{v_s,v_{c,}v_e}(cost_t(\boldsymbol{v_s}, \boldsymbol{v_c}, \boldsymbol{v_e}))$$
  $$under\ constraints_t(\boldsymbol{v_s}, \boldsymbol{v_c}, \boldsymbol{v_e})$$

  $$cost_t(\boldsymbol{v_s}, \boldsymbol{v_c}, \boldsymbol{v_e}) = cost_c(\boldsymbol{v_s}, \boldsymbol{v_c}) + k * cost_e(\boldsymbol{v_s}, \boldsymbol{v_e})$$

Here,
$\boldsymbol{v_s}$: shared variable vector
$\boldsymbol{v_c}$: cloud variable vector
$\boldsymbol{v_e}$: edge variable vector
$cost_c(\boldsymbol{v_s}, \boldsymbol{v_c})$: objective function of cloud optimization
$cost_e(\boldsymbol{v_s}, \boldsymbol{v_e})$: objective function of edge optimization
$cost_s(\boldsymbol{v_s}, \boldsymbol{v_c}, \boldsymbol{v_e})$: objective function of system optimization

The IoT edge optimization model minimizes the objective functions of cloud optimization, edge optimization, and system optimization by exchanging shared variable vectors with the intelligent protocol.

3.3 Intelligent Protocol

In the formulation mentioned above, the cloud variable vector $\boldsymbol{v_c}$ used only in the optimization process on the cloud side cannot be referred from the edge side. Likewise, the edge variable vector $\boldsymbol{v_e}$ used only on the edge side optimization process cannot be referred from the cloud side. For that reason, in the IoT edge optimization model, shared variable vector $\boldsymbol{v_s}$ are exchanged between cloud and edge by the intelligent protocol. Figure 2 shows the intelligent protocol for exchanging shared variable vectors $\boldsymbol{v_s}$ and negotiating between cloud and edge optimization processes. First, as shown in the figure, $\boldsymbol{v_s}$ which minimizes the cloud

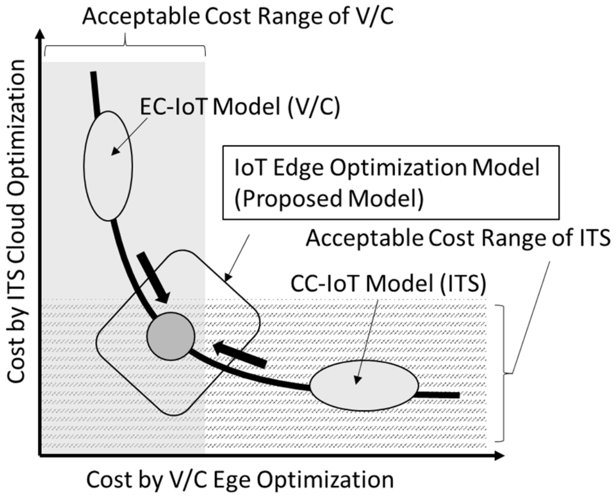


Figure 1. Balancing Cloud Optimization and Edge Optimization

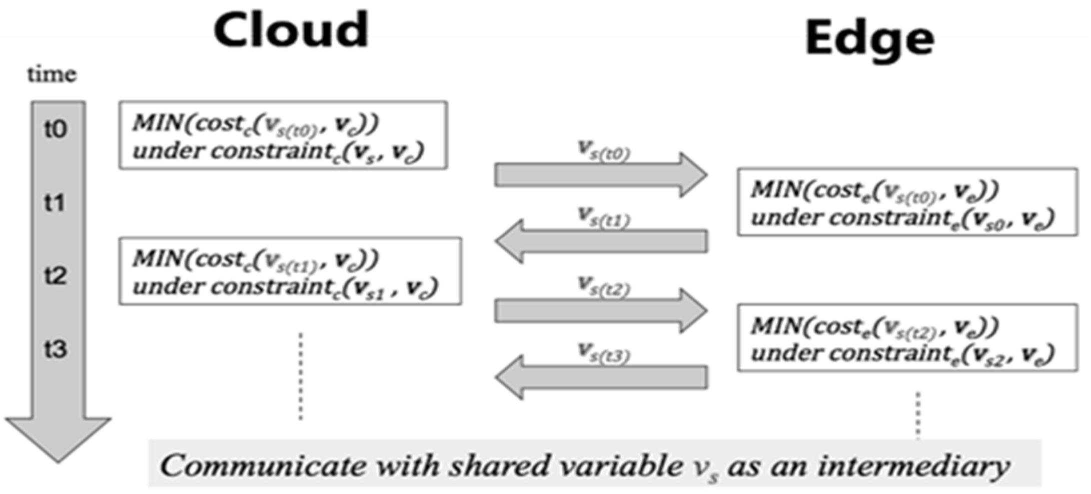

Cloud
Edge
time
t0
t1
t2
t3
MIN(cost_c(v_s(t0), v_c))
under constraint_c(v_s, v_c)
v_s(t0)
v_s(t1)
MIN(cost_e(v_s(t0), v_e))
under constraint_e(v_s0, v_e)
MIN(cost_c(v_s(t1), v_c))
under constraint_c(v_s1 , v_c)
v_s(t2)
v_s(t3)
MIN(cost_e(v_s(t2), v_e))
under constraint_e(v_s2, v_e)
Communicate with shared variable v_s as an intermediary

Figure 2. Intelligent Protocol of IoT Edge Optimization Model

optimization cost is transmitted from the cloud to the edge. On the edge side, adjust the value of $\boldsymbol{v_s}$ to minimize the edge optimization cost and send it back to the cloud. By repeating this negotiation, each optimization processes are balanced so that the optimal cost of the cloud, the edge, and the entire system are minimized.

4 EVALUATION

As described in Chapter 2, to realize effective next-generation mobility service, it is necessary to balance the optimization of cloud by ITS and the edge optimization by V/C. Here, representative cloud optimization is to eliminate the traffic congestion in the whole town, and representative edge optimization is to maintain the comfort of the driver of the vehicle by reaching the destination quickly. When the traffic volume on the road is low, it is easy to minimize the objective function of both optimizations. However, when the traffic volume increases on the road, it is necessary to balance the ease traffic congestion and drivers' comfort. Applying the IoT optimization model in which ITS and V/C negotiate by the intelligent protocol, it is expected to solve this problem.

To evaluate the effectiveness of the IoT edge optimization model proposed in Chapter 3, we conducted a simulation in which the traffic congestion mitigation control by ITS and the automatic operation by V/C cooperate. In this simulation, we used the iGraph package of R [6].

In this simulation, it is assumed that four vehicles simultaneously depart from different starting points on the road in a grid pattern including 25 intersections (nodes). The objective function of edge optimization by V/C is the sum of the moving times of all vehicles, and the objective function of cloud optimization by ITS is the sum cost of each node. The initial value of the weight of the path corresponding to the moving time between nodes is 2 for the peripheral path and 1 for the other internal paths. In addition, the cost of each node is set in response to the degree of traffic congestion, and the weight of the surrounding paths are added according to the cost of the node.

V/C determines the shortest path that the vehicle moves from the start node to the destination node by the Dijkstra method. If the cost value is set for the nodes on the path, the path determination algorithm adds the value to the weight of the path around the node. The path weight means a movement time among nodes. After determining the shortest path, V/C transmits the scheduled passing time, a cumulative value of the path weight, of each node to the ITS. If the node cost value is transmitted from the ITS, V/C determines the shortest path again by adding the cost to the surrounding pass. Although it is necessary to calculate the shortest path continuously while the vehicle is moving, we assume that V/C determines the shortest path before moving in this simulation.

The ITS presumes congestion degree of the node based on the expected passing time of the node receiving from the V / C of vehicles in the system. That is if two vehicles pass through at the same node simultaneously, the ITS adds 1 point to the cost of the node. If two or more vehicles pass at the same node simultaneously, the ITS adds the number of combinations of the simultaneous passage to the cost of the node. For example, if 4 vehicles pass at the node simultaneously, its node cost becomes 6. After calculating the cost of all the nodes, The ITS notifies the largest cost of the node to the V/C of all vehicles. If there are multiple of nodes with the largest cost, the ITS selects a node randomly and notifies it to the V/C.

Figure 3 shows the relationship between cloud optimal cost and the edge optimal cost for the optimization phases 1 to 8 as a summary of the simulation results. As shown in the figure, if the cloud optimization cost is 3 or less, the edge optimization cost will be 20 or more. In that case, although the number of congested nodes is limited to about 2, the traveling time of the car increases more than twice. Conversely, optimizing the edge optimization cost to be 10 or less, the cloud optimization cost becomes 6 to 10, and it turns out that the number of congested nodes increases. Also, it was found that when the cloud optimization and the edge optimization are balanced in the state of the optimization phase 5. In this balanced point, the cloud optimization cost is 4 and the edge optimization cost is 16.

5 DISCUSSION

As shown in the simulation result of Figure 3, there is a trade-off relationship between the cloud optimization by the ITS and the edge optimization by the V/C. In this simulation, negotiating between the cloud optimization process and the edge optimization process, we confirmed that it is possible to balance both optimizations in the optimization phase 5. In fact, whether the balanced state of two optimizations is valid depends on the allowable range for each optimization. In the proposed model, we can select various combinations of the cloud optimization and the edge optimization depends on stakeholders' requests. Since the simulation shown in this paper was aimed at verifying the principle of optimization balancing by IoT edge optimization model, the simulation model was simplified. Furthermore, to perform a practical simulation, it is necessary to consider the difference in the moving start time and the moving direction of vehicles in the system. That is, it is assumed that congestion occurs in a node when there are vehicles passing through the node at the same

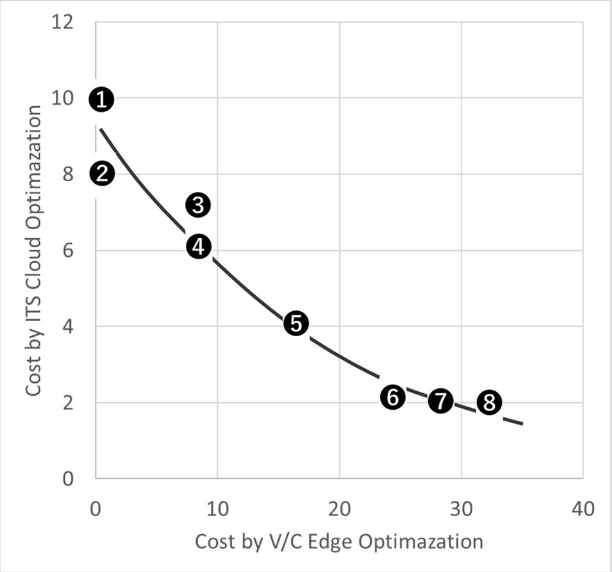


Figure 3 Balanced Optimization of Cloud and Edge

time in this simulation. However, in fact, the case such as to pass each other in the opposite direction and not to be congested at the node shall be considered.

The IoT edge optimization model can balance the cloud optimization and the edge optimization according to different requirements of the stakeholders of the system. In the energy management system, power companies are stakeholders on the cloud side and the supply balance of electric power is one of the objective functions of cloud optimization. Meanwhile, power consumers, such as buildings, factories, homes, etc., are stakeholders of the edge side and the comfort of the residents is one of the objective functions of edge optimization [5].

In next-generation mobility services for vehicles, municipalities and road managers are stakeholders on the cloud side and drivers of vehicles are stakeholders on the edge side. As the objective function of the cloud optimization, it is conceivable that the traffic congestion mitigation, traffic accident reduction, CO2 emission reduction, and so on. However, in this paper, relaxation of traffic congestion is regarded as a main objective function of the cloud optimization. Also, the shortest arrival at the destination was defined as a main objective function of the edge optimization as one of the comforts of the driver.

As shown in the simulation result of Figure 3, there is a trade-off relationship between the cloud optimization by the ITS and the edge optimization by the V/C. In this simulation, negotiating between the cloud optimization process and the edge optimization process, we confirmed that it is possible to balance both optimizations in the optimization phase 5. In fact, whether the balanced state of two optimizations is valid depends on the allowable range for each optimization. In the proposed model, we can select various combinations of the cloud optimization and the edge optimization depends on stakeholders' requests. Since the simulation shown in this paper was aimed at verifying the principle of optimization balancing by IoT edge optimization model, the simulation model was simplified. Furthermore, to perform a practical simulation, it is necessary to consider the difference in the moving start time and the moving direction of vehicles in the system. That is, it is assumed that congestion occurs in a node when there are vehicles passing through the node at the same time in this simulation. However, in fact, the case such as to pass each other in the opposite direction and not to be congested at the node shall be considered.

6 CONCLUSION

In this paper, we proposed an IoT edge optimization model that balances the cloud optimization and the edge optimization. Also, as an example of applying the proposed model, we showed simulation results for alleviating traffic congestion in automatic driving. As future work, we will extend the IoT optimization model for the time series optimization.

ACKNOWLEDGEMENT

Part of this work was carried out under the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University.

REFERENCES

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications,” IEEE Communication Surveys & Tutorials, vol.17, no.4, pp. 2347-2376, Jun. 2015.

[2] N. Shiratori, S. Kitagami, T. Suganuma, K. Sugawara, and S. Shimamoto, “Latest Developments of IoT Architecture,” The Institute of Electronics, Information and Communication Engineers, vol.100, no.3, pp.214-221, Mar. 2017 (in Japanese).

[3] P.G. Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere, “Edge-centric Computing: Vision and Challenges,” ACM SIGCOMM Computer Communication Review, vol. 45 issue 5, pp. 37-42, Oct. 2015.

[4] T. Suganuma, T. Oide, S. Kitagami, K. Sugawara, and N. Shiratori, “Multiagent-Based Flexible Edge Computing Architecture for IoT,” IEEE Network, Vol.21, Issue.1, pp. 16-23, Jan. 2018.

[5] T. Ogino, S. Kiatagami, and N. Shiratori, “Flexible IoT Edge Computing Model and Its Applications,” IPSJ SIG ZTechnical Report, vol.2018-cds-21 No.1, Jan.2018 (in Japanese).

[6] Get started with R igraph, http://igraph.org/r/ (accessed 14 May 2018).

Time synchronization method for rice cultivation management system with IoT specialized outdoor communication procedure

Koichi Tanaka†, Mikiko Sode‡, Masakatsu Nishigaki†, Tadanori Mizuno*

†Graduate School of Science and Technology, Shizuoka University, Japan
‡Global Information and Management, International College of Technology, Japan
*Faculty of Information Science, Aichi Institute of Technology, Japan

Abstract - In the field servers for rice fields that use LPWA technology, which require only batteries for their operation, time synchronization is an important factor in reducing the power consumption. Additionally, time synchronization is an important technique for increasing line use efficiency. In this paper, we describe a method of constructing a wireless network of an economical time-synchronized field server using LoRa for achieving low cost and describe the effect of reducing the power consumption. The performance of the time synchronization using GPS is accurate but considering the increase in manufacturing cost because of increase in the cost of the GPS module and the number of parts, we considered that the introduction of GPS is difficult in a field server for usage by a general farmer. Therefore, we propose a time synchronization method without GPS. From experimental results, we confirmed that the time was synchronized and transmission and reception of data between the master and the field server ensued normally. The power consumption of the field server was 108.4mWh per day, and it is theoretically possible to operate it for 691 consecutive days. Continuous operation of the server for 691 days is sufficient for monitoring rice cultivation work**.**

Keywords: Agriculture, Field Server, Sensor Network, Low Battery Consumption, Time Synchronization

1 INTRODUCTION

Owing to the advancing age of older farmers, it is necessary to pass down the knowledge and techniques of farming known to them, to the next generation [1]. Therefore, we are developing a field management system that will help this knowledge transfer [2]. It is reported that LoRa has low power consumption and long communication distance; therefore, it is suitable for communication in the field server for the rice fields [3]. Therefore, we adopted LoRa which does not require communication cost.

The field server needs to operate on the battery for six months, starting from the rice planting to reaping, because that rice fields do not have a power supply because of cost. In addition, it is difficult to install solar panels because solar panels are big and get in the way of farming.

Therefore, low power consumption is important for the field servers for rice fields. To operate for six months with no power supply, the field server needs to turn itself off except when sending or receiving the sensor data or other communication. This requires the intermittent operation communication protocol and the time synchronization method.

The time synchronization technique has been extensively studied previously [4, 5]. The method for the time synchronization using GPS has also been proposed [4]. To use this method, it is necessary to install a GPS receiving module in every field server, which leads to an increase in initial cost at the time of introduction. There is also a problem of increasing power consumption, therefore, it is difficult to use it for the rice field servers, for which lowering the introduction barrier is desired. Although TPSN [5] is proposed as the time synchronization method, it requires a long time for the time synchronization and is therefore difficult to use for the field servers where the power consumption demands are considerably high.

In this study, we propose the intermittent operation communication protocol and the time synchronization method to solve the aforementioned problems. In the proposed method, after the field server system transmits the sensor data to the master unit system, the master unit system on receiving the sensor data transmits the time correction signal to the field server system, thereby performing the time synchronization. In this paper, section 2 describes the system configuration, section 3 discusses the communication protocol and time synchronization, section 4 shows the operational test and result, and finally, section 5 summarizes the study.

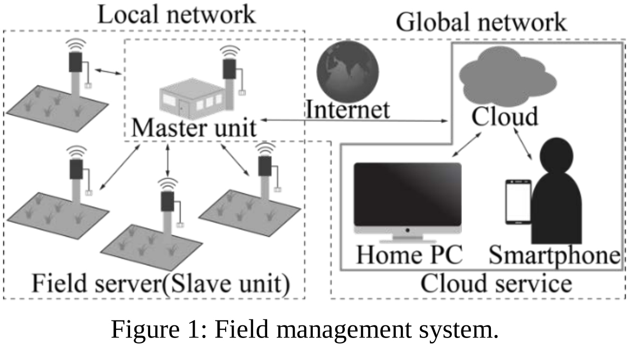


Figure 1: Field management system.

2 SYSTEM CONFIGURATION OF FIELD MANAGEMENT SYSTEM

The field management system is composed of the field server system, master unit system, and cloud service. Figure 1 shows the overall structure of the field management system. The field server system is installed in the rice fields and receives the sensor data. Further, the data is sent to the master unit system through the LoRa wireless network. The master unit system integrates the sensor data from the field server system and sends them to the cloud service through the 3G line or Wi-Fi. The cloud services are provided by the smartphone applications, tablet applications, and web pages. These services provide data to farmers to alert them about water levels, propose a suitable work plan, preserve work records etc.

Communication between the field server system and the master unit system using LoRa is capable of long-distance communication. LoRa has been found to have a practical communication distance of 3,000 to 4,000 m as shown by the basic communication characteristics survey [7]. The rice field of Ishikawa prefecture was assumed, and the linear distance between the parent machine and the field server was within 3,000 m. For this reason, we adopted LoRa, which enables direct communication between the field server and parent machine.

Figure 2 shows the positional relationship between each field and the office of the assumed the agricultural corporation in the Ishikawa prefecture. A, B, C, D, E, F, and G represent the position in each field, where the field servers are installed. P represents the location of the office; the master unit is installed in the office. The linear distances between the field servers A to G and the parent machine P are as follows: 397 m between A and P, 923 m between B and P, 943 m between C and P, 684 m between D and P, 1,150 m between E and P, 1,440 m between F and P, and 1,910 m between G and P.

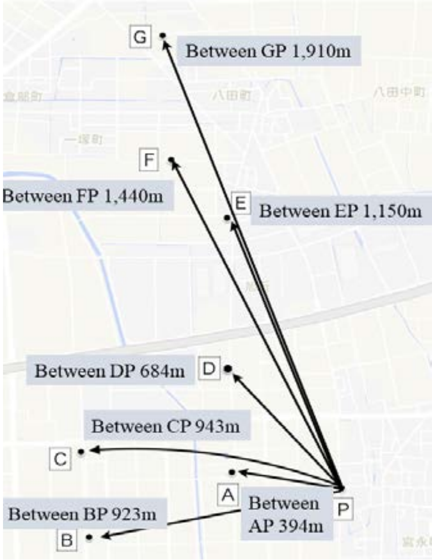


Figure 2: Position of each rice field and the office.

We will further explain the configuration of the field server system installed in the rice field and the master unit system installed in the office. Figure 3 shows the configuration of the field server system. The field server comprises the battery, power ON/OFF circuit, AVR microcomputer, LoRa module, various sensors, and SD card module. The field server is powered by the battery. To realize low power consumption, the power ON/OFF circuit operates only for several tens of seconds in one hour. The wireless modules and the sensors are controlled by the AVR microcomputer. Five types of sensors are mounted to measure the temperature, humidity, water level, soil temperature, and soil moisture content. The sensor data is stored in the SD card together with the time stamp. This is a function for reliably saving the data, considering the case where it cannot be transmitted to the master unit or where the time correction signal cannot be received. The power ON/ OFF circuit is composed of the PIC microcomputer and the FET; it controls power supply to the AVR microcomputer. The PIC microcomputer controls the FET by outputting HIGH/LOW at the GPIO pin. The time required for the power supply control is calculated and controlled by using the timer interrupt in the internal clock of the PIC microcomputer.

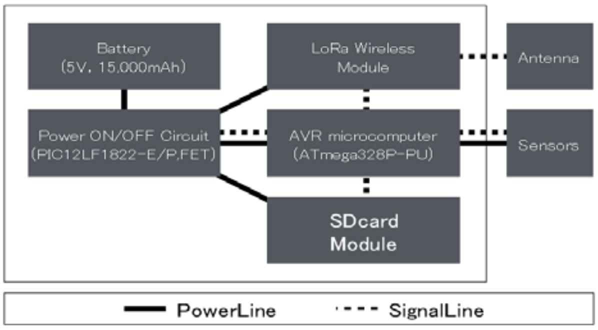


Figure 3: Field server system configuration.

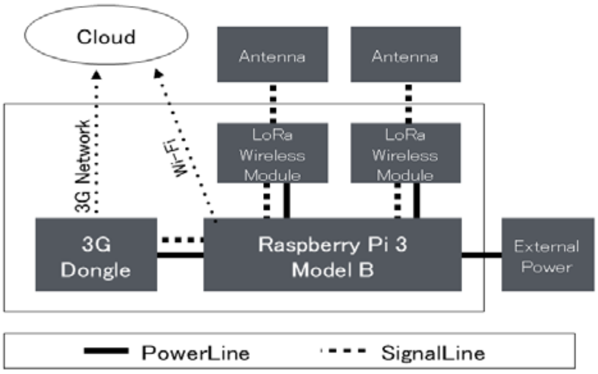


Figure 4: Master unit system configuration.

The configuration of the master unit system is shown in Figure 4. The master unit system is composed of a Raspberry Pi, LoRa module for transmission, LoRa module for reception, and a 3G dongle. When the field server system is turned on for the first time, the time is not held and the time is set after it is transmitted by the master unit system. Therefore, the master unit system always maintains the reception state. It is desirable that the master unit system can respond to communication from the field server system when the field server is installed. Further, the reception and transmission modes exist in the LoRa module, and it takes time to switch the modes. Therefore, by installing two different LoRa modules for receive and transmit, it is possible to reduce the waiting time of transmission and reception and maintain the reception state at all times.

3 COMMUNICATION PROTOCOL

3.1 Communication Protocol

The frame formats used for the communication are shown in Tables 1, 2, and 3. Table 1 shows the common frame format, consisting of the destination, the source, and the payload. Table 2 shows the format of sensor data transmission. Since there are five types of sensors in use, the sensor data are defined in the format of 1 to 5. The temperature, humidity, water level, soil temperature, and soil moisture content are entered in that order from the sensor data 1 to 5. The acquired five types of data can be stored in 2-byte units. Additionally, when the number of types of sensors increases, 2 bytes are added to the format of sensor data transmission. Table 3.3 shows the format of the time correction signal. The time stamp and the time correction signal transmitted from the master to the field server are stored in the payload.

Table 1: Common frame format.

Destination	Source	Payload
1Byte	1Byte	Variable

Table 2: Format of sending sensor data (Payload).

Sensor Data 1	Sensor Data 2	Sensor Data 3	Sensor Data 4	Sensor Data 5
2 Byte	2 Byte	2 Byte	2 Byte	2 Byte

Table 3: Format of correction time signal (Payload).

Timestamp (UNIX Time)	Correction time
4 Byte	2 Byte

The field server system starts once every hour. It gets the sensor data and transmits it to the master unit system. When transmission is successful, the master unit system sends the correction time to the field server system. The field server system further corrects its own internal clock for the time synchronization. Later, when the field server system receives the corrected time or go on operating time per hour of described later elapses, the power is turned off except for the power control circuit.

Figure 5 shows an example of three field servers system in which resending mode does not occur in any of the communications. First, the field server system A (hereinafter, FS-A) is activated. FS-A measures the sensor data and generates the sending packet according to the sensor data. Further, the packet is sent to the master unit system. The field server system has only one LoRa module, therefore it switches from the sending to the reception mode. This switching requires several seconds. After switching to the reception mode, the field server system waits until the set timeout period. Further, it receives the corrected time signal from the master unit system. When the master unit system receives sensor data from FS-A, it sends the corrected time signal to FS-A. FS-A corrects its internal time based on the received correction time signal. After the correction, FS-A goes into sleep mode even during the resending possible time. When the field server system fails to receive the corrected time signal or the master unit system fails to send the corrected time signal to the field server system it is necessary to resend it along while the possible resending time. Field server system B (FS-B) and field server system C (FS-C) perform in the same sequence as A.

The operating time per hour can be obtained from equation (1).

operating time = sensor data acquisition time +
sensor data transmission time +
time correction signal reception time +
retransmission time (1)

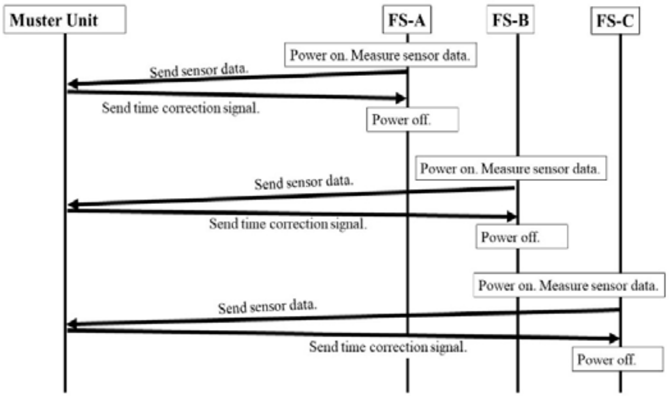

Muster Unit
FS-A
FS-B
FS-C
Power on. Measure sensor data.
Send sensor data.
Send time correction signal.
Power off.
Power on. Measure sensor data.
Send sensor data.
Send time correction signal.
Power off.
Power on. Measure sensor data.
Send sensor data.
Send time correction signal.
Power off.

Figure 5: Communication protocol sequence.

This operating time is written in advance to the AVR microcomputer and sent to the PIC microcomputer each time the AVR microcomputer is powered on. The PIC microcomputer uses this value to calculate the restart time. Here, the sensor data acquisition time is the time to measure the sensor data. The sensor data transmission time is the time to transmit the sensor data to the master unit. The time correction signal reception time is the time to receive the current time from the master unit. The retransmission time is the time to perform retransmission processing when sensor data cannot be transmitted to the master unit. In the proposed method, time synchronization is performed once every hour. It has been confirmed from the measurement results that there is an error of not more than ±10 seconds at the maximum in an hour [3]. Therefore, the error of acquisition time of sensor data is also ±10 seconds or less. This error is a problem-free range as the sensor acquisition time error for agriculture.

3.2 Operation of Resending Mode

When the field server cannot receive the time correction signal and the reception waiting time has elapsed, the field server performs the timeout operation. After the timeout, the field server waits for random seconds from 0.1 to 5.0 s and then retransmits. The following is the cause of the timeout.

1) When the master unit cannot receive the communication from the field server due to the radio wave attenuation and the noise cannot be demodulated

2) When a collision occurs in the transmission data due to overlapping of the transmission times of a plurality of field servers

The first one is that the cause of the noise is often temporary, so there is a high possibility that the problem will be solved if the transmission process is performed with shifted time. The second one can be prevented by accurate time synchronization.

When the noise or the collision occurs, the master unit does not send the time correction signal to the field server, therefore, the field server's reception standby timeout occurs. The field server that has timed out executes retransmission, but to prevent re-collision with the communication performed by the field server of the initial power-on, a random second standby time is provided. After executing the retransmission, the field server switches from the transmission to the reception mode and waits for reception. This operation is continued until the field server can receive the time correction signal from the master unit. However, to avoid collision of the communication with that of other field servers, the power is turned off forcibly when the other field server's operation is about to start.

The sequence operation in this case is shown in figure 6. In figure 6, the field server C (FS-C) has timed out and is retransmitting. If it is within the possible retransmission time, the processing of the transmission and reception standby is repeated until transmission/reception is completed.

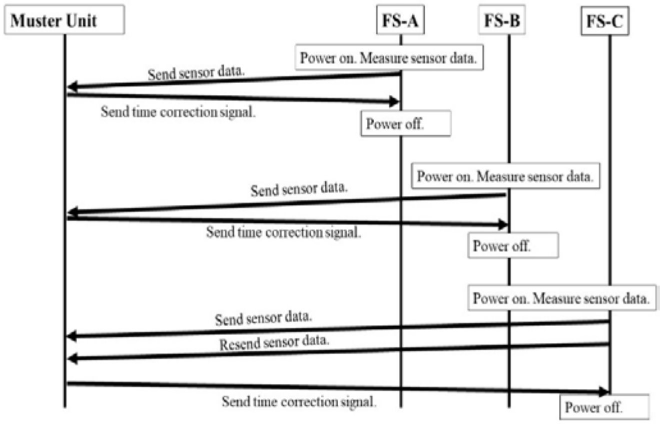

Master Unit
FS-A
FS-B
FS-C
Power on. Measure sensor data.
Send sensor data.
Send time correction signal.
Power off.
Power on. Measure sensor data.
Send sensor data.
Send time correction signal.
Power off.
Power on. Measure sensor data.
Send sensor data.
Resend sensor data.
Send time correction signal.
Power off.

Figure6: Sequence diagram at sensor data collision.

3.3 Time Synchronize Signal

The field server can transmit sensor data at an arbitrary timing because the master unit is always on standby for reception. The master unit performs the time synchronization with the NTP server beforehand and acquires accurate time. The master unit generates the time correction signal according to the time, receives the sensor data from the field server, and then transmits the time correction signal.

The format of time correction signal is shown in Table 3. The time stamp is entered with 4-byte UNIX time. It is used to write the sensor data to the EEPROM or the SD card of the AVR microcomputer. Since the correction time is used for correcting the time within the PIC microcomputer, the time shifted for each field server from the current time is set as 0 to 3599 s in 2 bytes. CRC etc. is used for detecting and correcting communication errors that are not defined in the format because they are added by the LoRa communication module.

Figure 7 shows the mechanism of the time synchronization between the field server and master unit. The master unit sends the correct time obtained by the NTP server to field server in the corrected time format. Upon receiving the correction time, the field server transfers the correction time via the AVR microcomputer to the PIC microcomputer in the power ON/OFF circuit. To prevent the correction time from starting simultaneously with other field servers, the current time is shifted appropriately.

A formula for calculating the start time is shown in equation (2). Regarding equation (2), each field server has a uniquely assigned field server identifier (below), and the correction time to transmit to the field server is obtained by subtracting from the current time. Depending on the power-on time of the field server, the time becomes a negative value; however, in this case, a value of 3,600 is added.

$$t_2 = t_1 - 30 \cdot \mathrm{FSID} \qquad (2)$$

The FSID can be used in the range of 0x00 to 0x77, and the field servers are started in the ascending order of FSID. By using FSID, simultaneous activation of each field server is prevented, and transmission signals of the field server are prevented from colliding. It became possible to transmit once every 30 seconds and about 100 field servers are able to connect one master unit.

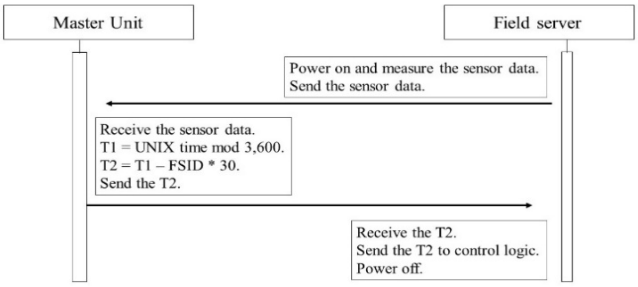


Figure7: Time synchronization mechanism between master unit and field server.

The PIC microcomputer controlling the power ON/OFF circuit always counts 0 to 3,599 s with the internal clock. When the time within the PIC microcomputer reaches 3,600 (0) s, power is supplied to the AVR microcomputer. Upon receiving the time correction signal from the master unit, the field server corrects the time within the PIC microcomputer to the correction time transmitted from the master unit and continues counting. This leads to the time synchronization between the field server and master unit. Even during the second and subsequent runs, when the time within the PIC microcomputer reaches 3,600 (0) s, power is supplied to the AVR microcomputer.

If the master unit fails to normally receive data from the field server due to a communication error etc., the master unit maintains the reception standby state of the sensor data without transmitting the time correction signal.

4 EXPERIMENTAL RESULTS

4.1 Verification of Communication Protocol

We conducted the 7-day operation test to confirm whether the proposed communication protocol works as expected between the master unit and field server. Following are the points for the verification:

1) Confirm whether the master unit can return the time correction signal to the field server within the reception waiting time of the field server.

2) The field server that received the time correction signal confirms whether to shift to the sleep state immediately.

3) Confirm whether each field server properly changes the timing according to the time correction signal and starts at the specified time.

In the verification of the communication protocol, we used seven field servers, which is the same number used in our field. The distance between the field server and the master unit is centered on the master unit and all field servers are installed within a radius of 1 m. The sensor data sent from the field server to the master unit was saved in the verification cloud.

Table 4: Verification results of communication protocol.

Classification	Number of communications
Send sensor data	1,185
Number of resending	9
Completion of time synchronization (resend 1 time)	9

Table 4 shows the verification results of the communication protocol. The number of operational days is seven. The field server gets sensor data at an hourly interval, which is further sent to the master unit. The field server successfully sent the data 1,185 times. Of the 1,185 times, only nine were retransmission. Even when the first communication failed, reception of sensor data was successful from all field servers through retransmission. We confirmed that the protocol works for seven days without problems.

We implemented the designed communication protocols and carried out the operational test for a period of two months in an actual field. Figure 2 shows the measurement result at point C. A master unit was installed at point A. In this experiment, we confirmed that environmental data can be acquired every hour. The field server was equipped with sensors that can measure temperature, humidity, water level, soil temperature, and soil moisture content. The height at which the field server was installed was approximately 1 m from the ground surface to accurately measure the temperature. The height at which the master unit was installed was set to approximately 0.5m. It was confirmed that the measurement can be performed without problems, and data can be transmitted to the master unit. Owing to the fact that there is a communication failure at the rate of approximately 15.8%, the time correction may not be performed. The time synchronization was carried out when the fault was solved and it was confirmed that the protocol was operating properly.

We examined the difference between the assumed startup time of the field server and the actual startup time. The results are shown in figure 8. The result displays the representative pattern of eight days from the operational test of two months. From this result, it is understood that when the time correction is performed, the error is suppressed to about in tens of seconds. Moreover, it is understood that the error is suppressed to 0 s in most communication between the master unit to field servers.

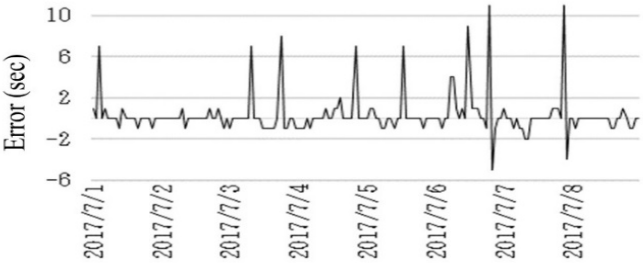


Figure 8: Time error of field server.

The time error of the PIC microcomputer is the maximum at 6.512 s in an hour from actual measurement [6]. This error is accumulated without time synchronization; however, in the field management communication protocol proposed here, this error is within the range in which collision with other field servers does not occur. Therefore, it is confirmed that time synchronization is effective in this communication protocol, and it is possible to reduce the increase in time error, which is proportional to the usage time. As a result, it became possible to transmit once every 30 seconds, and became possible to connect about 100 field servers to one master unit.

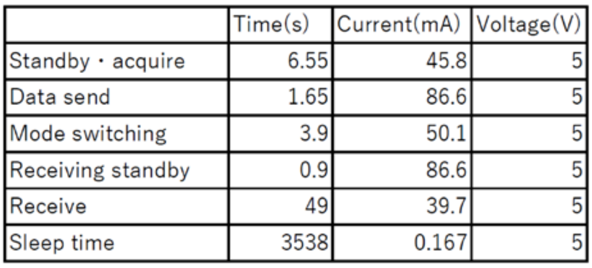
4.2 Evaluation of Power Consumption

Apart from the verification of the communication protocol, we conducted an experiment to verify the power consumption. The purpose of the verification is to obtain the power consumption during the operation. First, we measured the voltage, current value, and processing time for each operation mode. Table 5 shows the measured results [7]. The data transmission mode is the most power consuming. It can be confirmed that the sleep time mode has the lowest power consumption among all.

Next, we calculated the power consumption and number of working days. Equation (3) shows the power consumption W [mWh] . Here, V_1 is the rated voltage [V] of the field server system. I_a is the electric current [mA] during the sensor stabilization standby and the sensor acquisition. t_1 is the time[s] during the sensor stabilization standby and the sensor acquisition. I_b is the electric current [mA] during the transmission of sensor data. t_2 is the electric time [s] during the transmission of sensor data. I_c is the electric current [mA] during the mode switching. t_3 is the time [s] during the mode switching. I_d is the electric current [mA] during the standby reception. t_4 is the electric current [mA] during the standby reception. I_e is the electric current [mA] during the data reception. t_5 is the time[s] the during the data reception. I_g is the electric current [mA] during the system sleep state.

$$W = (V_1\{(I_a \cdot t_1) + (I_b \cdot t_2) + (I_c \cdot t_3) + (I_d \cdot t_4) + (I_e \cdot t_5)\} + V_1 \cdot I_g\{3600 - (t_1 + t_2 + t_3 + t_4 + t_5)\} \ ) \ / \ 3600 \qquad (3)$$

Table 5: Measurement results of power consumption.



	Time(s)	Current(mA)	Voltage(V)
Standby・acquire	6.55	45.8	5
Data send	1.65	86.6	5
Mode switching	3.9	50.1	5
Receiving standby	0.9	86.6	5
Receive	49	39.7	5
Sleep time	3538	0.167	5

We derived the number of operating days theoretically. The power consumption is 4.52 mWh per hour; the consumption being 108.4mWh per day. Therefore, theoretically, the field server can operate for approximately 691 days, assuming the electric quantity of the portable battery charger to be 75000 mWh. Although the number of operating days has the theoretical value, it seems that the field server is able to operate for six months, which is the requirement of the agricultural corporation.

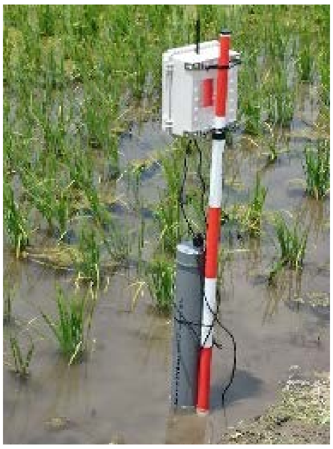


Figure 9: A field server system in a rice field.

We conducted the operational test in actual rice fields using the 7 field servers of figure 2. The picture of the field server system installed in the rice field is shown in figure 9. We confirmed that the field server system works correctly from rice planting to rice reaping.

In IEEE 802.15.4e [8], two types of time synchronization methods, Beacon and Channel Hopping are defined. In the time synchronization defined in both methods, it is required that all nodes belonging to the network always synchronize the time within an error of ±1 ms, thereby realizing the time division access method. On the other hand, in the proposed method, time synchronization is performed between the master unit and each field server, but time synchronization between the field servers is not performed. Therefore, time synchronization accuracy of about ±10 seconds is sufficient, it is not necessary to hold hardware for special time synchronization and it is easy to put into practical use.

5 CONCLUSION

We proposed a new communication protocol, constructed a local wireless network, and conducted the experiment. In the field servers for the rice field using the LPWA technology, which require only batteries for operation, the proposed time synchronization is an important technology for the purpose of reducing the power consumption. Additionally, the proposed time synchronization is an important technique for increasing the line use efficiency. It was seen from the experimental results that the power consumption of the field server is 108.4mWh per day. Therefore, it was confirmed that the method can continuously work for 691 days based on our calculations. The time synchronization is effective and was able to decrease the timing error in direct proportion to the operating time. This protocol is valid for the rice cultivation management systems because the field server is stable and can operate for a long time. Therefore, it meets farmers' expectation to utilize a reasonable field server.

REFERENCES

[1] Ministry of Agriculture, Forestry and Fisheries of Japan, Statistics of agricultural labor, Accessed on 2017-6-2. [Online]. Available: http://www.maff.go.jp/j/tokei/sihyo/data/08.html.

[2] Kiyokazu Kurosawa, Isamu Iizima, Yoshiki Amemiya, Shunya Yamamichi, Masaharu Toyota and Mikiko Sode Tanaka, “Development of Operational Control System for Rice Cultivation Equipped with Activity History Function”, IEICE Technical Report, vol. 116, no. 346, CS2016-55, pp. 59-64 (2016).

[3] Yuta Kawakami, Masaharu Toyota, Keitaro Terada, Keiko Matsumoto and Mikiko Sode Tanaka, “A study of the optimal agricultural field communication using Sub-GHz wireless technology”, IEICE Technical Report, vol. 116, no. 382, NS2016-138, pp. 107-112.

[4] Hao Guo and Peter Crossley “Design of a Time Synchronization System Based on GPS and IEEE 1588 for Transmission Substations”, IEEE Transactions on power delivery, vol. 32, no. 4, (2017).

[5] S. Ganeriwal, R. Kumar and M. B. Srivastava, “Timing-sync Protocol for Sensor Networks”, Proceedings of the 1st ACM Conference on Embedded Network Sensor Systems (SenSys'03), Los Angeles, California (2003).

[6] Keitaro Terada, Masaharu Toyota, Tadaaki Hirata, Yuya Takada, Keiko Matsumoto and Mikiko Sode Tanaka, “Proposal of communication protocol for field management using LoRa”, DICOMO2017, pp. 1671-1678, 2017/6/28-30.

[7] Yumeto Kojima and Mikiko Sode Tanaka, “Current value measuring device for field server of field management using LoRa”, IEICE Society Conference 2018, 2018/9/11-14.

[8] IEEE Standard for Local and metropolitan area networks--Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC sublayer [Online]. Available: https://ieeexplore.ieee.org/document/6185525/.

Proposal of IoT System with SmartContract on BlockChain

Tetsuo Furuichi [*], Hiroshi Mineno[**]

[*]e-Cloud Computing&Co. / Graduate School of Informatics, Shizuoka University, Japan
furuichi.tetsuo.15@shizuoka.ac.jp
[**]Faculty of Informatics, Shizuoka University, Japan
mineno@inf.shizuoka.ac.jp

Abstract – In recent years, the demand for information security for the IoT system has been increasing. For information security, we made an IoT system using BlockChains and evaluated the performance as an IoT system in several types of node connection. And we propose the configuration method with the best performance. Specifically, by using SmartContract which is one function of the BlockChain as IoT Gateway, we implemented the authentication of IoT module and data transfer function, realizing the integrity and availability of information security.

Also, in systems using BlockChains already in practical use, the overhead associated with data transfer delay and mining cycle time is disadvantageous, but from the evaluation results, we have developed a realistic IoT system configuration using BlockChains, and propose the requirements of BlockChains specialized for IoT.

Keywords: IoT, BlockChain, SmartContract

1 INTRODUCTION

The Internet of Things (IoT) has already been used for factory and industrial products, and effects in agriculture have also been reported. Recently, it has begun to be widely used as consumer products targeting home and healthcare. By connecting an embedded device to a server via a network, it has become possible to acquire, accumulate and analyze information that has not been used before. This information is useful for improving our living environment and revitalizing industrial activities and further development is expected in the future [1]. However, due to the rapidly widely used IoT technology, some problems have arisen. In this section, we explain situations where the IoT request is becoming complicated due to diversification, and the problem that the IoT device is the next attack target of PCs and smartphones in information security. We introduce an approach using BlockChain technology which is one of the countermeasures for these problems.

1.1 Component structure of IoT module

Many current IoT modules consist of simple parts of sensors and microcontrollers for lower price and lower power consumption. In recent years, with the development of smart phones and tablet mobile devices, SoC (System on Chip) capable of running Linux has become a practical price, and modules using them can be used as IoT devices. These IoT oriented SoC have the merit of being able to execute advanced processing with high load which could not be done on the edge side so far. However, specifications have become complicated, and development time have been increasing. Many IoT frameworks have been used as methods to decrease those problems.

1.2 IoT security threat

As the number of IoT devices is increases, many information security incidents including leakage and tampering of IoT data utilizing the vulnerability of these devices are reported much. In addition, there are various attack methods using the vulnerability of the IoT system [2], not only functions and performance but also information security measures including attack detection function and update function are important elements of IoT system development. In the future, from the beginning of the design it is necessary to practice "security by design" method to counter information security.

1.3 BlockChain technology

Encryption technology has evolved and abundant hardware capable of executing cryptographic algorithms at practical speed has become commonly available on PCs. By implementing these cryptographic techniques, BlockChains that realized a distributed ledger system, which is one of the distributed processing technologies, have received attention in recent years. In the BlockChain, we have a mechanism whereby mutual checking and sharing of each transaction in the BlockChain cannot tamper with transaction information without centralization. Many virtual currencies using BlockChains have already been announced. Besides using BlockChains in the finance field, application to primary industries including agriculture is also being considered. Food traceability in the agricultural sector is expected to utilize BlockChains, and some consortium activities have already been started overseas as well. Tamper-proof reliable information such as the place of production, producer, production process and distribution route of agricultural products is useful information for consumers as well. In particular, we focused attention on functions that cannot be tampered with as an information security system of BlockChains.

1.4 Application of BlockChain to IoT field

As requirements items for the IoT system, functions, performance, information security can be assumed. Performance depends on the type of IoT, and it is necessary to consider the latency time of data acquisition and data bandwidth. Also, at the same time, the IoT system needs to take measures against information security.

In recent years, IoT is also being used for agricultural products production described in the previous chapter, and data reliability is desired which cannot be tampered with by other persons or even operators even in history information related to IoT.

Information security measures against an unspecified large number of attacks can be implemented in various ways. However, in a conventional IoT system, it is difficult to implement a tamperproof mechanism including an operator. In the prevention of tampering with all subjects, BlockChain technology is one of effective means for ensuring completeness. In addition, although many IoT configuration modules are composed of sensors, IoT Gateways, and cloud servers, each module may cause malfunction individually for each reason, and availability measures of each module are also important. In particular, since the IoT Gateway constituted by the SmartContract of the BlockChain does not limit the hardware to be executed, it is possible to realize not only the function of preventing tampering inherent in the BlockChain but also availability.

The confidentiality preservation of the IoT data on the communication line is a function necessary to prevent data leakage. Because confidentiality cannot be secured in the BlockChain, it is necessary to respond in another way.

2 PREVIOUS RESEARCH AND TASK

We describe IoT technology and BlockChain technology in this section.

2.1 IoT

The growth of smartphones has reduced the cost of sensors and the cost of infrastructure has declined due to the development of cloud computing. And due to promotion from German Industrie 4.0 and various countries including Japan, the IoT market is further expanding. IoT, which has become popular, has begun to create new value by connecting everything including people and things to the network [1]. In particular, the IoT fields are diverse. 1) Health / medical monitoring, 2) Machine monitoring, 3) Natural / environmental monitoring, 4) Human motion detection, 5) Machine motion detection / control, etc. There are various required real-time performances depending on the purpose.

The IoT system consists of 1) Sensor/actuator, 2) Pre-processing, 3) Communication, 4) Accumulation, 5) Analysis, 6) Inference, etc. The standard IoT system was specialized in sensors, communication. For low power consumption, miniaturization, low cost and high reliability, these are configured using one chip microcontrollers, and in many cases, they are implemented without OS or with RTOS (Real Time OS). The SoC having low power consumption and high performance can be used at low cost, and the advanced IoT system running Linux is used. As the system becomes more sophisticated, it has become possible to execute advanced processing that is executed on the cloud server side and that requires additional processor load. Concerning information security for the IoT system, confidentiality, integrity, and availability, which are generally said three important subjects, are important. In order to deal with the malware for IoT mentioned earlier, securing access rights, prevention of tampering, and response to emergencies are becoming important.

2.2 BlockChain

The BlockChain is a system capable of tracing all transaction histories by a distributed consensus building mechanism with network participants. So far, many kinds of virtual currency using BlockChain are distributed. Here we introduce Bitcoin, the most famous BlockChain, and Ethereum, a BlockChain system that can be applied to applications.

Bitcoin was developed based on the BlockChain technology posted by a person named Satoshi Nakamoto in 2008. It started operation in 2009, and is a famous BlockChain for virtual currency [3]. Bitcoin is a system composed of a BlockChain node called Bitcoin client and a Bitcoin network. The transaction information issued by the Bitcoin client is sent as a transaction to the Bitcoin network, and minor, which is a kind of Bitcoin client, miners, so that the block is generated. And that block is approved from multiple nodes of the Bitcoin network.

Ethereum, proposed by Viralik Buterin's white paper in November 2013, is a BlockChain and makes it possible to build applications by SmartContract [4]. While Bitcoin is specialized in moving ownership of cryptographic currency, Ethereum is characterized by being able to create and execute distributed applications called SmartContracts as well as moving cryptographic currencies [5] [6].

A BlockChain can be said a distributed database that realizes a "distributed ledger" that distributes and manages transactions as exchange information on a distributed network. We manage and operate a list of sequential data called "blocks" that summarizes those transactions on multiple nodes. Moreover, the validity of the block is secured by the mining processing using the distributed consensus algorithm. PoW (Proof-of-Work) is mainstream in the current distributed consensus algorithm. Under the agreement of the configuration node of the network, difficulty values are set and have a mechanism to adjust the mining time. In addition to PoW, a distributed consensus algorithm is studying methods that do not spend processor resources or power called PoS (Proof-of-Stake) or PoI (Proof-of-Importance).

Whereas Bitcoin has a mechanism specialized for virtual currency trading, there is a BlockChain that can handle SmartContracts, which is a type of program shared on the BlockChain, as well as virtual currency transactions. Ethereum is one of them, and each node can access virtual currency transaction, mainly to execute virtual program.

2.3 IoT + BlockChain

Focusing on the convenience of distributed management of BlockChain and the characteristics of the virtual currency, the degree of expectation for adaptation to IoT is increasing. In the field of electric power systems, there are cases where BlockChain efficiently perform IoT updates on an ongoing basis [7]. To cope with IoT, a mechanism is developed to cover a BlockChain client program with a wrapper, and by using a network different from the BlockChain, a weak data transfer of the BlockChain is handled [8]. According to research to use IoT in Smart Home, the merit of information security is larger than the overhead of processing BlockChains [9]. In order to easily manage the configuration of the IoT device as a research concretely using SmartContract, an IoT system that has the mechanism of RSA key management in Ethereum's SmartContract has been reported [10].

3 ABOUT THE DEVELOPED IOT SYSTEM

In the IoT system, a registration of IoT devices and delivery of data in a specific network is an important function. In this research, we focused on the information security of the BlockChain and used geth which is Ethereum's BlockChain client program of virtual currency which has already been put into practical use. We worked with geth, developed an API with the basic function of the IoT module interface, and implemented it as a BlockChain IoT module in the embedded system.

3.1 BlockChain IoT module

Requirements for a practical IoT system that meets this condition are flexibility, connectivity, extensibility and security, for example. Then, the following five points were defined as basic requirements. 1) use of general-purpose network, 2) use of general-purpose hardware and OS, 3) pre-registration of connectable sensor modules, 4) prompt response at failure, abnormality, 5) secure access to the network. In particular, in consideration of requirements of 3) to 5), this prototype system was constructed as a SmartContract for management using virtual currency in the BlockChain.

This hardware is based on embedded Linux. The basic specifications are ARM-CPU, 1 GB of the main memory and 16 GB or more of the external storage, with network function and sensor interface. This time, we prepared a module based on Raspberry Pi 3b.

Ethereum's geth program running with this module is implemented in Go language and executes on Linux / Unix environment. The system is connected via a TCP/IP based network, and its physical communication method may be wired or wireless. Each device (hereinafter "Node") needs to assign a unique IP address. Node has made it possible to run JavaScript API and application script which has interface of geth program and sensor and communication function with external cloud server. And handling of sensors and communication authority to an external network was set as a separate policy from Node-to-Node connection using BlockChain, and flexibility of connectivity as an IoT device was secured.

The connection with the BlockChain is the jurisdiction of the upper hierarchy without hardware dependency. Therefore, when the corresponding module fails, even if the alternative hardware has different performance, it can connect to the BlockChain.

3.2 Sensor data flow

Each Node having a unique IP address forms a BlockChain by mutually connecting with the geth program executed on Node. Each node has an address that is an account for accessing the BlockChain, and communication among Nodes is performed among addresses as transactions. This system has IoT management API function to SmartContract which is positioned as middleware existing in the BlockChain and to operate as the IoT system throughout the BlockChain network. **Figure 1** shows the flow of processing of one example of the system using the BlockChain IoT module developed this time. Data is sent from a node having a sensor (Sensor Node) to a cloud server via a BlockChain network via a node having a gateway function (Gateway Node).

Specifically, at Sensor Node, 1) IoT Application reads information from the connected sensor via Sensor Library. 2) The IoT Application passes the data to the SmartContract on the BlockChain via the BlockChain Client (geth) and the network. 3) Upon receiving the data, the SmartContract sends an event to the IoT Application on the GateWay Node side via the network of GateWay Node and the BlockChain Client (geth). 4) The IoT Application on the GateWay Node side reads the sensor data stored from the SmartContract, and 5) sends the data to the cloud server using the IoT Library.

Characteristically, in this connection, the Sensor Node and the Gateway Node are able to exchange information with the unique address of the BlockChain to be connected and the information of only the address of the SmartContract to be connected. That is, both Nodes can communicate even if they do not know IP address each other. And the SmartContract is virtually executed by hardware to be mined. In this example, Mining Node becomes the execution machine of SmartContract. At least one Mining Node is required.

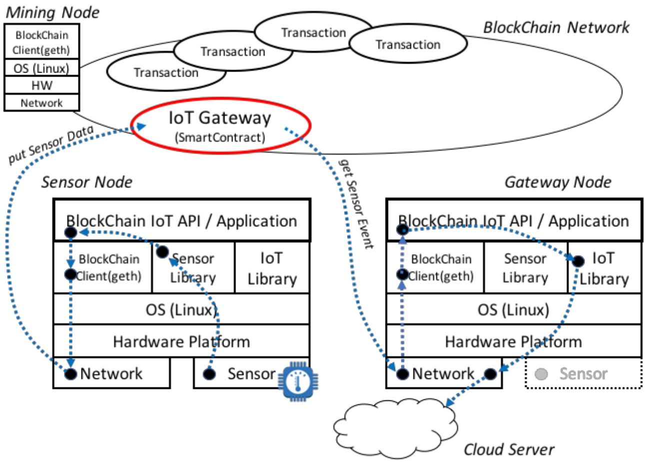


Figure 1: Data flow between SmartContract and node

3.3 IoT Gateway by SmartContract

In this system, IoT module's authentication function and data delivery function are implemented in SmartContract of the BlockChain, and this method is different implementation from the IoT device interface so far.

By connecting to the BlockChain via BlockChain client software, the account address becomes the mutually approved unique ID. Then, the data exchanged by that account is handled as a transaction, and it is handled in the same way as signing with the account address. In consequence, the interface in the BlockChain can ensure the integrity of information security.

This IoT Gateway implemented in the SmartContract is a distributed management API for connecting with the virtual network for IoT, and it is registered and authenticated in the BlockChain as a transaction in advance. And the BlockChain information to be connected is preset in the BlockChain IoT module developed this time. This module can physically connect to the TCP/IP network and join a configured BlockChain to synchronize with the active BlockChain and to communicate with its registered SmartContract.

As shown in Figure 2, this IoT Gateway roughly has two functions. The first one is a sensor management function and the second one is a sensor data delivery function. The IoT Gateway is registered as a transaction in the BlockChain with the preset Owner Node account address. The account address of that Owner Node will be the privileged user of the registered the IoT Gateway and will have administrator privileges. And sensor registration is executed with its Owner Node account address.

(1) Registration of Sensor Node

One of the IoT management functions is registration of connectable Sensor Nodes. This setting is proceeded by Owner, the SmartContract creator. Owner Node sets the account address of a connectable Sensor Node as an argument of the registration function and registers with the IoT Gateway by generating a transaction.

(2) Activation of Sensor Node

The Sensor Node registered in the IoT Gateway uses his account address to access the IoT Gateway and activates its own sensor data channel. Activation is programmed to spend virtual currency as fees. Then it is possible to set the number of data that can be buffered and set the fees for sensor data exchanges.

(3) Sending sensor data

The virtual currency is also required when Sensor Node registers data received from sensor device to IoT Gateway as fees. This specification can expect the effect of suppressing unnecessary data registration to the IoT system. Since we are assuming small capacity data delivery this time, we decided to treat direct data as a transaction in the BlockChain. When the sensor outputs a large amount of data, since the load of the transaction in the BlockChain is large, it is necessary to have a separate data delivery mechanism.

In this implementation, when sensor data from Sensor Node is stored in a preset number of buffers in the IoT Gateway, this IoT Gateway asserts an event to notice outside. The Gateway Node is set to receive the event via the geth program and receives data from the SmartContract when it receives the event. This access requests fees of the virtual currency of the amount set beforehand by the Sensor Node at the time of activation, and this setting can be expected to prevent unscrupulous data access without permission.

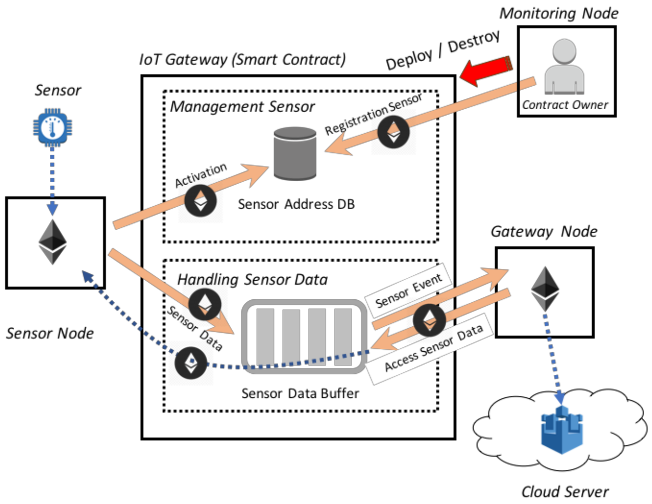


Figure 2: IoT Gateway by SmartContract

4 PROTOTYPE IMPLEMENTATION AND EVALUATION

To clarify the usage conditions of the configuration model discussed in Chapter 3, we combined the IoT system with the existing BlockChain and the evaluated characteristics as a system. Especially we focused on the sensor data transfer latency time from a viewpoint of one IoT system, and evaluated the dependency relationship with the logical network configuration and the number of mining nodes.

4.1 Evaluation environment and method

Figure 3 shows the configuration of the environment used for this evaluation. Ethereum Private Net was constructed by implementing Ethereum 's client program geth on the server' s container, Note PC and IoT module. Furthermore, IoT SmartContract, which is an IoT API, is implemented in the BlockChain net and it was created in advance as a transaction in the BlockChain.

Three types of Nodes were prepared for evaluation. The first one is a general Node which is a contract Owner. The second one is a Sensor Node that uploads sensor data to the BlockChain. The third one is a Gateway Node that takes sensor data from the BlockChain. Each node assigns an account address of the BlockChain.

The evaluation script sends test data to the SmartContract and measures the time the event returns from the SmartContract. These sending and measurement are executed for the specified number of times. Mining Node for blocking transactions in the BlockChain is a Node running on a PC or a container connected to the network. And the execution instruction of the Mining process was manually performed.

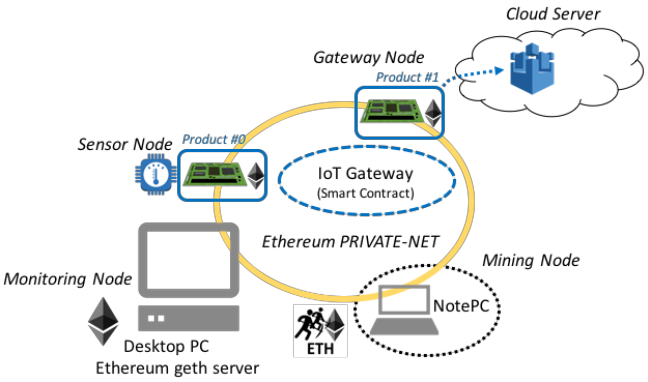


Figure 3: Overview of Logical nodes connection

4.2 Evaluation contents and results

In this research, two kinds of data transfer latency times were evaluated.

(1) Data transfer latency time by physical connection

This evaluation measures the data transfer latency time in "Node - SmartContract - Node" by changing the physical connection form of Node in the BlockChain. We made 100 data accesses in each physical connection form.

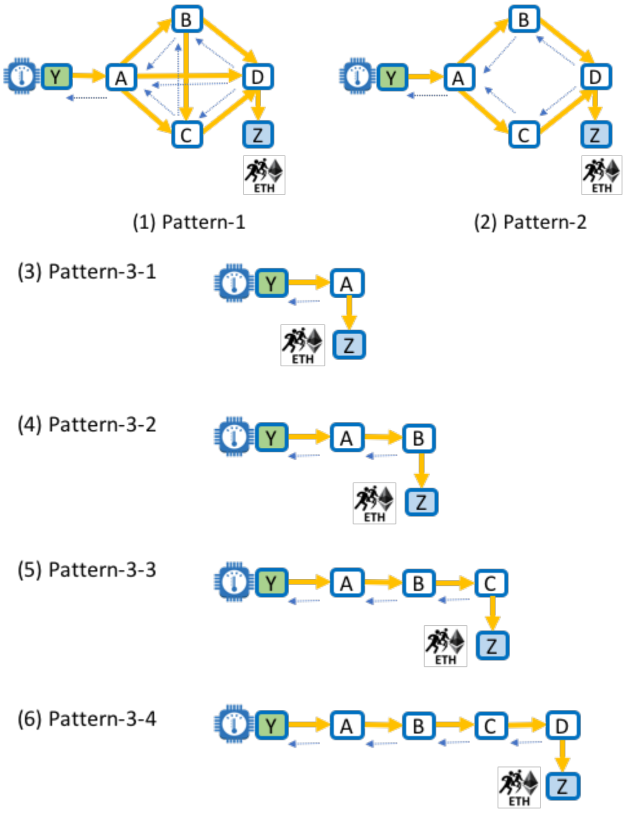
Six types of physical connection were prepared. Figure 4 shows these connection modes. 'Y' in the green box is a Node with sensor data. 'Z' in the blue box is Mining Node. Pattern-1 indicates that the constituent Nodes A to D are mutually connected. In Pattern-2, configuration nodes A - D are connected in an annular shape. Pattern-3-1 to Pattern-3-4 are connection embodiments in which the number of Mining Nodes is changed from Node having sensor data. The yellow arrows indicate the expected direction of propagation of the issued transaction. The blue dotted arrows indicate the flow of the transaction after Mining.

Figure 5 shows the latency time results for each connection pattern. There was no significant difference between the minimum value and the average value of the latency time in any of the physical connection patterns, but as for the maximum value, the deflection becomes larger as the physical distance becomes farther.

(2) Data transfer latency time by mining number

Next, we measured the latency time when changing the number of mining nodes with the same net structure. Figure 6 shows the pattern of the BlockChain network evaluated. As well as the evaluation of (1), the latency time of data transfer was measured with a pattern in which Mining Node was increased from 1 to 4 patterns.

Figure 7 shows the latency time in each mining pattern. The variance of the latency time when the mining number = 1 is wide, but as the mining number increases, the variation of the latency time is contained.



(1) Pattern-1

(2) Pattern-2

(3) Pattern-3-1

(4) Pattern-3-2

(5) Pattern-3-3

(6) Pattern-3-4

Figure 4: Various Connection Patterns with 1-Miner

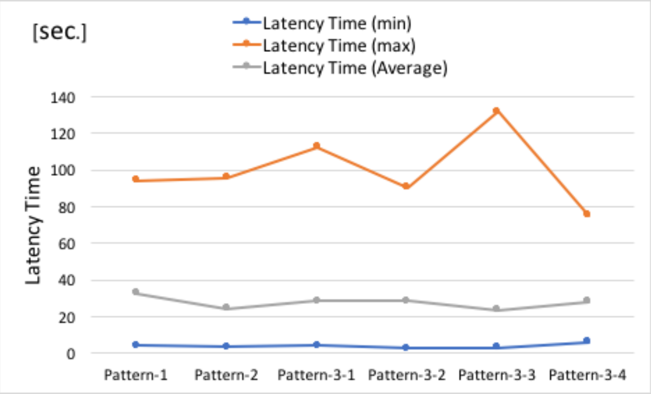
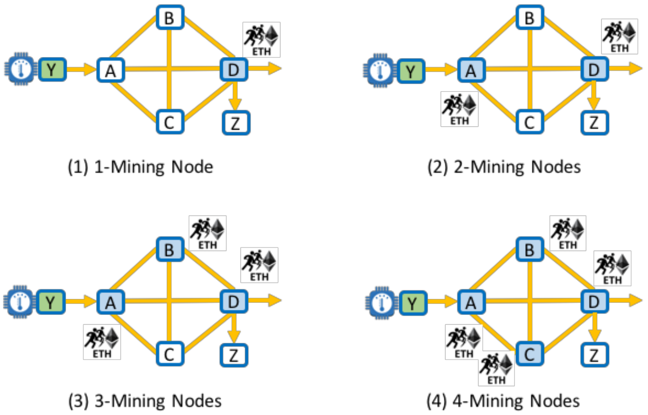


Figure 5: Latency time of each Pattern with 1-Miner



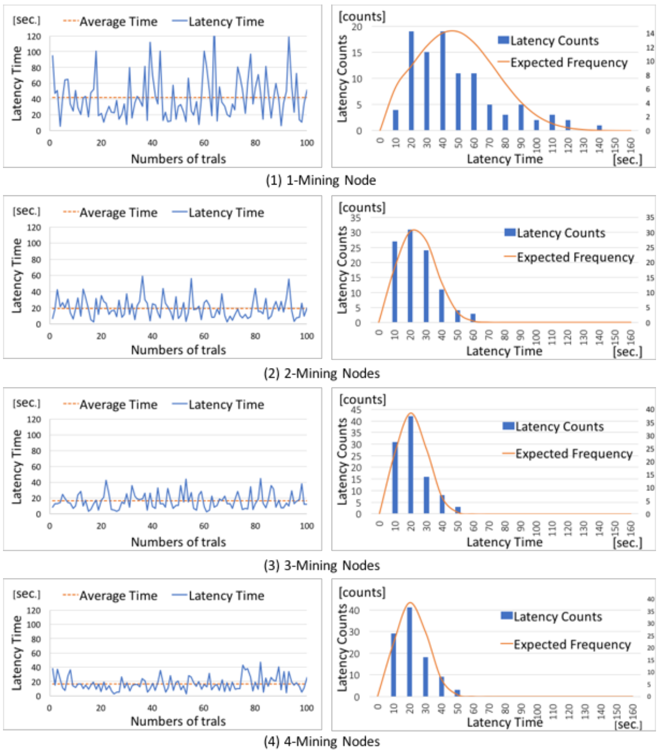
(1) 1-Mining Node

(2) 2-Mining Nodes

(3) 3-Mining Nodes

(4) 4-Mining Nodes

Figure 6: Various patterns of Mining Nodes



(1) 1-Mining Node

(2) 2-Mining Nodes

(3) 3-Mining Nodes

(4) 4-Mining Nodes

Figure 7: Latency time on various Mining Patterns

5 CONCLUSION

In this prototype, by implementing the function of IoT Gateway on the SmartContract of the BlockChain, we were able to operate normally as an IoT system. However, we found that the latency time of IoT data transmission may take more than one minute. The result of the evaluation shows that the scope of application of this IoT system is not general purpose and limited. For example, this system is suitable for monitoring natural environment, health, etc., which does not expect real-time reaction, and it is not suitable for machine control and warning detection expecting quick response. However, since the propagated data inherits the basic function of the BlockChain, no one can tamper with it.

Moreover, when using existing BlockChain as the IoT system, each node has a logical connection link with plural nodes, a stable mining cycle can be realized by preparing a plurality of mining nodes, and It is possible to reduce variations in latency time of data propagation. Although this evaluation used Ethereum, it is BlockChain which is used for ordinary virtual currency transactions, and it is a system assuming operation at PC and server, but it was able to operate comfortably in IoT system as well. However, the computer resources to use seems to be slightly heavy in the current IoT system, but future hardware evolution will reduce those problems.

There are improvements in mining and SmartContract implementation when assuming BlockChains specialized for IoT. In mining, control of the mining cycle is necessary to reduce the power consumption of the IoT system. With this function, if data transmission does not come for a while, mining can be stopped. In addition, granting of mining node privileges is necessary to stop illegal acquisition of virtual currency.

For SmartContracts, there are improvements in the method of calling it. In the current BlockChain, the SmartContract is called with an address indicated by a number. However, due to a malfunction or improvement of the SmartContract itself, when version upgrade is done, it is necessary to change the address of the new SmartContract in some way. In order to solve this problem, it is necessary to call a SmartContract name or a secure delivery method of a SmartContract address.

By using the BlockChain, although overhead of the BlockChain itself occurs, it is thought that the feature that transaction cannot be tampered can be effectively used in the IoT system as a method of "Security by Design".

REFERENCES

[1] Dave Evans, The Internet of Everything How More Relevant and Valuable Connections Will Change the World, Cisco IBSG, (2012).

[2] Afreen Fatima Mohammed, Security Issues in IoT, IJSRSET Volume 3 Issue 8, http://ijsrset.com/paper/3369.pdf, (2017).

[3] Satoshi Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, https://bitcoin.org/bitcoin.pdf, (2008).

[4] Vitalik Buterin, A NEXT GENERATION SMART CONTRACT & DECENTRALIZED APPLICATION PLATFORM, http://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf, white paper, (2014).

[5] Ethereum home page, https://www.ethereum.org/ , (Accessed on 06/03/2018)

[6] Ethereum Homestead Documentation, http://www.ethdocs.org/en/latest/ , (Accessed on 06/03/2018)

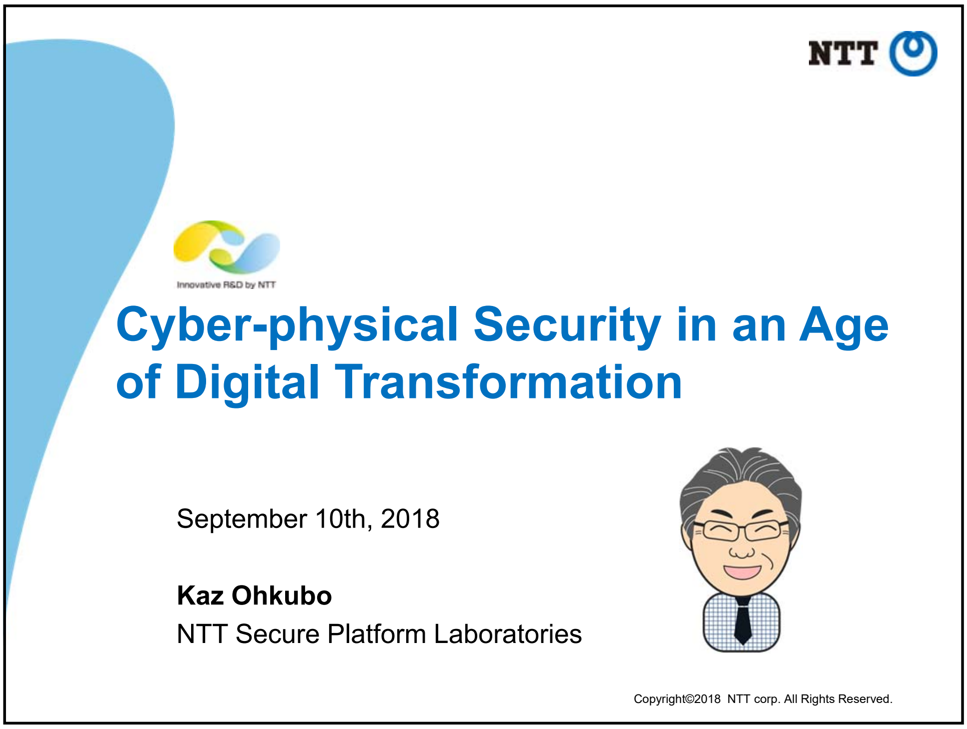
[7] K. Christidis and M. Devetsikiotis, Blockchains and smart contracts for the internet of things, IEEE Access, 4:2292--2303 (2016).

[8] M. A. Walker, A. Dubey, A. Laszka, and D. C. Schmidt, PlaTIBART: a platform for transactive IoT blockchain applications with repeatable testing, in 4th Workshop on Middleware and Applications for the IoT (M4IoT), December 2017.

[9] Ali Dorri, Salil S Kanhere, Raja Jurdak, and Praveen Gauravaram, Blockchain for IoT security and privacy: The case study of a smart home, In Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE International Conference, 618‑623 (2017).

[10] S. Huh, S. Cho, S. Kim, Managing IoT devices using blockchain platform, 2017 19th International Conference on Advanced Communication Technology (ICACT), pp. 464-467 (2017).

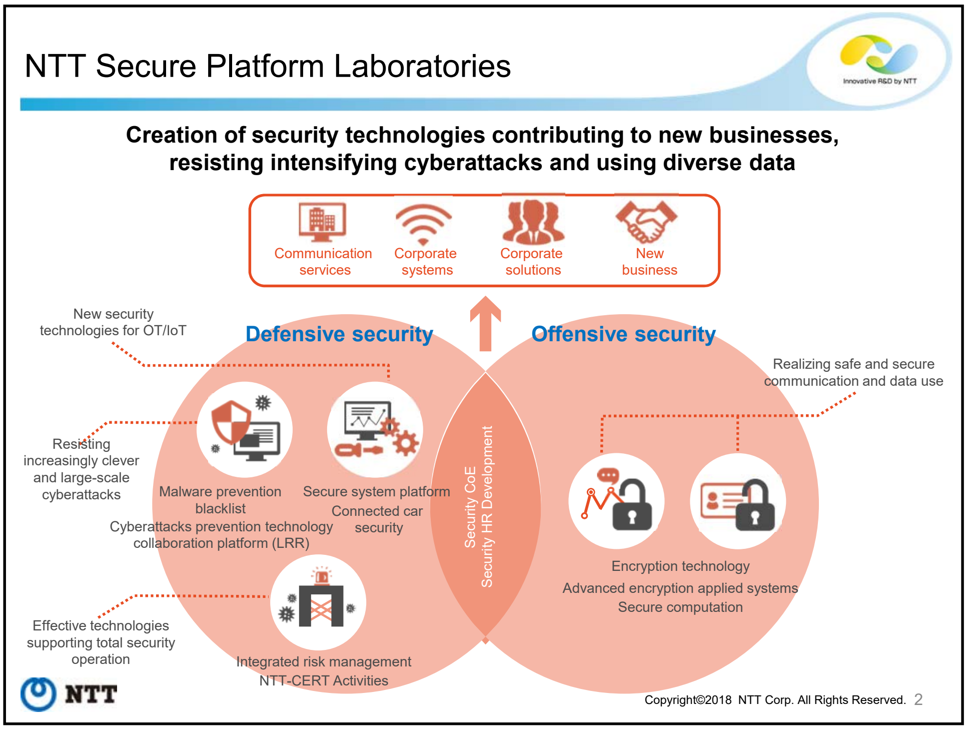
Keynote Speech 1: Mr. Kazuhiko Ohkubo (Vice President / Head of NTT Secure Platform Laboratories)



Cyber-physical Security in an Age of Digital Transformation

September 10th, 2018

Kaz Ohkubo
NTT Secure Platform Laboratories



NTT Secure Platform Laboratories

Creation of security technologies contributing to new businesses, resisting intensifying cyberattacks and using diverse data

Communication services
Corporate systems
Corporate solutions
New business

New security technologies for OT/IoT

Defensive security

Offensive security

Realizing safe and secure communication and data use

Resisting increasingly clever and large-scale cyberattacks

Malware prevention blacklist
Cyberattacks prevention technology collaboration platform (LRR)

Secure system platform
Connected car security

Security CoE
Security HR Development

Encryption technology
Advanced encryption applied systems
Secure computation

Effective technologies supporting total security operation

Integrated risk management
NTT-CERT Activities

2

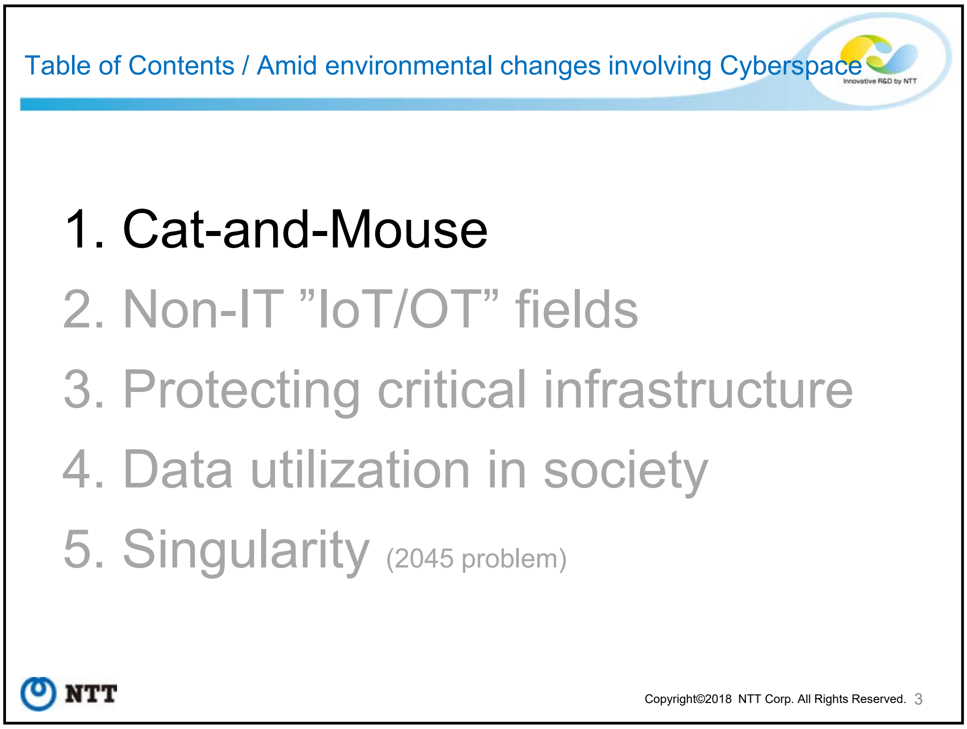
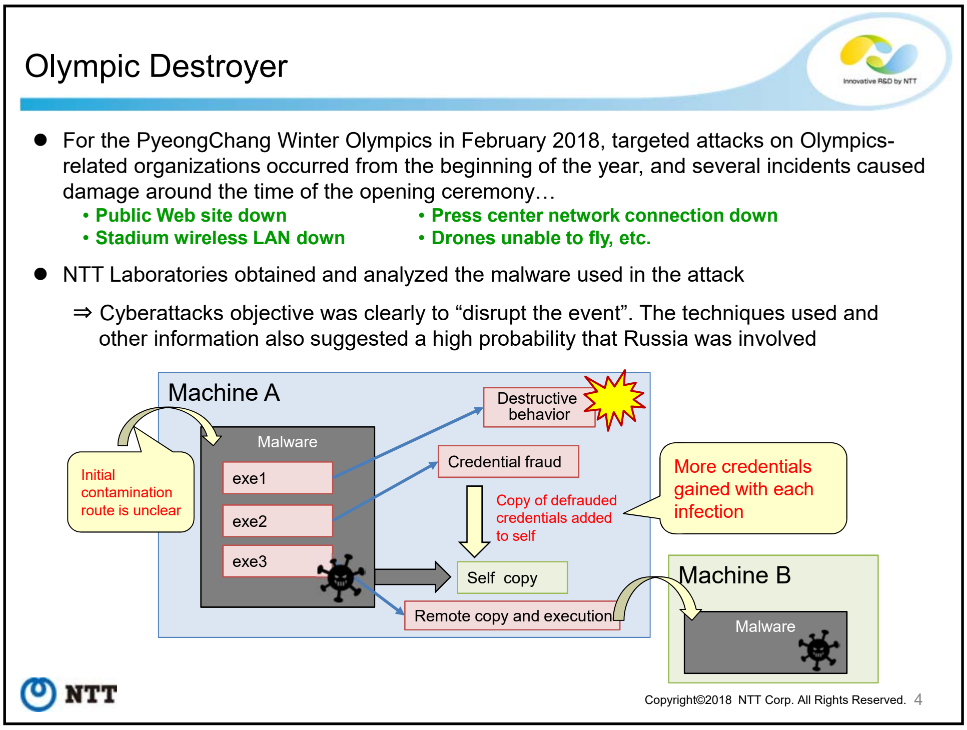


Table of Contents / Amid environmental changes involving Cyberspace

1. Cat-and-Mouse
2. Non-IT ”IoT/OT” fields
3. Protecting critical infrastructure
4. Data utilization in society
5. Singularity (2045 problem)

NTT



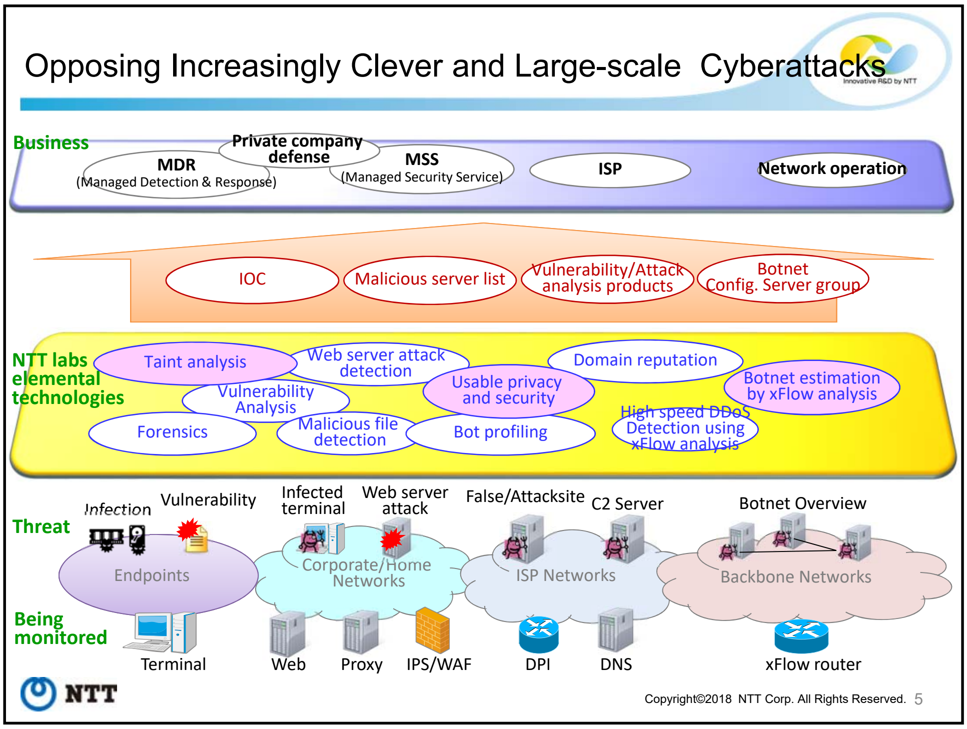
Olympic Destroyer

- For the PyeongChang Winter Olympics in February 2018, targeted attacks on Olympics-related organizations occurred from the beginning of the year, and several incidents caused damage around the time of the opening ceremony…
  - **Public Web site down**
  - **Press center network connection down**
  - **Stadium wireless LAN down**
  - **Drones unable to fly, etc.**
- NTT Laboratories obtained and analyzed the malware used in the attack

  ⇒ Cyberattacks objective was clearly to “disrupt the event”. The techniques used and other information also suggested a high probability that Russia was involved

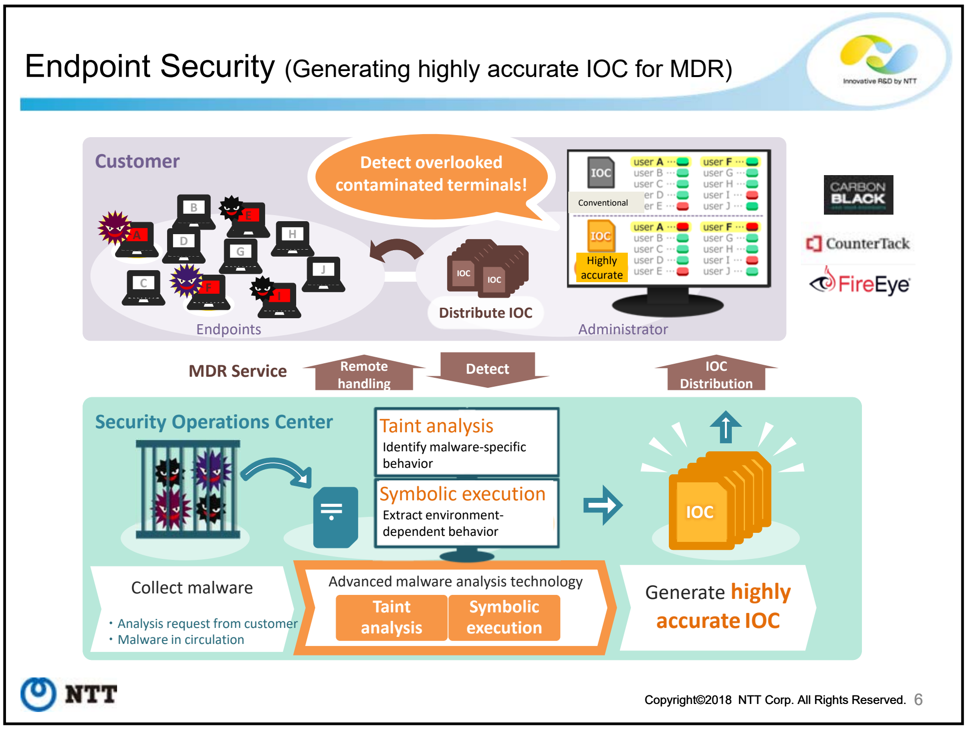
Machine A
Initial contamination route is unclear
Malware
exe1
exe2
exe3
Destructive behavior
Credential fraud
Copy of defrauded credentials added to self
Self copy
Remote copy and execution
More credentials gained with each infection
Machine B
Malware

NTT



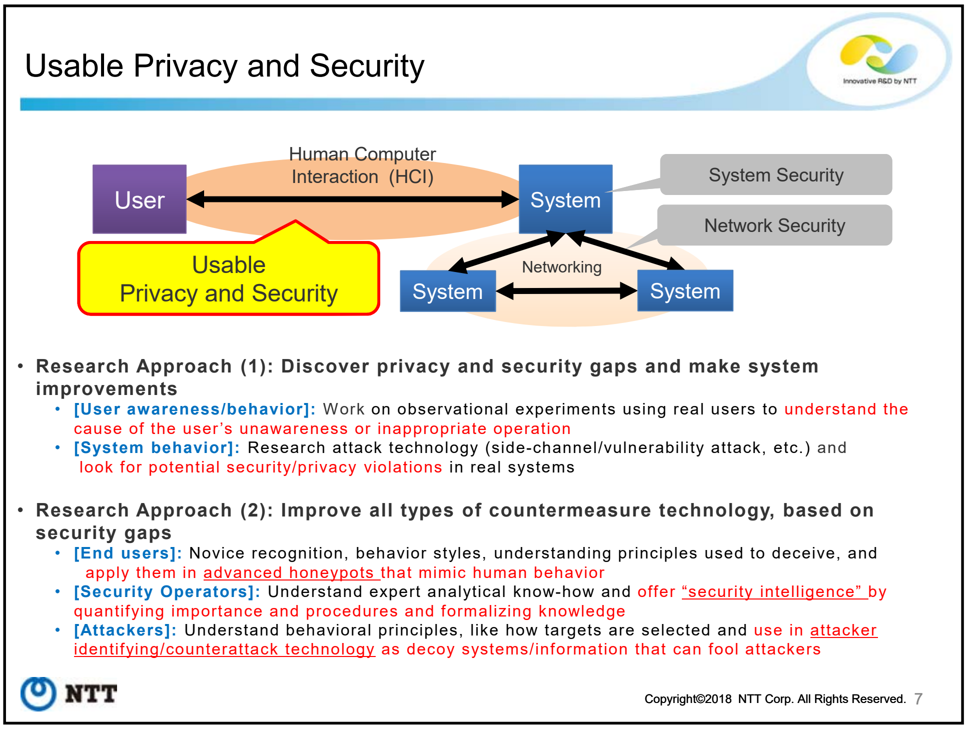
Opposing Increasingly Clever and Large-scale Cyberattacks

5



Endpoint Security (Generating highly accurate IOC for MDR)

6



Usable Privacy and Security

Human Computer Interaction (HCI)

User

System

System Security

Network Security

Usable Privacy and Security

Networking

System

System

- **Research Approach (1): Discover privacy and security gaps and make system improvements**
  - **[User awareness/behavior]:** Work on observational experiments using real users to understand the cause of the user's unawareness or inappropriate operation
  - **[System behavior]:** Research attack technology (side-channel/vulnerability attack, etc.) and look for potential security/privacy violations in real systems

- **Research Approach (2): Improve all types of countermeasure technology, based on security gaps**
  - **[End users]:** Novice recognition, behavior styles, understanding principles used to deceive, and apply them in advanced honeypots that mimic human behavior
  - **[Security Operators]:** Understand expert analytical know-how and offer "security intelligence" by quantifying importance and procedures and formalizing knowledge
  - **[Attackers]:** Understand behavioral principles, like how targets are selected and use in attacker identifying/counterattack technology as decoy systems/information that can fool attackers

NTT

7

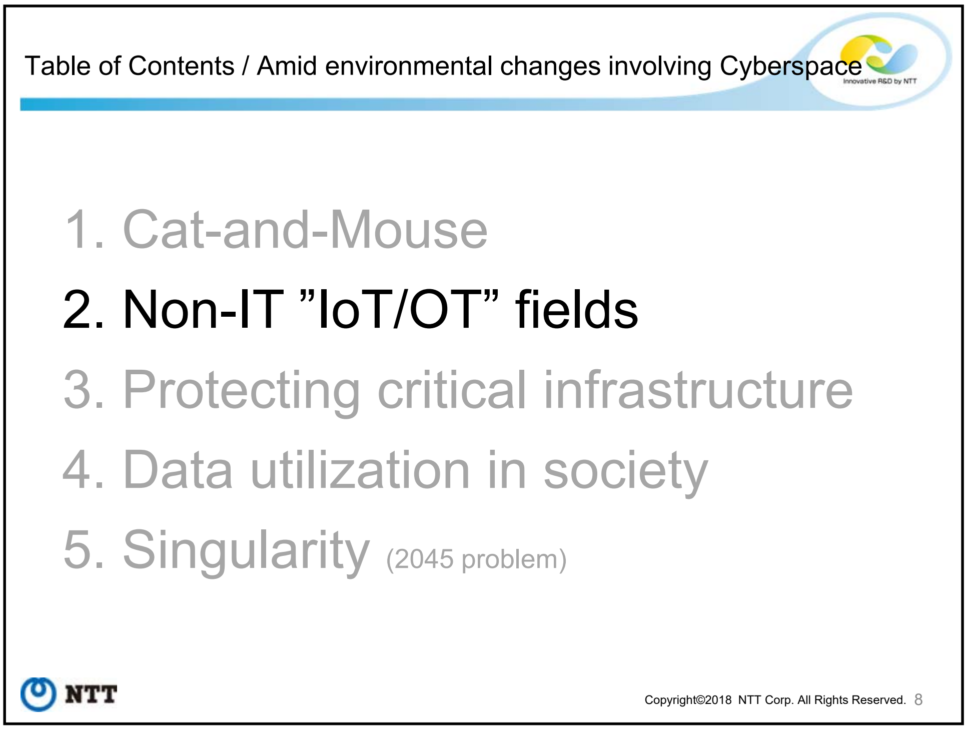


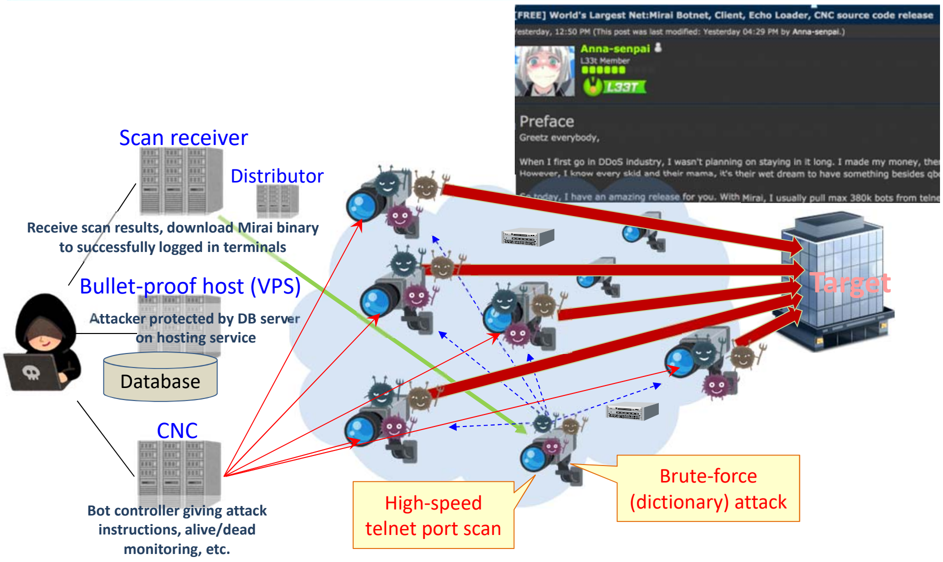
Table of Contents / Amid environmental changes involving Cyberspace

1. Cat-and-Mouse
2. Non-IT "IoT/OT" fields
3. Protecting critical infrastructure
4. Data utilization in society
5. Singularity (2045 problem)

NTT

8

Mirai Attack Mechanism - Analysis by NTT Laboratories -



[FREE] World's Largest Net:Mirai Botnet, Client, Echo Loader, CNC source code release

Yesterday, 12:50 PM (This post was last modified: Yesterday 04:29 PM by Anna-senpai.)

Anna-senpai

L33t Member

L33T

Preface

Greetz everybody,

When I first go in DDoS industry, I wasn't planning on staying in it long. I made my money, the

However, I know every skid and their mama, it's their wet dream to have something besides qb

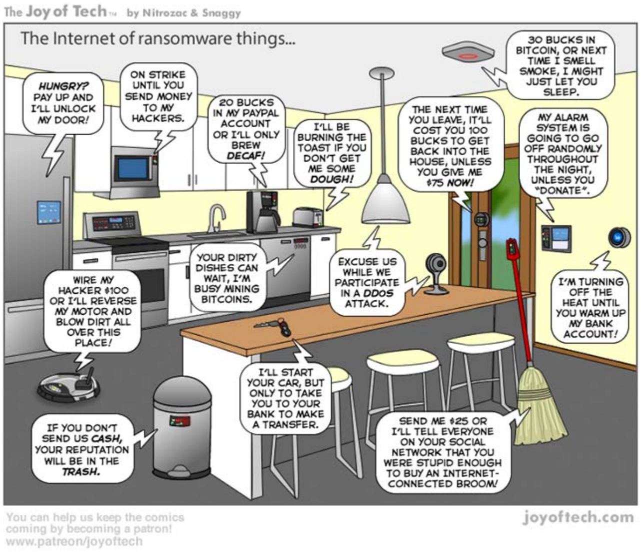
So today, I have an amazing release for you. With Mirai, I usually pull max 380k bots from telne

Scan receiver
Distributor
Receive scan results, download Mirai binary to successfully logged in terminals
Bullet-proof host (VPS)
Attacker protected by DB server on hosting service
Database
CNC
Bot controller giving attack instructions, alive/dead monitoring, etc.
Target
High-speed telnet port scan
Brute-force (dictionary) attack

https://krebsonsecurity.com/2016/10/source-code-for-iot-botnet-mirai-released/

9

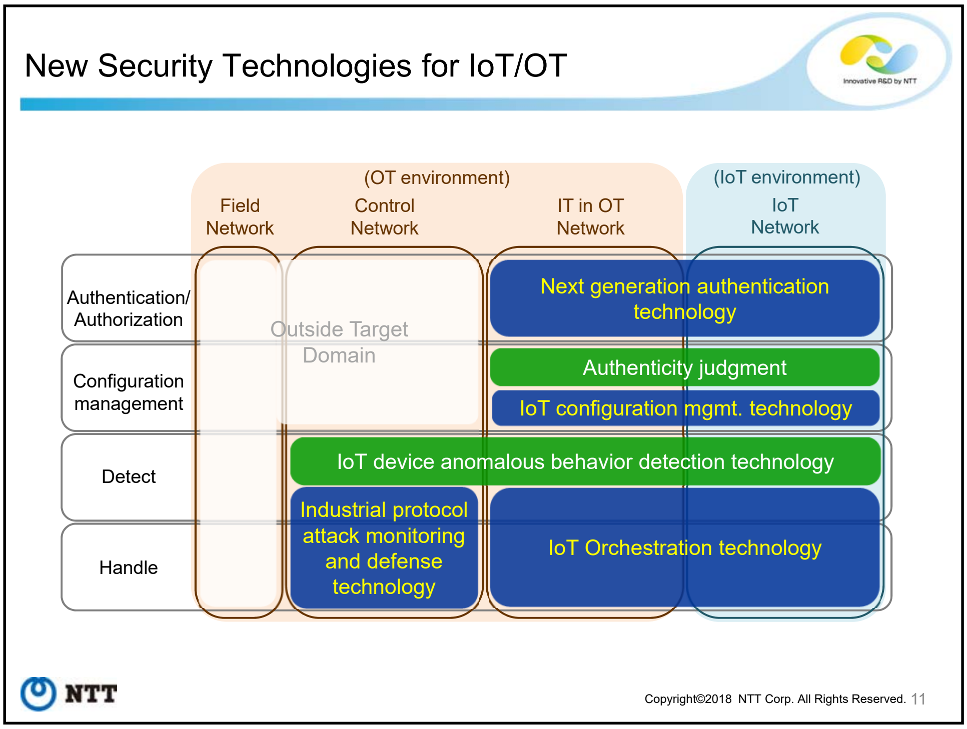
IoT Ransomware?! - Coming Soon -



The Joy of Tech by Nitrozac & Snaggy
The Internet of ransomware things...
HUNGRY? PAY UP AND I'LL UNLOCK MY DOOR!
ON STRIKE UNTIL YOU SEND MONEY TO MY HACKERS.
20 BUCKS IN MY PAYPAL ACCOUNT OR I'LL ONLY BREW DECAF!
I'LL BE BURNING THE TOAST IF YOU DON'T GET ME SOME DOUGH!
THE NEXT TIME YOU LEAVE, IT'LL COST YOU 100 BUCKS TO GET BACK INTO THE HOUSE, UNLESS YOU GIVE ME $75 NOW!
30 BUCKS IN BITCOIN, OR NEXT TIME I SMELL SMOKE, I MIGHT JUST LET YOU SLEEP.
MY ALARM SYSTEM IS GOING TO GO OFF RANDOMLY THROUGHOUT THE NIGHT, UNLESS YOU "DONATE".
YOUR DIRTY DISHES CAN WAIT, I'M BUSY MINING BITCOINS.
EXCUSE US WHILE WE PARTICIPATE IN A DDOS ATTACK.
I'M TURNING OFF THE HEAT UNTIL YOU WARM UP MY BANK ACCOUNT!
WIRE MY HACKER $100 OR I'LL REVERSE MY MOTOR AND BLOW DIRT ALL OVER THIS PLACE!
I'LL START YOUR CAR, BUT ONLY TO TAKE YOU TO YOUR BANK TO MAKE A TRANSFER.
IF YOU DON'T SEND US CASH, YOUR REPUTATION WILL BE IN THE TRASH.
SEND ME $25 OR I'LL TELL EVERYONE ON YOUR SOCIAL NETWORK THAT YOU WERE STUPID ENOUGH TO BUY AN INTERNET-CONNECTED BROOM!
You can help us keep the comics coming by becoming a patron! www.patreon/joyoftech
joyoftech.com

https://www.reddit.com/r/Bitcoin/comments/56m0qu/the_internet_of_ransomware_things/

10



New Security Technologies for IoT/OT

(OT environment)

Field Network

Control Network

IT in OT Network

(IoT environment)

IoT Network

Authentication/ Authorization

Configuration management

Detect

Handle

Outside Target Domain

Next generation authentication technology

Authenticity judgment

IoT configuration mgmt. technology

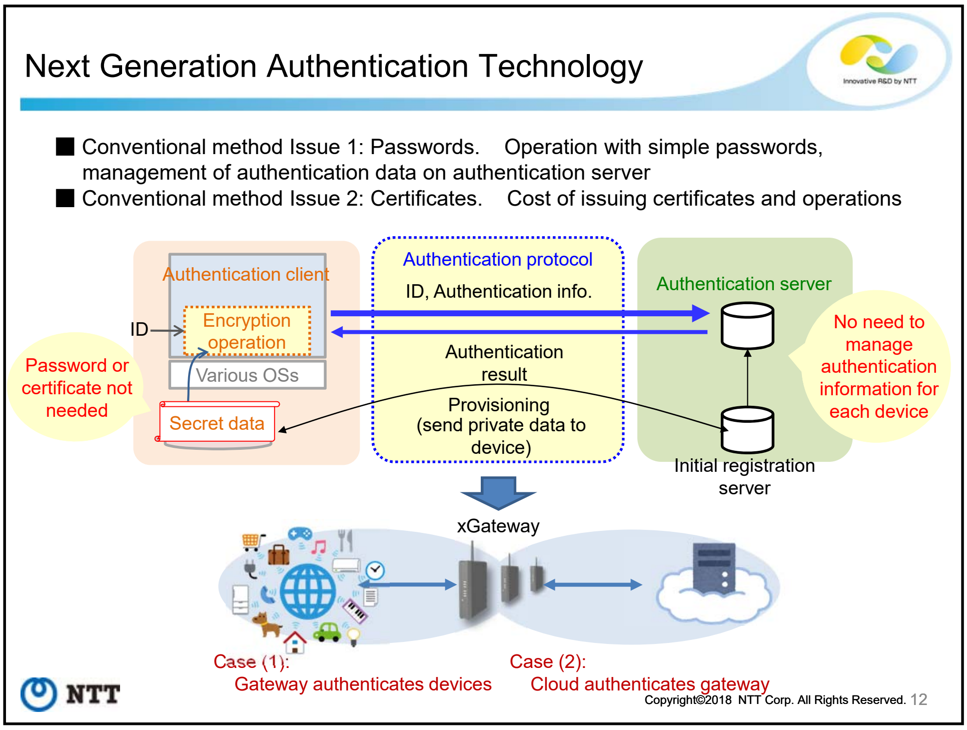
IoT device anomalous behavior detection technology

Industrial protocol attack monitoring and defense technology

IoT Orchestration technology

NTT

11



Next Generation Authentication Technology

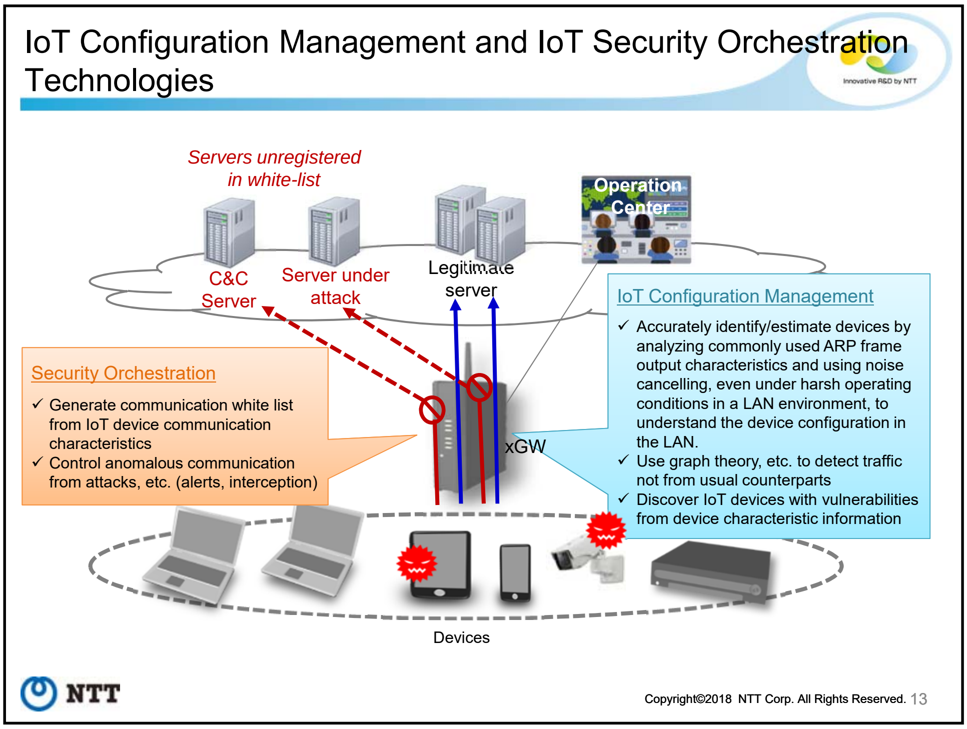
■ Conventional method Issue 1: Passwords. Operation with simple passwords, management of authentication data on authentication server

■ Conventional method Issue 2: Certificates. Cost of issuing certificates and operations

Authentication client
ID
Encryption operation
Various OSs
Secret data
Password or certificate not needed
Authentication protocol
ID, Authentication info.
Authentication result
Provisioning (send private data to device)
Authentication server
Initial registration server
No need to manage authentication information for each device
xGateway
Case (1): Gateway authenticates devices
Case (2): Cloud authenticates gateway

NTT

12



IoT Configuration Management and IoT Security Orchestration Technologies

Servers unregistered in white-list

C&C Server

Server under attack

Legitimate server

Operation Center

xGW

Devices

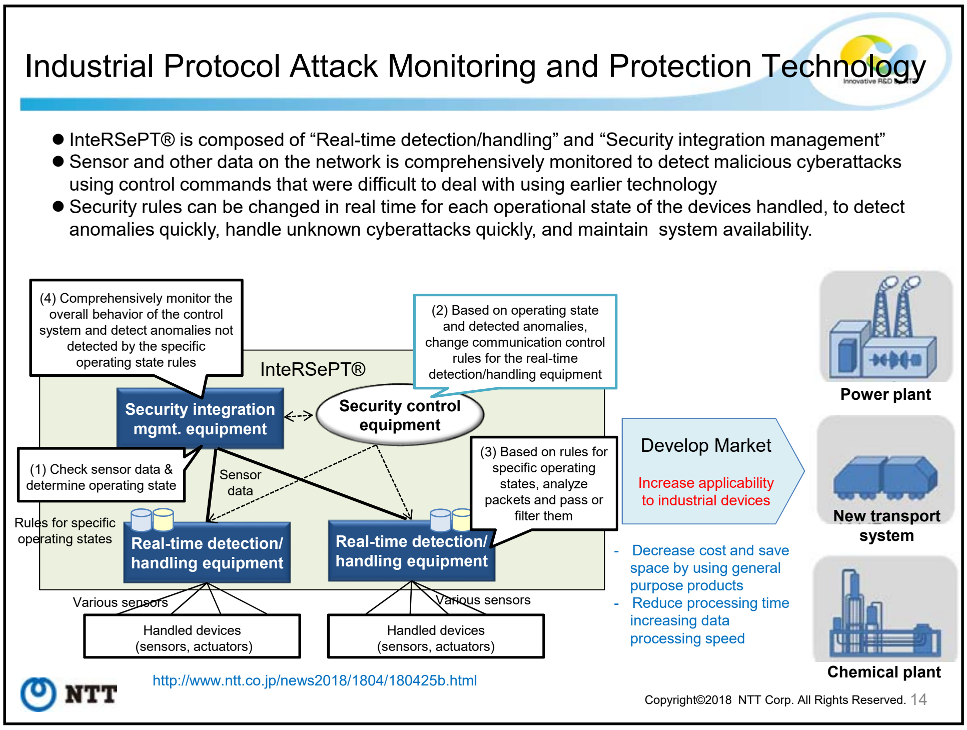
Security Orchestration

- ✓ Generate communication white list from IoT device communication characteristics
- ✓ Control anomalous communication from attacks, etc. (alerts, interception)

IoT Configuration Management

- ✓ Accurately identify/estimate devices by analyzing commonly used ARP frame output characteristics and using noise cancelling, even under harsh operating conditions in a LAN environment, to understand the device configuration in the LAN.
- ✓ Use graph theory, etc. to detect traffic not from usual counterparts
- ✓ Discover IoT devices with vulnerabilities from device characteristic information

13



Industrial Protocol Attack Monitoring and Protection Technology

- ● InteRSePT® is composed of “Real-time detection/handling” and “Security integration management”
- ● Sensor and other data on the network is comprehensively monitored to detect malicious cyberattacks using control commands that were difficult to deal with using earlier technology
- ● Security rules can be changed in real time for each operational state of the devices handled, to detect anomalies quickly, handle unknown cyberattacks quickly, and maintain system availability.

(4) Comprehensively monitor the overall behavior of the control system and detect anomalies not detected by the specific operating state rules

(2) Based on operating state and detected anomalies, change communication control rules for the real-time detection/handling equipment

InteRSePT®

Security integration mgmt. equipment

Security control equipment

(1) Check sensor data & determine operating state

Sensor data

(3) Based on rules for specific operating states, analyze packets and pass or filter them

Rules for specific operating states

Real-time detection/ handling equipment

Real-time detection/ handling equipment

Various sensors

Various sensors

Handled devices (sensors, actuators)

Handled devices (sensors, actuators)

Develop Market

Increase applicability to industrial devices

- Decrease cost and save space by using general purpose products
- Reduce processing time increasing data processing speed

Power plant

New transport system

Chemical plant

http://www.ntt.co.jp/news2018/1804/180425b.html

14

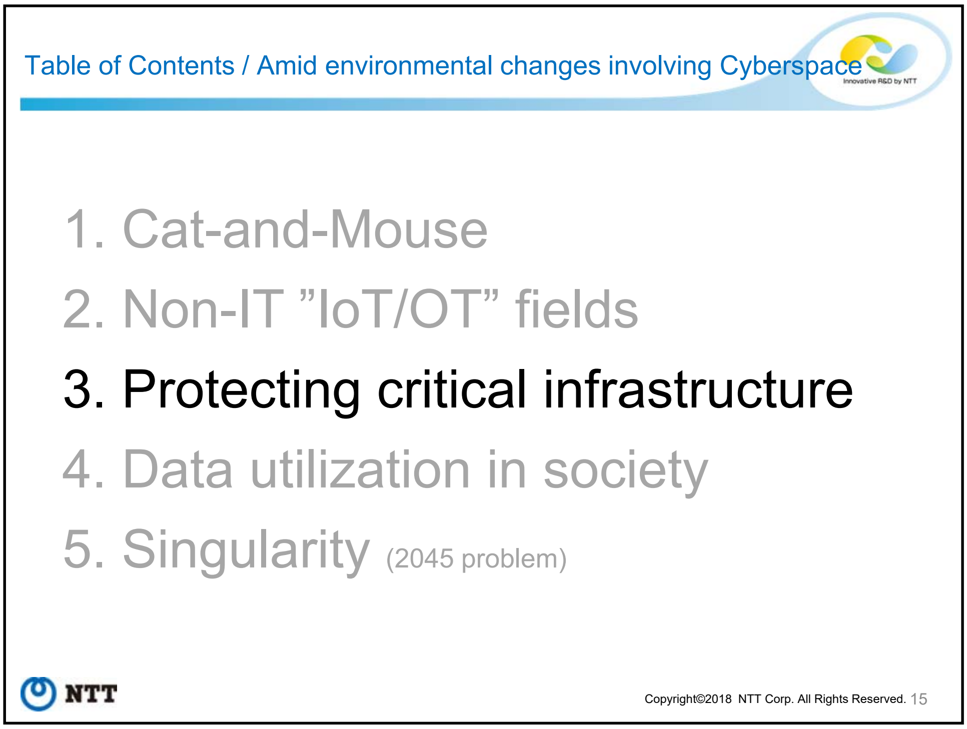
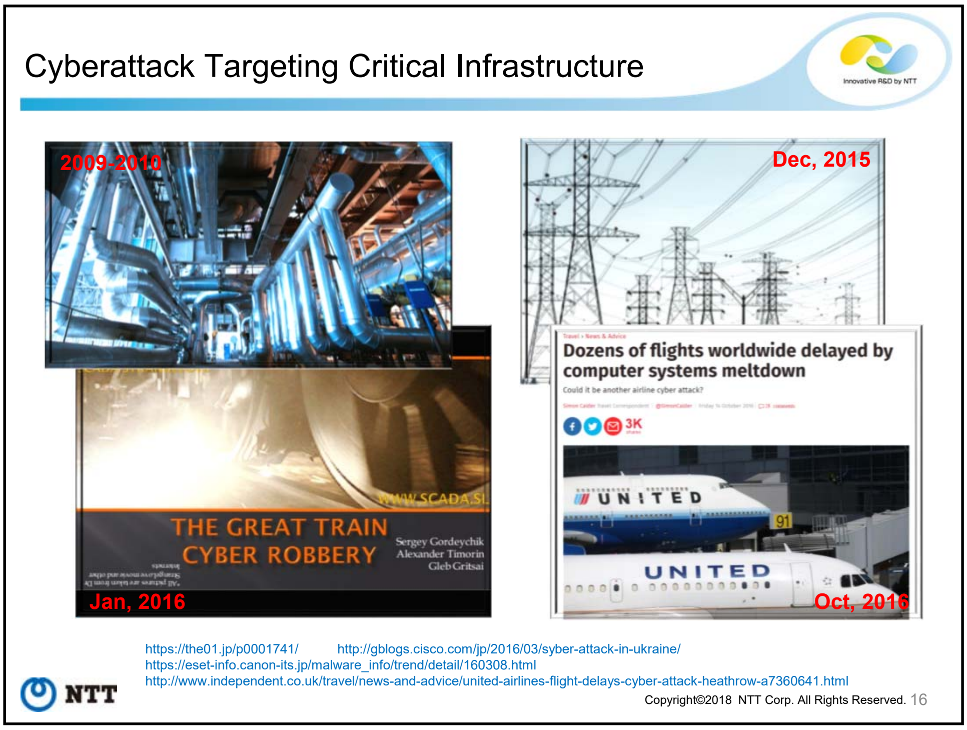


Table of Contents / Amid environmental changes involving Cyberspace

1. Cat-and-Mouse
2. Non-IT ”IoT/OT” fields
3. Protecting critical infrastructure
4. Data utilization in society
5. Singularity (2045 problem)

Innovative R&D by NTT
NTT



Cyberattack Targeting Critical Infrastructure

2009-2010

Dec, 2015

THE GREAT TRAIN CYBER ROBBERY

Sergey Gordeychik
Alexander Timorin
Gleb Gritsai

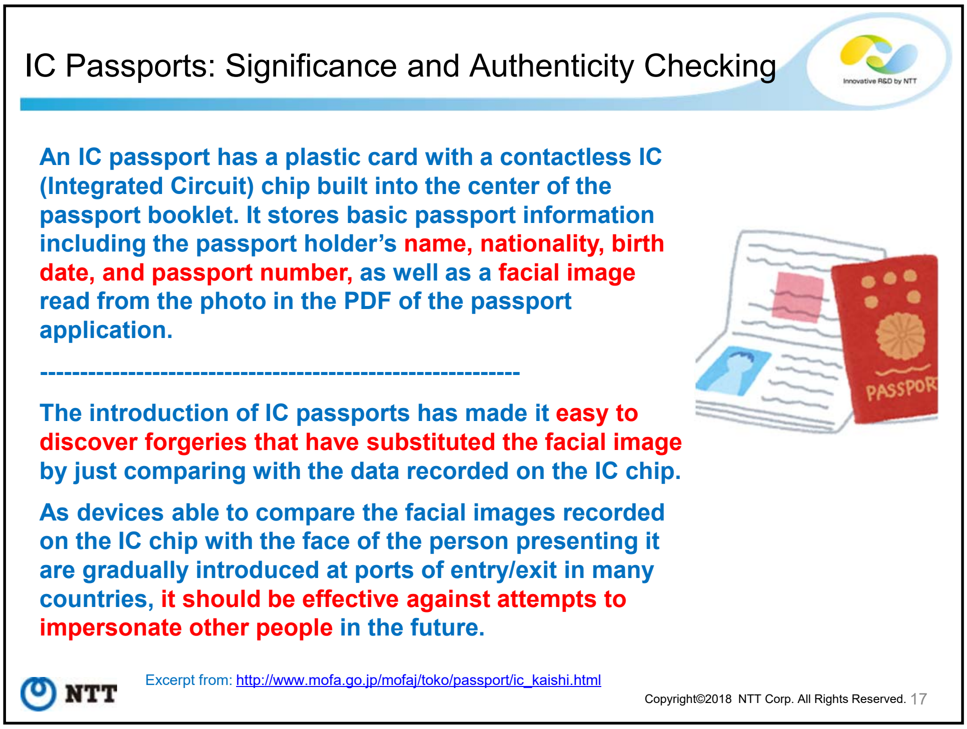
Jan, 2016

Dozens of flights worldwide delayed by computer systems meltdown

Could it be another airline cyber attack?

Oct, 2016

https://the01.jp/p0001741/ http://gblogs.cisco.com/jp/2016/03/syber-attack-in-ukraine/
https://eset-info.canon-its.jp/malware_info/trend/detail/160308.html
http://www.independent.co.uk/travel/news-and-advice/united-airlines-flight-delays-cyber-attack-heathrow-a7360641.html



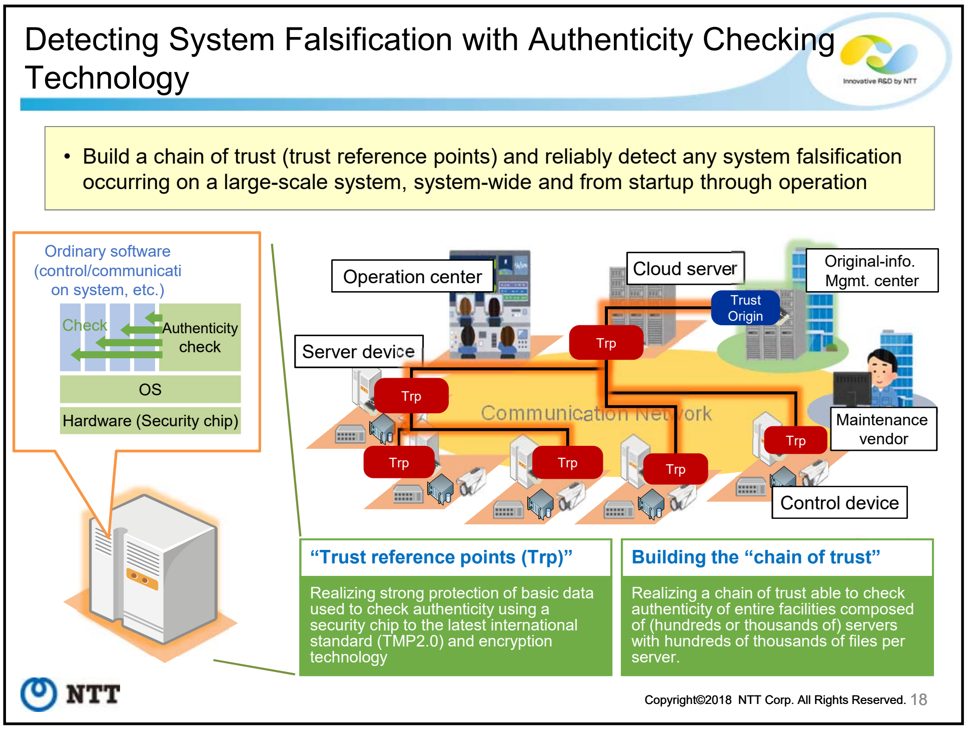
IC Passports: Significance and Authenticity Checking

An IC passport has a plastic card with a contactless IC (Integrated Circuit) chip built into the center of the passport booklet. It stores basic passport information including the passport holder's name, nationality, birth date, and passport number, as well as a facial image read from the photo in the PDF of the passport application.

The introduction of IC passports has made it easy to discover forgeries that have substituted the facial image by just comparing with the data recorded on the IC chip.

As devices able to compare the facial images recorded on the IC chip with the face of the person presenting it are gradually introduced at ports of entry/exit in many countries, it should be effective against attempts to impersonate other people in the future.

Excerpt from: http://www.mofa.go.jp/mofaj/toko/passport/ic_kaishi.html



Detecting System Falsification with Authenticity Checking Technology

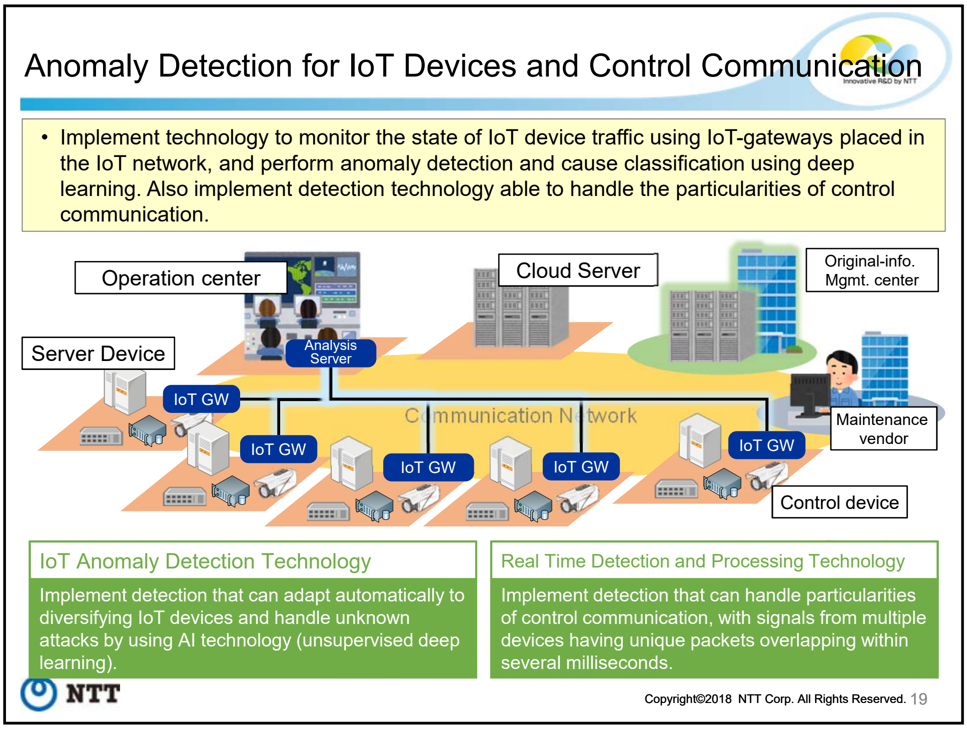
- Build a chain of trust (trust reference points) and reliably detect any system falsification occurring on a large-scale system, system-wide and from startup through operation

"Trust reference points (Trp)"

Realizing strong protection of basic data used to check authenticity using a security chip to the latest international standard (TMP2.0) and encryption technology

Building the "chain of trust"

Realizing a chain of trust able to check authenticity of entire facilities composed of (hundreds or thousands of) servers with hundreds of thousands of files per server.



Anomaly Detection for IoT Devices and Control Communication

- Implement technology to monitor the state of IoT device traffic using IoT-gateways placed in the IoT network, and perform anomaly detection and cause classification using deep learning. Also implement detection technology able to handle the particularities of control communication.

Operation center

Cloud Server

Original-info. Mgmt. center

Server Device

Analysis Server

IoT GW

Communication Network

Maintenance vendor

Control device

IoT Anomaly Detection Technology

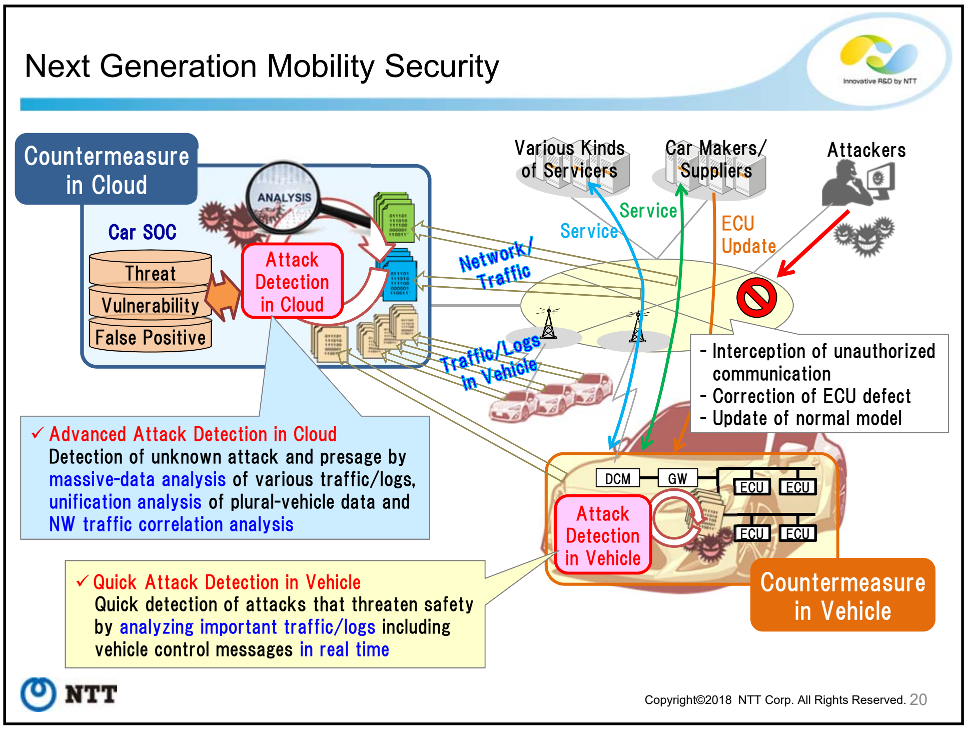
Implement detection that can adapt automatically to diversifying IoT devices and handle unknown attacks by using AI technology (unsupervised deep learning).

Real Time Detection and Processing Technology

Implement detection that can handle particularities of control communication, with signals from multiple devices having unique packets overlapping within several milliseconds.

NTT

19



Next Generation Mobility Security

Countermeasure in Cloud

Car SOC

Threat

Vulnerability

False Positive

ANALYSIS

Attack Detection in Cloud

Various Kinds of Servicers

Car Makers/ Suppliers

Attackers

Service

Service

ECU Update

Network/ Traffic

Traffic/Logs in Vehicle

- Interception of unauthorized communication
- Correction of ECU defect
- Update of normal model

DCM

GW

ECU

ECU

ECU

ECU

Attack Detection in Vehicle

Countermeasure in Vehicle

✓ Advanced Attack Detection in Cloud

Detection of unknown attack and presage by massive-data analysis of various traffic/logs, unification analysis of plural-vehicle data and NW traffic correlation analysis

✓ Quick Attack Detection in Vehicle

Quick detection of attacks that threaten safety by analyzing important traffic/logs including vehicle control messages in real time

NTT

20

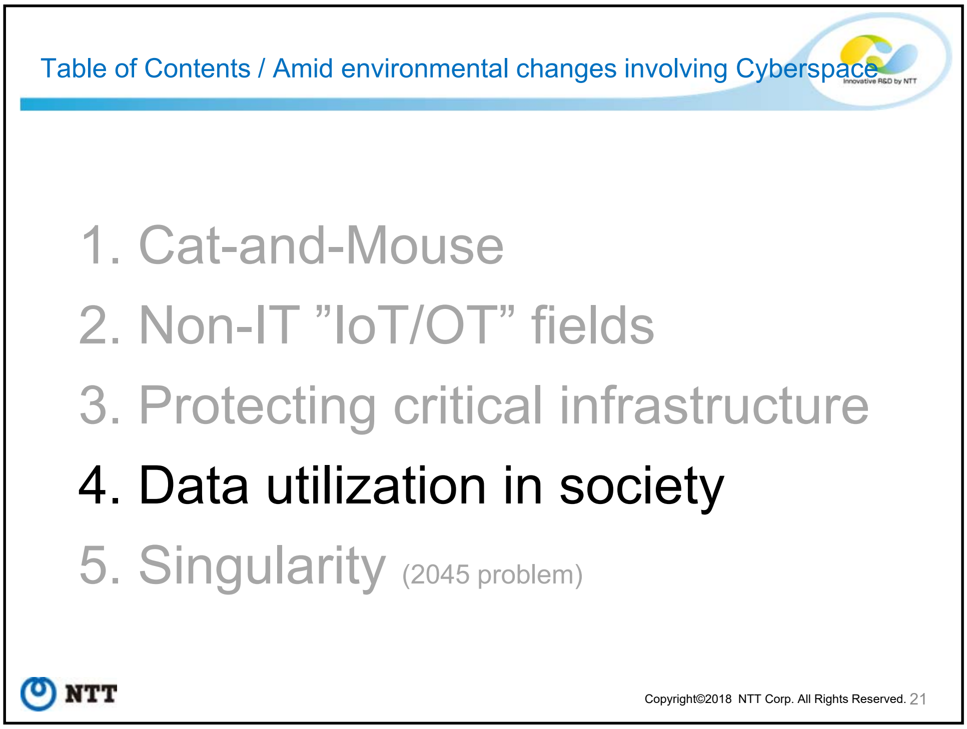
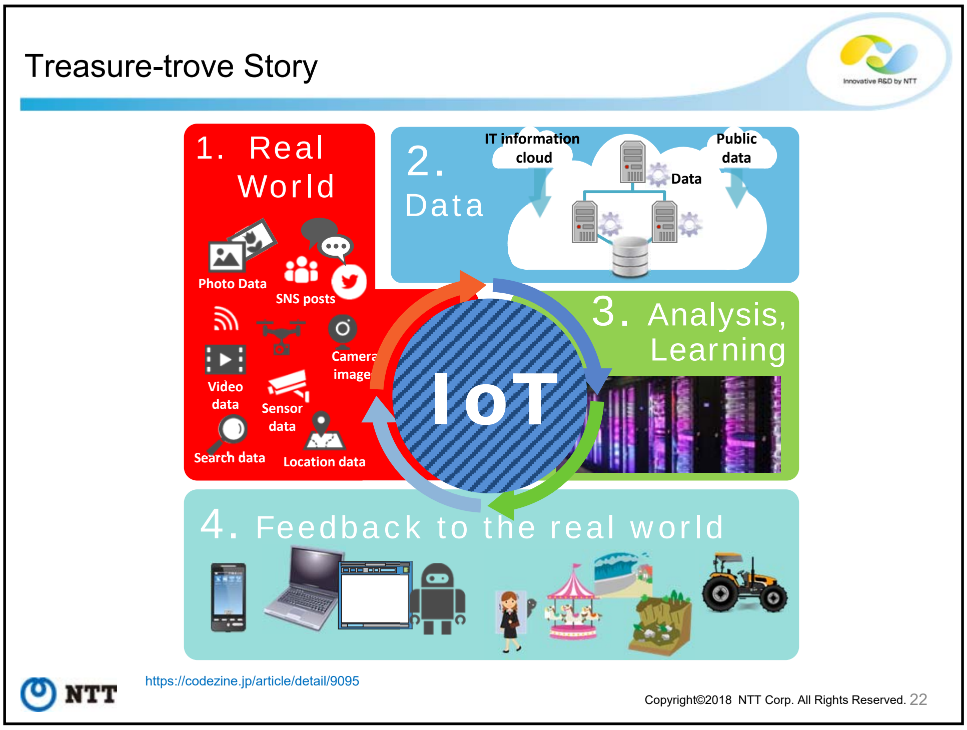


Table of Contents / Amid environmental changes involving Cyberspace

1. Cat-and-Mouse
2. Non-IT ”IoT/OT” fields
3. Protecting critical infrastructure
4. Data utilization in society
5. Singularity (2045 problem)

NTT

21



Treasure-trove Story

1. Real World

Photo Data

SNS posts

Video data

Sensor data

Camera image

Search data

Location data

2. Data

IT information cloud

Public data

Data

3. Analysis, Learning

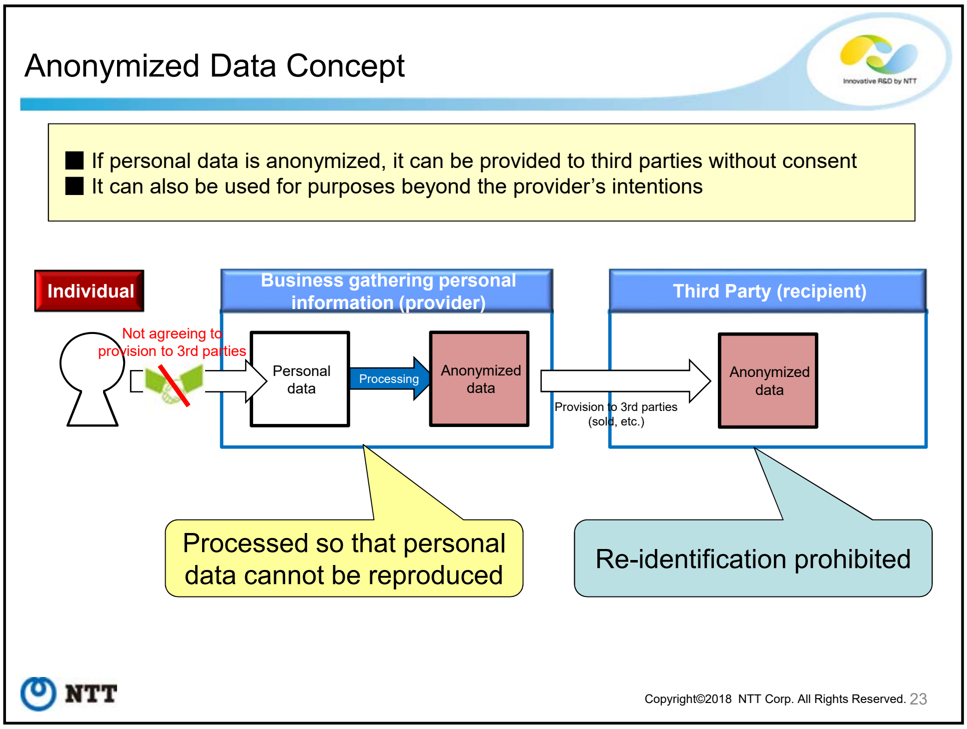
IoT

4. Feedback to the real world

NTT

https://codezine.jp/article/detail/9095

22



Anonymized Data Concept

- ■ If personal data is anonymized, it can be provided to third parties without consent
- ■ It can also be used for purposes beyond the provider's intentions

Individual

Business gathering personal information (provider)

Third Party (recipient)

Not agreeing to provision to 3rd parties

Personal data

Processing

Anonymized data

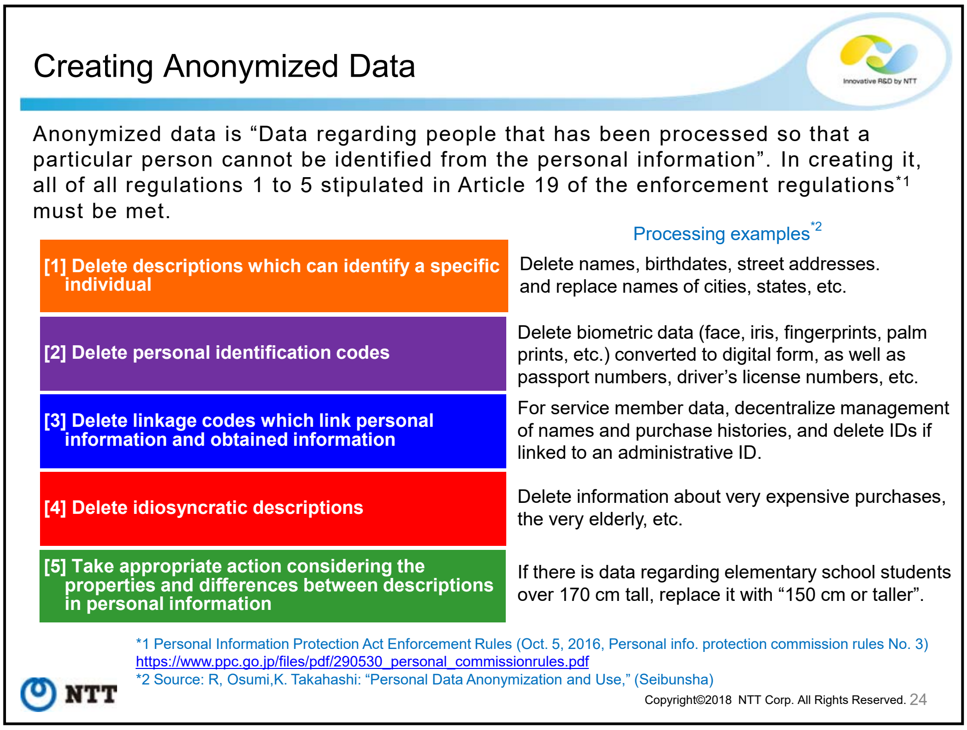
Provision to 3rd parties (sold, etc.)

Anonymized data

Processed so that personal data cannot be reproduced

Re-identification prohibited

NTT



Creating Anonymized Data

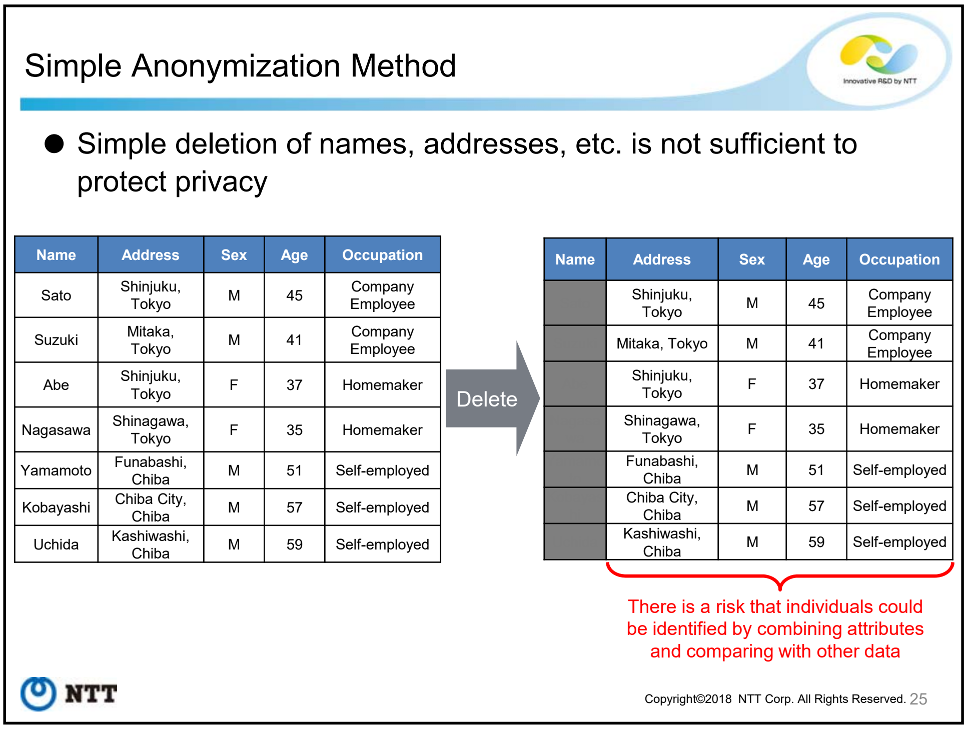
Anonymized data is "Data regarding people that has been processed so that a particular person cannot be identified from the personal information". In creating it, all of all regulations 1 to 5 stipulated in Article 19 of the enforcement regulations*1 must be met.

	Processing examples*2
[1] Delete descriptions which can identify a specific individual	Delete names, birthdates, street addresses. and replace names of cities, states, etc.
[2] Delete personal identification codes	Delete biometric data (face, iris, fingerprints, palm prints, etc.) converted to digital form, as well as passport numbers, driver's license numbers, etc.
[3] Delete linkage codes which link personal information and obtained information	For service member data, decentralize management of names and purchase histories, and delete IDs if linked to an administrative ID.
[4] Delete idiosyncratic descriptions	Delete information about very expensive purchases, the very elderly, etc.
[5] Take appropriate action considering the properties and differences between descriptions in personal information	If there is data regarding elementary school students over 170 cm tall, replace it with "150 cm or taller".

*1 Personal Information Protection Act Enforcement Rules (Oct. 5, 2016, Personal info. protection commission rules No. 3) https://www.ppc.go.jp/files/pdf/290530_personal_commissionrules.pdf

*2 Source: R, Osumi,K. Takahashi: "Personal Data Anonymization and Use," (Seibunsha)

NTT



Simple Anonymization Method

- Simple deletion of names, addresses, etc. is not sufficient to protect privacy

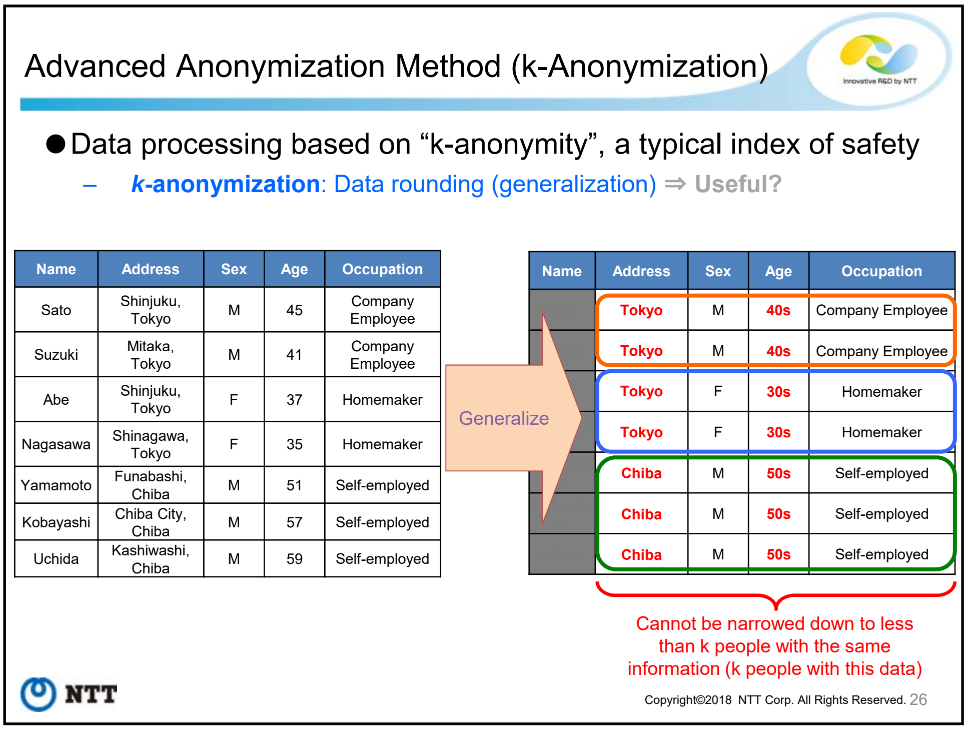
Name	Address	Sex	Age	Occupation
Sato	Shinjuku, Tokyo	M	45	Company Employee
Suzuki	Mitaka, Tokyo	M	41	Company Employee
Abe	Shinjuku, Tokyo	F	37	Homemaker
Nagasawa	Shinagawa, Tokyo	F	35	Homemaker
Yamamoto	Funabashi, Chiba	M	51	Self-employed
Kobayashi	Chiba City, Chiba	M	57	Self-employed
Uchida	Kashiwashi, Chiba	M	59	Self-employed

Name	Address	Sex	Age	Occupation
	Shinjuku, Tokyo	M	45	Company Employee
	Mitaka, Tokyo	M	41	Company Employee
	Shinjuku, Tokyo	F	37	Homemaker
	Shinagawa, Tokyo	F	35	Homemaker
	Funabashi, Chiba	M	51	Self-employed
	Chiba City, Chiba	M	57	Self-employed
	Kashiwashi, Chiba	M	59	Self-employed

There is a risk that individuals could be identified by combining attributes and comparing with other data

NTT

25



Advanced Anonymization Method (k-Anonymization)

- Data processing based on “k-anonymity”, a typical index of safety
  - ***k*-anonymization**: Data rounding (generalization) ⇒ **Useful?**

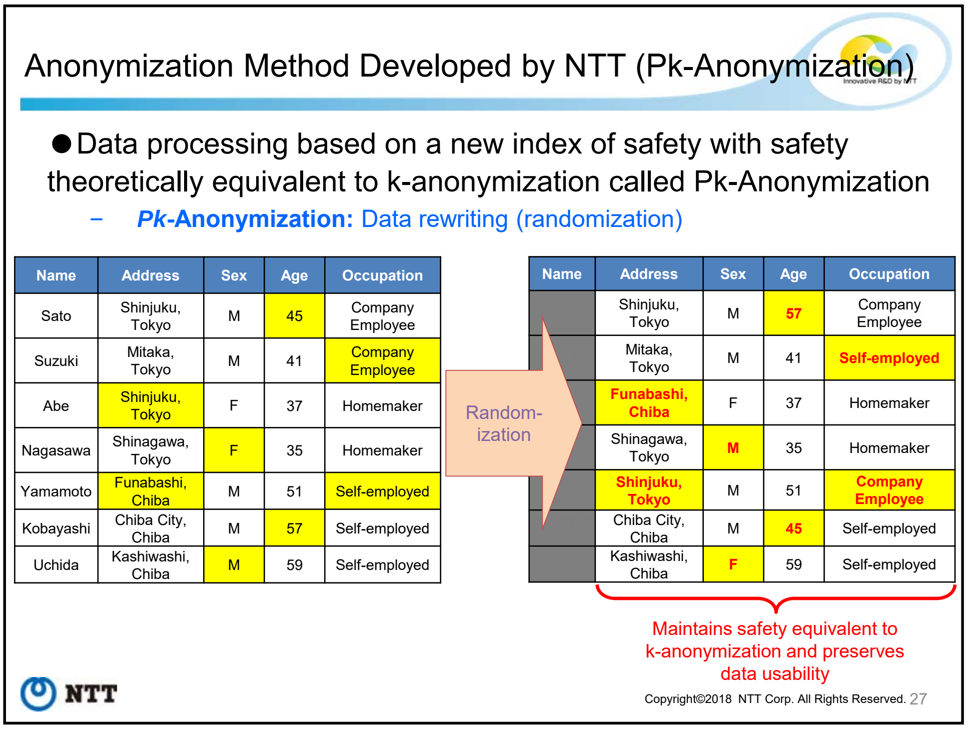
Name	Address	Sex	Age	Occupation
Sato	Shinjuku, Tokyo	M	45	Company Employee
Suzuki	Mitaka, Tokyo	M	41	Company Employee
Abe	Shinjuku, Tokyo	F	37	Homemaker
Nagasawa	Shinagawa, Tokyo	F	35	Homemaker
Yamamoto	Funabashi, Chiba	M	51	Self-employed
Kobayashi	Chiba City, Chiba	M	57	Self-employed
Uchida	Kashiwashi, Chiba	M	59	Self-employed

Name	Address	Sex	Age	Occupation
	Tokyo	M	**40s**	Company Employee
	Tokyo	M	**40s**	Company Employee
	Tokyo	F	**30s**	Homemaker
	Tokyo	F	**30s**	Homemaker
	Chiba	M	**50s**	Self-employed
	Chiba	M	**50s**	Self-employed
	Chiba	M	**50s**	Self-employed

Cannot be narrowed down to less than k people with the same information (k people with this data)

NTT

26



Anonymization Method Developed by NTT (Pk-Anonymization)

- ●Data processing based on a new index of safety with safety theoretically equivalent to k-anonymization called Pk-Anonymization
  - ***Pk*-Anonymization:** Data rewriting (randomization)

Name	Address	Sex	Age	Occupation
Sato	Shinjuku, Tokyo	M	45	Company Employee
Suzuki	Mitaka, Tokyo	M	41	Company Employee
Abe	Shinjuku, Tokyo	F	37	Homemaker
Nagasawa	Shinagawa, Tokyo	F	35	Homemaker
Yamamoto	Funabashi, Chiba	M	51	Self-employed
Kobayashi	Chiba City, Chiba	M	57	Self-employed
Uchida	Kashiwashi, Chiba	M	59	Self-employed

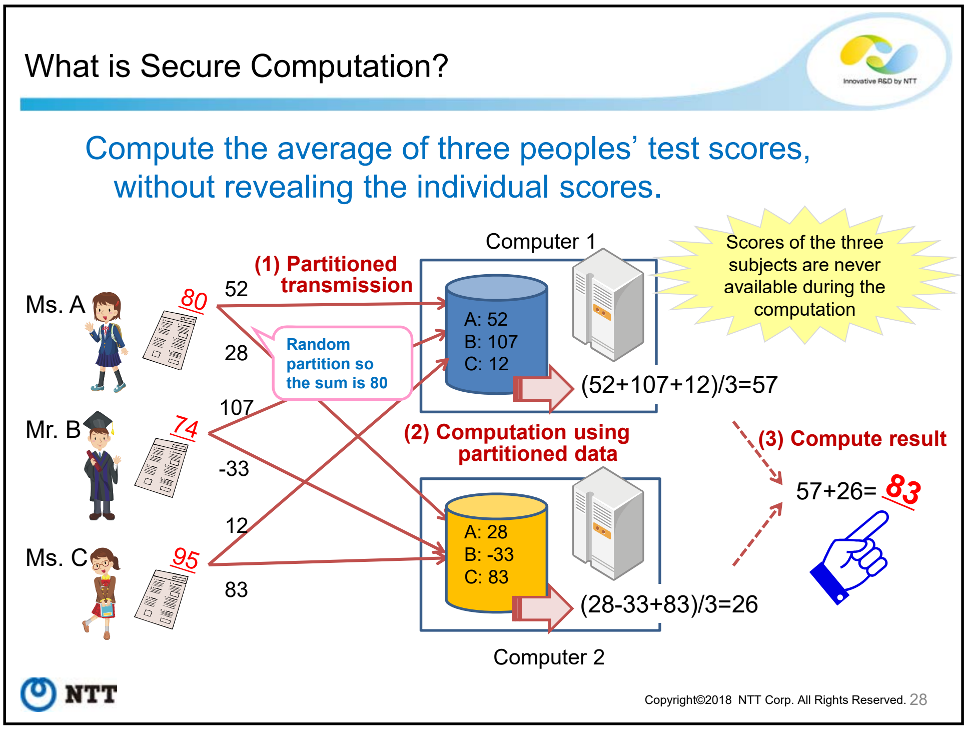
Random-ization

Name	Address	Sex	Age	Occupation
	Shinjuku, Tokyo	M	**57**	Company Employee
	Mitaka, Tokyo	M	41	**Self-employed**
	Funabashi, Chiba	F	37	Homemaker
	Shinagawa, Tokyo	**M**	35	Homemaker
	Shinjuku, Tokyo	M	51	**Company Employee**
	Chiba City, Chiba	M	**45**	Self-employed
	Kashiwashi, Chiba	**F**	59	Self-employed

Maintains safety equivalent to k-anonymization and preserves data usability

NTT

27



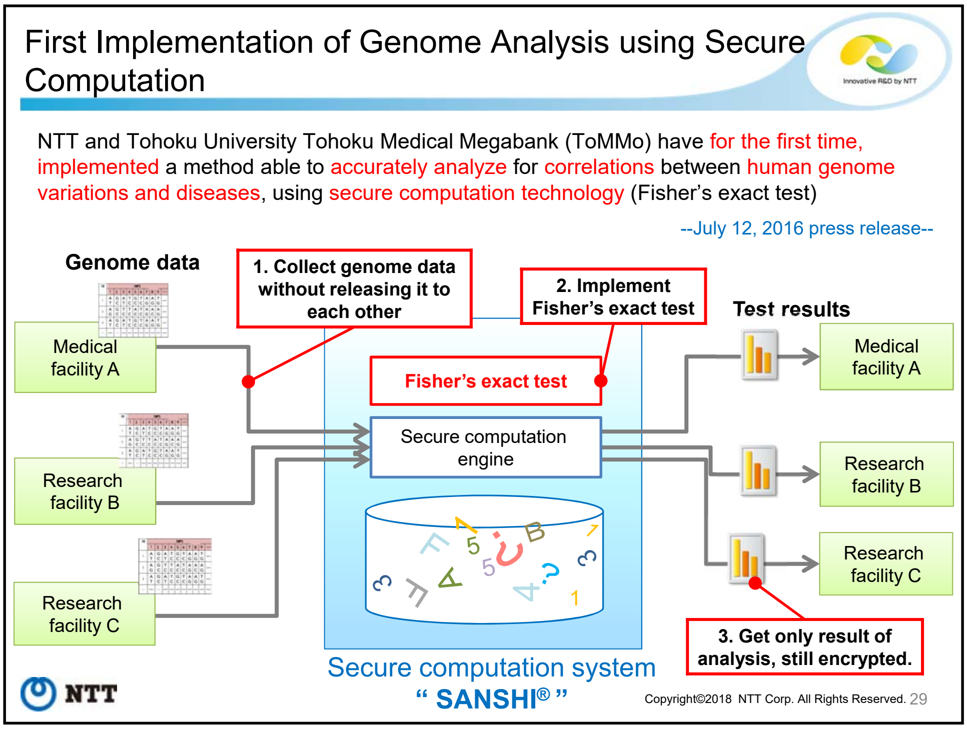
What is Secure Computation?

Compute the average of three peoples' test scores, without revealing the individual scores.

Ms. A
80
Mr. B
74
Ms. C
95
52
28
107
-33
12
83
(1) Partitioned transmission
Random partition so the sum is 80
Computer 1
A: 52
B: 107
C: 12
(52+107+12)/3=57
Scores of the three subjects are never available during the computation
(2) Computation using partitioned data
A: 28
B: -33
C: 83
(28-33+83)/3=26
Computer 2
(3) Compute result
57+26= 83

NTT

28



First Implementation of Genome Analysis using Secure Computation

NTT and Tohoku University Tohoku Medical Megabank (ToMMo) have for the first time, implemented a method able to accurately analyze for correlations between human genome variations and diseases, using secure computation technology (Fisher's exact test)

--July 12, 2016 press release--

29

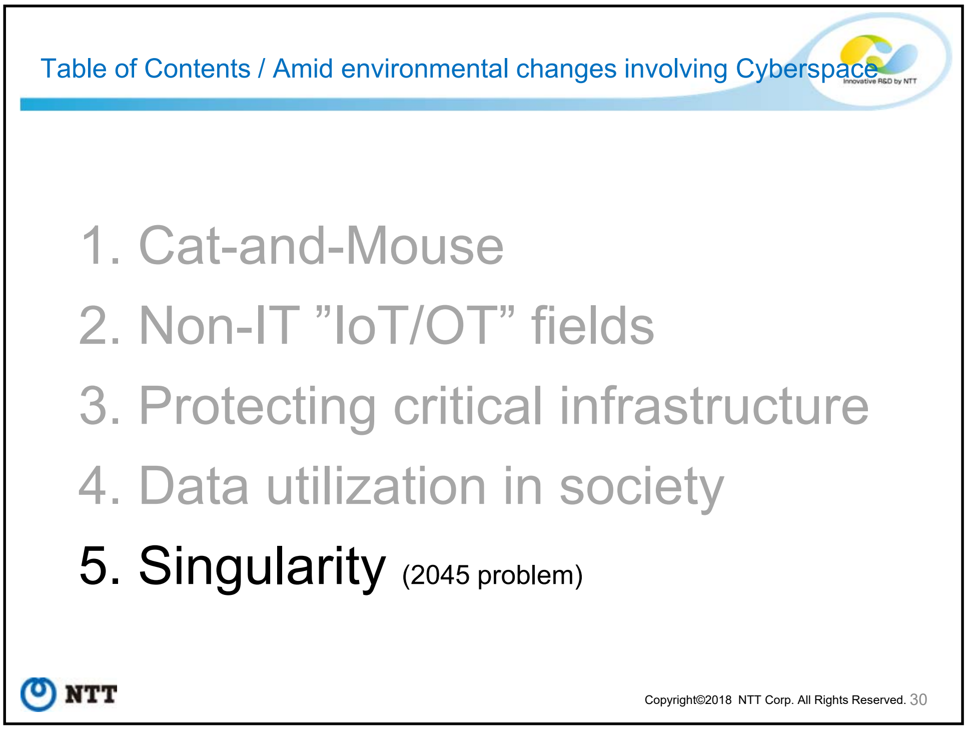


Table of Contents / Amid environmental changes involving Cyberspace

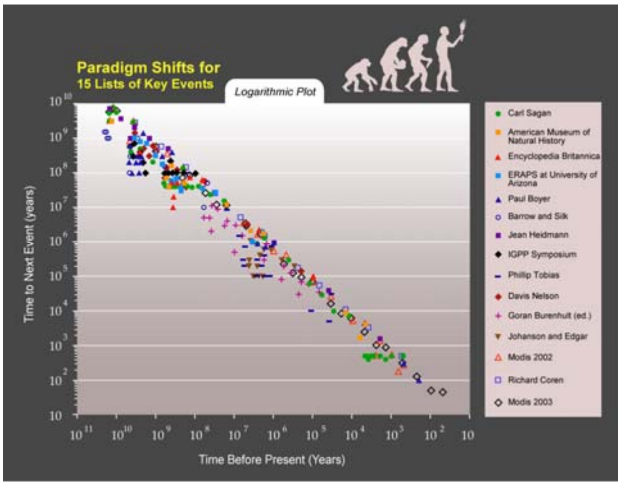
1. Cat-and-Mouse
2. Non-IT "IoT/OT" fields
3. Protecting critical infrastructure
4. Data utilization in society
5. Singularity (2045 problem)

Innovative R&D by NTT
NTT

30

Singularity (2045 problem) - Kurzweil's Law of Accelerating Change -

By 2045, a $1,000 computer will have performance of approximately 10 peta FLOPs, which is ten billion times that of the human brain and a sufficient base for AI to reach a technical singularity

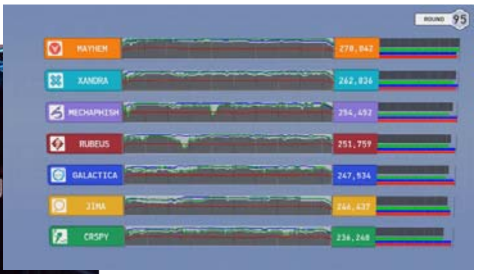


Source: Wikipedia "Technilogical singularity"

NTT

Copyright©2018 NTT Corp. All Rights Reserved. 31

Appearance of AI at the World's Largest Security Hacking Conference (2017)…

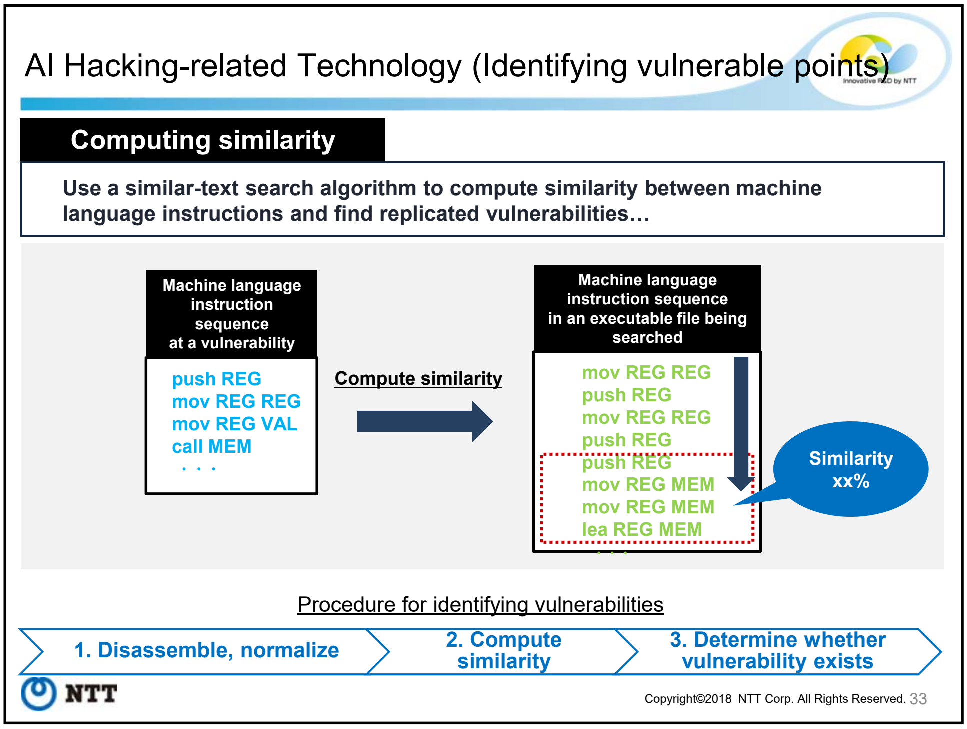


Research on automation of hacking, AI hacking, etc. has already begun. The fact that Mayhem placed 14th at DEFCON, solving problems without human help in a contest with the best hacking teams in the world shows how far R&D has progressed…

https://the01.jp/p0003285/
https://roboteer-tokyo.com/archives/5734

NTT

Copyright©2018 NTT Corp. All Rights Reserved. 32



AI Hacking-related Technology (Identifying vulnerable points)

Computing similarity

Use a similar-text search algorithm to compute similarity between machine language instructions and find replicated vulnerabilities…

Machine language instruction sequence at a vulnerability

push REG
mov REG REG
mov REG VAL
call MEM
・・・

Compute similarity

Machine language instruction sequence in an executable file being searched

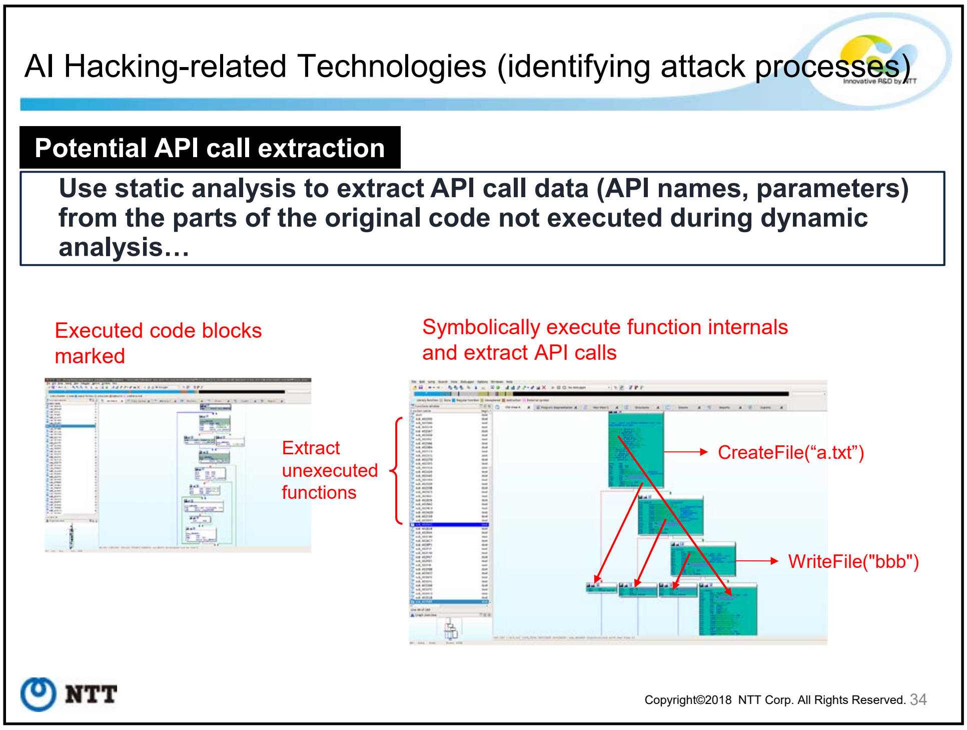
mov REG REG
push REG
mov REG REG
push REG
push REG
mov REG MEM
mov REG MEM
lea REG MEM

Similarity xx%

Procedure for identifying vulnerabilities

1. Disassemble, normalize
2. Compute similarity
3. Determine whether vulnerability exists

NTT



AI Hacking-related Technologies (identifying attack processes)

Potential API call extraction

Use static analysis to extract API call data (API names, parameters) from the parts of the original code not executed during dynamic analysis…

Executed code blocks marked

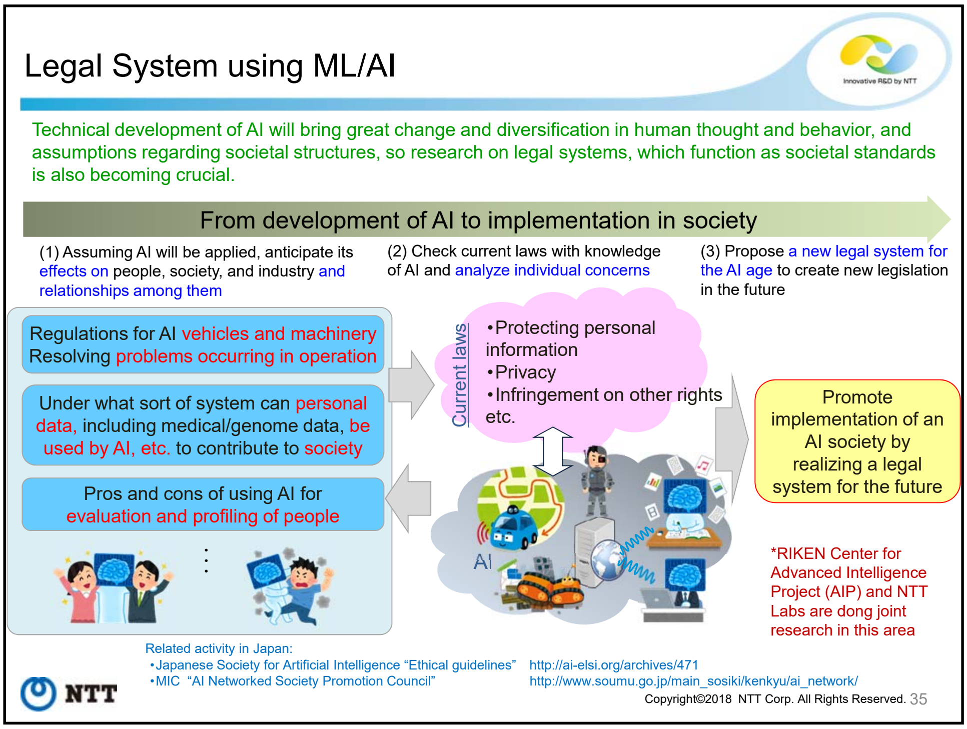
Symbolically execute function internals and extract API calls

Extract unexecuted functions

CreateFile(“a.txt”)

WriteFile("bbb")

NTT



Legal System using ML/AI

Technical development of AI will bring great change and diversification in human thought and behavior, and assumptions regarding societal structures, so research on legal systems, which function as societal standards is also becoming crucial.

From development of AI to implementation in society

(1) Assuming AI will be applied, anticipate its effects on people, society, and industry and relationships among them

(2) Check current laws with knowledge of AI and analyze individual concerns

(3) Propose a new legal system for the AI age to create new legislation in the future

Regulations for AI vehicles and machinery
Resolving problems occurring in operation

Under what sort of system can personal data, including medical/genome data, be used by AI, etc. to contribute to society

Pros and cons of using AI for evaluation and profiling of people

:

Current laws

•Protecting personal information
•Privacy
•Infringement on other rights
etc.

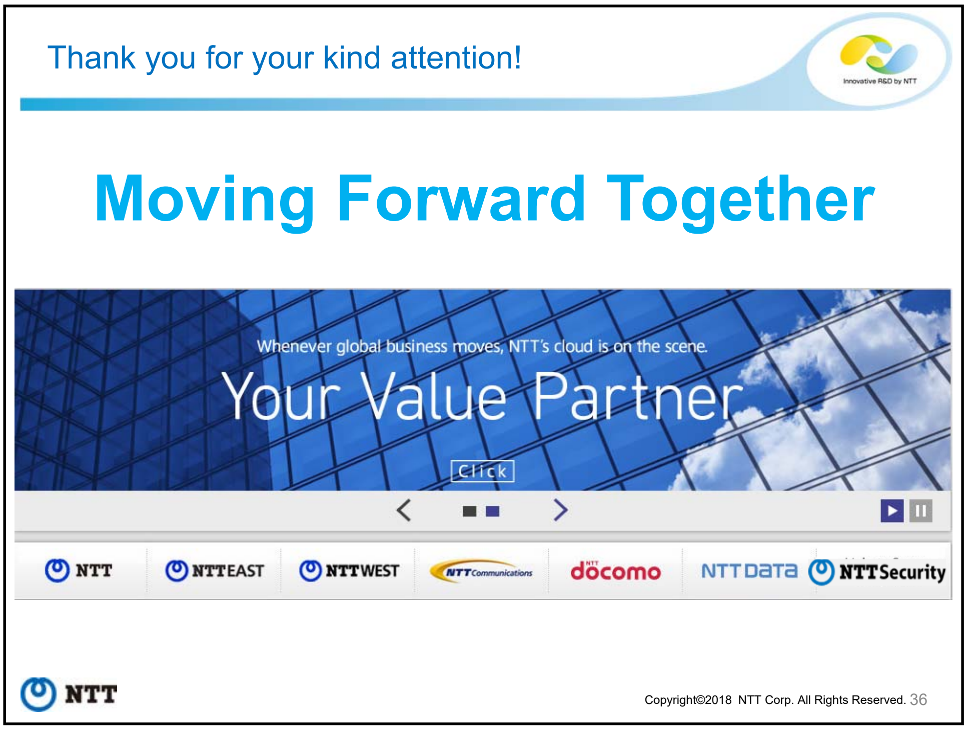
Promote implementation of an AI society by realizing a legal system for the future

*RIKEN Center for Advanced Intelligence Project (AIP) and NTT Labs are dong joint research in this area

Related activity in Japan:
•Japanese Society for Artificial Intelligence "Ethical guidelines" http://ai-elsi.org/archives/471
•MIC "AI Networked Society Promotion Council" http://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/

NTT

Copyright©2018 NTT Corp. All Rights Reserved. 35



Thank you for your kind attention!

Moving Forward Together

NTT

Copyright©2018 NTT Corp. All Rights Reserved. 36

Session 3:
Multimedia Systems

(Chair: Katsuhiko Kaji)

A Method for Video Advertisement Insertion on Smartphone

Yoshia Saito[*]

[*]Faculty of Software and Information Science, Iwate Prefectural University, Japan
y-saito@iwate-pu.ac.jp

Abstract –
In this research, we propose a method for video advertisement insertion on smartphone. The method have an algorithm which estimates a comfort timing to insert a video advertisement using the acceleration data at the time of watching a video. To create the algorithm, we formulated three hypotheses; (1) Viewers who watch video contents on smartphones sitting on chairs change their posture when they take a short breath, (2) The postural change can be detected by analyzing acceleration data from the accelerometer of the smartphone, (3) The timings of the postural changes correspondent with timings of a scene change on the video content and one of the timings is an appropriate timing to insert a video advertisement. We conducted a preliminary experiment and found these hypotheses were true. On the basis of the finding, we proposed an algorithm which estimates a comfort timing using the acceleration data to detect a viewer's postural change. The evaluation results showed accuracy rate of the proposed algorithm was 86% and useful in terms of practical usage.

Keywords: Video Advertisement, Insertion Algorithm, Smartphone, Accelerometer

1 INTRODUCTION

In recent years, video sharing services introduce a business model which inserts video advertisements in their video contents in the same manner as TV. The business model becomes popular with increase of smartphone users. Major video sharing services such as YouTube [1] and niconico [2] provide smartphone applications to users and the smartphone applications insert video advertisements to make a profit. Therefore, many smartphone users have to watch video advertisements on the video sharing services.

There are three types of video advertisements, which are pre-roll, post-roll and mid-roll. The pre-roll video advertisement inserts a video advertisement before video start. The post-roll inserts a video advertisement after video end. The mid-roll inserts a video advertisement viewing a video content like TV commercials and becomes popular in recent years. Adobe reports the mid-roll video advertisement is engaging commercials which have high completion rate [3]. The mid-roll video advertisements will be used even further in the next several years.

We have proposed a video advertisement insertion method which does not interfere with video viewing to make viewers accept the video advertisements [4]. In the previous work, it analyzes characteristics of viewers' comments for the video content. It enables viewers to watch videos more comfortably without feeling of interruption of their video viewing by the video advertisement insertion. However, there are two issues in the previous method. The first issue is that the previous method does not apply to various video sharing services. This is because it needs special kind of viewers' comments with playback time such as comments on the niconico. The second issue is that the previous method has room for improvement of viewers' experience. This is because it does not personalize the timing of video advertisement insertion in spite of difference of the right timings for each viewer.

In order to solve these issues, based on the fact that the number of users watching videos on smartphones has increased rapidly, it is worthwhile considering a method using smartphone sensor information. In this research, we use acceleration information from the accelerometer of the smartphone at the time of watching video contents. We try to find relationship between the acceleration information and appropriate timings for video advertisement insertion. Utilizing the relationship, we propose a method which estimates an appropriate timing to insert a video advertisement for each viewer.

2 RELATED WORK

2.1 Video Advertisement Insertion

There are studies of interactive advertising to provide interactivity to the advertising [5-8]. The interactive advertising allows selecting appropriate ads according to the viewers and changing video length and display methods. Our research is regarded as one of technologies for interactive advertising. Tao Mei et al. [9] proposed a scheme of appropriate video ad insertion for online videos. In this research, the appropriate timing for video advertisement insertion is determined detecting an unattractive video shot boundary. The unattractive video shot boundary is detected by importance of the scene audio-visually. Since this research analyzes video image and audio in detail, the processing cost will be high when it applies to a large number of videos on the video sharing services. Our study aims to find other approaches which estimate the appropriate timing for video advertisement insertion without heavy audio-visual processing.

3 HYPOTHESES FORMULATION

In this section, we indicate possibility that there is a relationship between human motion and degree of interest. We also describe existing techniques of sensing for human motion to choose what sensor is appropriate for estimating human motion on smartphones. Then, we formulate hypotheses in order to create a new method.

3.1 Human Motion and Degree of Interest

There are several techniques for estimating human motion by sensor devices on the smartphone. There are also some studies which show a relationship between human motion and degree of interest.

The relationship between eye motion and degree of interest is well-known [10-13]. The data of eye motion can be acquired by eye-tracking techniques. However, high accurate eye-tracking techniques require special devices for eye tracking or strict restrictions of the measuring environment. It is difficult for smartphone users to prepare for the special devices and strict restrictions force inconvenience upon the users. For these reason, the eye motion is not suitable to estimate viewers' degree of interest on the smartphone.

The relationship between posture and degree of interest is mentioned [14]. People change their posture at intervals from 15 to 20 minutes at the time of sitting because of fatigue [15, 16]. Meanwhile, their postural change hardly occurs when they are interested on something. We can apply this knowledge to a method which estimates an appropriate timing to insert a video advertisement for each viewer if we can detect viewers' postural change by smartphone sensors.

There are a lot of techniques to estimate body motion using sensors. Visual analysis using video data taken by camera devices is one of the techniques. However, usage of camera devices causes large power consumption and it is a disadvantage especially on the smartphone. Visual analysis is not appropriate to estimate body motion. Estimation of the body motion using acceleration information is popular and light-weight techniques [17, 18]. Most of smartphones have accelerometer and many researchers study estimating body motion of the smartphone users from the acceleration information. These research shows various states of smartphone users such as sitting, standing, walking, running, going up and down the stairs and so on can be discriminated. Usage of accelerometer to detect postural change of smartphone users is reasonable.

3.2 Hypotheses

Our previous work shows appropriate timings for video advertisement insertion corresponded with timings of scene changes on the video content. The viewers may change their posture in scene changes on the video content. We formulate three hypotheses as follows.

1. Viewers who watch video contents on smartphones sitting on chairs change their posture when they take a short breath.
2. The postural change can be detected by analyzing acceleration data from the accelerometer of the smartphone.
3. The timings of the postural changes correspondent with timings of a scene change on the video content and one of the timings is an appropriate timing to insert a video advertisement.

If these hypotheses are true, we can create a new method which estimates an appropriate timing for each individual to insert a video advertisement by analyzing acceleration data from the accelerometer of the smartphone.

4 PRELIMINARY EXPERIMENT

We conduct a preliminary experiment to reconfirm if there are difference of right timings to insert a video advertisement for each viewer and test the hypotheses.

4.1 Methodology

We select 5 videos at random from several videos which were used in the previous work. We ask 3 participants for cooperation, who are in their twenties and thirties and have used some video sharing services. At first, we explain about the experiment to the participants. They watch the 5 videos in random order on a smartphone sitting on a chair. We shoot a video to record their watching situation. In the smartphone, an application which records acceleration data from the accelerometer is running. After watching the videos, we carry out a questionnaire survey to ask the participants top 3 comfort timings if a video advertisement is inserted. We explain the participants the length of the video advertisement is about 15 seconds. Then, we interview the participants showing the recorded video. Table 1 shows the overview of the experiment.

Table 1: Overview of the experiment

Participants	3 people in 20s and 30s who have used video sharing services
Smartphone	Nexus 5
Videos	Video 1: 3D action game [19] Video 2: 2D action game [20] Video 3: 3D action game [21] Video 4: 2D action game [22] Video 5: 3D action game [23]
Condition	Sitting on a rotary chair with backrest (seat height: 30 cm)
Procedure	1. Receive an explanation about the overview of the experiment 2. Watch the 5 videos in random order 3. Answer top 3 comfort timings if a video advertisement is inserted 4. Take an interview about the comfort timings.

4.2 Results

At first, we reconfirm whether there are differences of right timings to insert a video advertisement for each viewer or not. Tables 2-6 show the top 3 comfort timings in the 5 videos for each participant. These results show there are differences of the right timing for each participant. We found the participants had characteristic features to select the comfort timings. The participant A tended to select the 1st comfort timing in early scenes. The participant B tended to select the 1st comfort timing in late scenes. We reconfirmed that it was necessary to personalize the timing of video advertisement insertion because there was difference of the right timings for each viewer.

We also checked the recorded video. The participants changed their posture in the scene change and the timings of the postural change matched their comfort timing. The postural changes rapidly increased the resultant acceleration of 3-axis. These results show monitoring the rapid increase of the resultant acceleration can detect postural changes of the viewers and estimate one of their comfort timings to insert a video advertisement in the video content.

Table 2: Comfort timings for each participant in video 1

Video 1

	Participant A	Participant B	Participant C
1st comfort timing	01:47	07:00	01:08
2nd comfort timing	00:09	06:43	01:47
3rd comfort timing	06:30	01:47	06:43

Table 3: Comfort timings for each participant in video 2

Video 2

	Participant A	Participant B	Participant C
1st comfort timing	01:00	09:49	01:23
2nd comfort timing	01:41	10:41	10:41
3rd comfort timing	09:41	05:45	09:41

Table 4: Comfort timings for each participant in video 3

Video 3

	Participant A	Participant B	Participant C
1st comfort timing	01:23	10:50	07:09
2nd comfort timing	02:19	11:11	10:50
3rd comfort timing	07:09	06:14	11:11

Table 5: Comfort timings for each participant in video 4

Video 4

	Participant A	Participant B	Participant C
1st comfort timing	01:39	07:10	01:39
2nd comfort timing	01:44	09:38	07:10
3rd comfort timing	07:10	10:16	03:45

Table 6: Comfort timings for each participant in video 5

Video 5

	Participant A	Participant B	Participant C
1st comfort timing	03:41	10:51	05:44
2nd comfort timing	05:44	06:50	06:50
3rd comfort timing	06:50	11:15	11:15

5 PROPOSED METHOD

The preliminary experiment shows the possibility of estimating a comfort timing for each viewer to insert a video advertisement utilizing their postural change which can be detected by the rapid increase of the resultant acceleration of 3-axis. On the basis of the finding, we create an algorithm for video advertisement insertion.

5.1 Algorithm for Video Advertisement Insertion

Figure 1 shows a flowchart of the algorithm for video advertisement insertion using an accelerometer on the smartphone. In the algorithm, it calculates a test statistic based on acceleration values for outlier detection which means occurring the viewer's postural change. The test static $\mathrm{T_i}$ can be calculated using $\mathrm{Acceleration}_t$, E_t, SD_t at time *t*. $\mathrm{Acceleration_t}$ denotes the resultant acceleration of 3-axis at time *t*. E_t denotes the average of the resultant acceleration from the video start to time *t*. SD_t denotes the standard deviation of the resultant acceleration from the video start to time *t*. T_t is calculated by the following equation.

$$\mathrm{T}_t = |\ (\ \mathrm{Acceleration_t} - \mathrm{E}_t\ )\ |\ /\ \mathrm{SD}_t$$

The outlier can be detected when the following inequality is completed.

$$\mathrm{T}_t > 2 * \mathrm{SD}_t$$

If the outlier is detected, the algorithm waits for reaching a next shot boundary. In the next shot boundary, the video content is stopped and a video advertisement starts. After completion of the video advertisement, the video content restarts and the algorithm terminates the process of the video advertisement insertion.

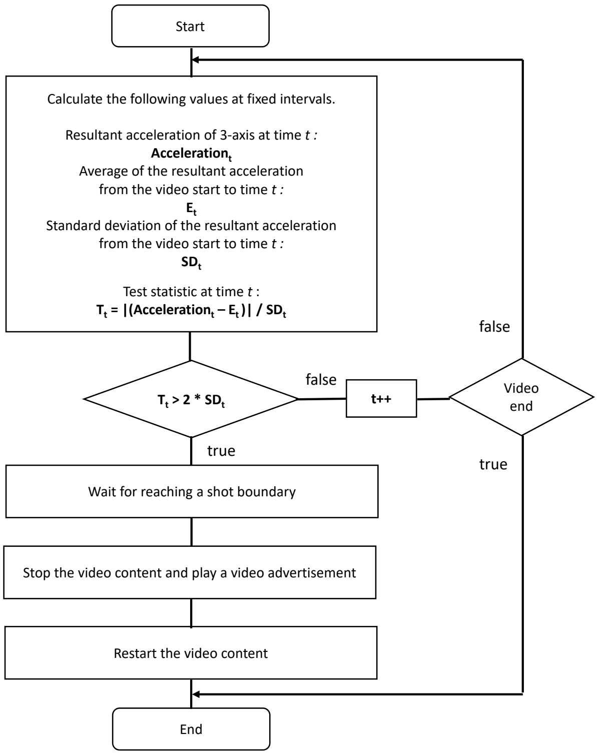

Start
Calculate the following values at fixed intervals.
Resultant acceleration of 3-axis at time t :
Acceleration_t
Average of the resultant acceleration
from the video start to time t :
E_t
Standard deviation of the resultant acceleration
from the video start to time t :
SD_t
Test statistic at time t :
T_t = |(Acceleration_t – E_t)| / SD_t
T_t > 2 * SD_t
false
t++
Video end
false
true
true
Wait for reaching a shot boundary
Stop the video content and play a video advertisement
Restart the video content
End

Figure 1: The flowchart of the algorithm for video advertisement insertion

5.2 System Design

We assume the algorithm for video advertisement insertion is implemented on the smartphone as an application for video viewing. A video advertisement and acceleration data are input to the algorithm. The algorithm outputs less than one comfort timing for each viewer in the video content.

Figure 2 shows the system design of the proposed method. The proposed system have two servers, which are for video sharing and video advertisement. Video uploaders submit their video contents to the video sharing server. Advertising sponsors provide video advertisements to the video advertisement server. Video viewers have smartphones with a smartphone application for video viewing which has the algorithm for video advertisement insertion. The smartphone application plays a video content from the video sharing server. Shot boundaries of the video content are detected by existing techniques for shot boundaries detection and it lies outside the scope of our research. While the video viewer is watching the video content, the algorithm for video advertisement insertion monitors acceleration data of the smartphone. The smartphone application stops the video content temporarily and starts a video advertisement from the video advertisement server when the algorithm estimates the timing is comfort for the viewer. After that, the video content restarts without any more video advertisement in the viewing. Note that we suppose the algorithm works only when the video viewer is sitting on a chair and the viewer watches the video holding the smartphone without video skip. In case of other conditions, new routines should be added to the algorithm.

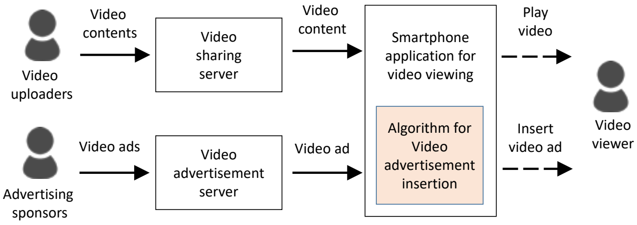


Figure 2: System design of the proposed method

6 EVALUATION

We evaluate the performance of the proposed algorithm for video advertisement insertion using an accelerometer on the smartphone. Comparing the previous algorithm, we verify the proposed algorithm estimates a better comfort timing for each viewer.

6.1 Methodology

The evaluation is conducted in the same manner as the preliminary experiment. We use 5 videos which are same videos in the preliminary experiment. We ask 10 participants for cooperation, who are in their twenties and thirties and have used some video sharing services. At first, we explain about the experiment to the participants. They watch the 5 videos in random order on a smartphone sitting on a chair. We shoot a video to record their watching situation. In the smartphone, an application which records acceleration data from the accelerometer is running. After watching the videos, we carry out a questionnaire survey to ask the participants top 3 comfort timings if a video advertisement is inserted. We explain the participants the length of the video advertisement is about 15 seconds.

After getting data of acceleration and comfort timings, we estimated a comfort timing for each viewer by using the proposed algorithm. We also estimated a comfort timing by using the previous algorithm which used viewers' comments on the niconico. Then, we compared the result of the proposed algorithm with one of the previous algorithm.

6.2 Results

Table 7 shows the result of comparing an estimated timing of the proposed algorithm with top 3 comfort timings for each viewer. Table 8 shows the result of comparing an estimated timing of the previous algorithm with top 3 comfort timings for each viewer. In these tables, “1st” means the estimated timing coincides with the 1st comfort timing for the viewer. The same applies to “2nd” and “3rd”. “n/a” means the estimated timing does not coincide with any top 3 comfort timings. We regard the estimated timing is an appropriate timing if it coincides with one of the top 3 comfort timings.

From Table 7, the proposed algorithm could estimate 43 appropriate timings of the 50 chances. The accuracy rate of the proposed algorithm was 86%. On the other hand, the previous algorithm could estimate only 22 appropriate timings of the 50 chances as shown in Table 8. The accuracy rate of the previous algorithm was 44%. Comparing these results, the proposed algorithm improved the accuracy rate more than 40%. Personalization of a timing to insert a video advertisement contributed the improvement of accuracy rate.

Table 7: Result of comparing an estimated timing of the proposed algorithm with top 3 comfort timings for each viewer

	Video 1	Video 2	Video 3	Video 4	Video 5
Participant A	1st	2nd	1st	2nd	n/a
Participant B	1st	1st	2nd	n/a	n/a
Participant C	1st	3rd	1st	1st	1st
Participant D	3rd	1st	3rd	n/a	1st
Participant E	1st	1st	2nd	1st	3rd
Participant F	1st	1st	1st	2nd	2nd
Participant G	2nd	3rd	1st	1st	n/a
Participant H	1st	2nd	1st	n/a	1st
Participant I	1st	n/a	3rd	2nd	3rd
Participant J	2nd	1st	1st	1st	1st

Table 8: Result of comparing an estimated timing of the previous algorithm with top 3 comfort timings for each viewer

	Video 1	Video 2	Video 3	Video 4	Video 5
Participant A	n/a	n/a	n/a	1st	n/a
Participant B	n/a	n/a	1st	n/a	3rd
Participant C	n/a	n/a	2nd	1st	n/a
Participant D	n/a	n/a	3rd	3rd	n/a
Participant E	n/a	3rd	1st	3rd	n/a
Participant F	n/a	2nd	3rd	n/a	2nd
Participant G	3rd	n/a	1st	2nd	n/a
Participant H	n/a	n/a	1st	n/a	2nd
Participant I	n/a	n/a	n/a	3rd	n/a
Participant J	3rd	3rd	1st	n/a	n/a

7 CONCLUSION

In this paper, we proposed a method for video advertisement insertion using acceleration data on smartphone. From the preliminary experiment, we got 3 findings; (1) viewers who watch video contents on smartphones sitting on chairs change their posture when they take a short breath, (2) the postural change can be detected by analyzing acceleration data from the accelerometer of the smartphone, (3) the timings of the postural changes correspondent with timings of a scene change on the video content and one of the timings is an appropriate timing to insert a video advertisement. Then, we created an algorithm for video advertisement insertion and designed its system. From the evaluation results, we found the proposed algorithm improved the accuracy rate more than 40% comparing with the previous algorithm. This result showed the effectiveness of the proposed algorithm.

For the future work, we will try to study other algorithms even if the viewer does not sit on a chair. We also study a method which switch mid-roll to pre-roll or post-roll video advertisement when there are not clear scene changes in the video content.

REFERENCES

[1] YouTube，http://www.youtube.com/

[2] niconico，http://www.nicovideo.jp/

[3] 2012 adobe digital video advertising report, https://blogs.adobe.com/primetime/files/2013/11/Monetization-Report_FINAL1.pdf

[4] Yoshia Saito: A method for video advertisement insertion with audience comments on action game videos, International Workshop on Informatics (IWIN2017), pp.153-158 (2017).

[5] K. Risden, M. Czerwinski, S. Worley, L. Hamilton, J. Kubiniec, H. Hoffman, N. Mickel and E. Loftus: Interactive advertising: patterns of use and effectiveness, SIGCHI, pp. 219-224 (1998).

[6] Jong Woo Kim and Stephen Du: Design for an Interactive Television Advertising System, Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Vol. 2 (2006).

[7] J. Lloyd: I-Ads - a new approach, European Conference on Interactive Television (2003).

[8] Panagiotis Giotis, George Lekakos: Effectiveness of Interactive Advertising Presentation Models, EuroITV '09, pp.157-160 (2009).

[9] Tao Mei: VideoSense-Towards Effective Online Video Advertising, ACM Multimedia'07, pp.1075-1084 (2007).

[10] Jeffrey Heer and Stuart K. Card: Efficient user interest estimation in fisheye views, CHI '03 Extended Abstracts on Human Factors in Computing Systems, pp. 836-837 (2003).

[11] Anthony Santella and Doug DeCarlo: Robust clustering of eye movement recordings for quantification of visual interest, Proceedings of the 2004 symposium on Eye tracking research & applications, pp. 27-34 (2004).

[12] Tina Walber, Chantal Neuhaus, Steffen Staab, Ansgar Scherp and Ramesh Jain: Creation of individual photo selections: read preferences from the users' eyes, Proceedings of the 21st ACM international conference on Multimedia, pp. 629-632 (2013).

[13] Vanessa Georges, François Courtemanche, Sylvain Senecal, Thierry Baccino, Marc Fredette and Pierre-Majorique Leger: UX Heatmaps: Mapping User Experience on Visual Interfaces, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4850-4860 (2016).

[14] Selene Mota and Rosalind W. Picard: Automated Posture Analysis for detecting Learner’s Interest Level, Computer Vision and Pattern Recognition Workshop (2003)

[15] P.Branton: Behaviour, Body mechanics and Discomfort, Ergonomics, Vol.12, No.2 (1969).

[16] Hidetoshi Watanabe, Masao Ando and Takashi Takahashi: Transition of sitting posture over time, J. Archit. Plann. Environ. Eng., AIJ, No. 474, pp. 107-114 (1995).

[17] Toshiki Iso and Kenichi Yamazaki: Gait analyzer based on a cell phone with a single three- axis accelerometer, ACM the 8th Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI2006), pp. 141-144 (2006).

[18] Ronit Slyper and Jessica K. Hodgins: Action capture with accelerometers, Proceedings of the 2008 ACM SIGGRAPH/Eurographics, pp. 193-199 (2008).

[19] niconico, http://www.nicovideo.jp/watch/sm5457137

[20] niconico, http://www.nicovideo.jp/watch/sm6979644

[21] niconico, http://www.nicovideo.jp/watch/sm2750853

[22] niconico, http://www.nicovideo.jp/watch/sm4895582

[23] niconico, http://www.nicovideo.jp/watch/sm8481759

Experiments to investigate the effects of scents on vection perception

Aoi Aruga[*], Yuichi Bannai[**], and Takeharu Seno[***]s

[*] Graduate School of Engineering, Kanagawa Institute of Technology, Japan
[**] Department of Information Media, Kanagawa Institute of Technology, Japan
[***] Faculty of Design, Kyushu University, Japan
Emails: S1885007@cce.kanagawa-it.ac.jp, bannai@ic.kanagawa-it.ac.jp, seno@design.kyushu-u.ac.jp

Abstract - Vection is an illusion that gives the feeling of motion in the absence of bodily movement. This phenomenon may occur with presentation of a screen displaying patterns with optical flows. In an immersive environment that uses a head mounted display (HMD), including virtual reality systems, vection is frequently induced. Additionally, it has been reported that certain scents can affect the perception of bodily movement and are thought to have some influence on vection. We aimed to investigate the effect of scents on vection perception. Subjects were seated in front of an olfactory display while wearing an HMD and were presented with moving images to induce vection under several conditions: scent presentation, sound presentation, and presentation of no additional stimuli. We found that the scent stimulus did not affect the perception of vection. However, there is a tendency that there is a difference in the stimulus affected depending on the vection direction. Furthermore, it appears that lavender may promote the perception of vection.

Keywords: vection, sense of smell, scent, sense of sight, HMD, olfactory display

1 INTRODUCTION

Vection is the visually induced illusion of self-motion [1], which may be felt when viewing a screen that is displaying patterns with optical flows. It is known that vection is induced not only by visual stimulation but also by auditory and somatosensory stimulation. Sakamoto et al. (2004) reported that sound images with movement from front to back or back to front induces linear self-motion perception and that the self-motion direction is influenced by the direction of the motion of auditory stimuli [2]. This shows that auditory information also has a great influence on self-motion. Vection is also caused by somatosensory sensation. Murata et al. (2014) performed experiments in which participants wore eye masks and were presented with white noise through a pair of earphones. A horse riding machine was used to produce bodily movement in the participant. Almost all the participants in this condition felt a forward motion sensation on presentation of a constant stream of air to their front [3].

In recent years, many trials have been conducted in which scents are displayed in conjunction with movies using virtual reality (VR) systems. In an immersive environment, such as a VR game using a head mounted display (HMD), vection is frequently induced.

Additionally, it has been reported that certain scents affect the movement of the body and are thought to influence vection, where the stimulus causes the sense of motion. An experimental example has been reported, where elderly people, aged 65 years or older, in nursing homes, were presented with the scent of lavender for 12 months. The proportion of people who fell and the incidence rate per person in these participants was lower than those who were not presented with lavender. Moreover, the Cohen-Mansfield Agitation Inventory score, which is used as an indicator of dementia, was also lower in the group that was presented with lavender [4]. Sakamoto et al. (2012) hypothesize that the scent of lavender may reduce the likelihood of falls and agitation in nursing home residents. However, few studies have investigated the relationship between vection and olfactory stimulation.

Subjects were seated in front of an olfactory display while wearing an HMD and were presented with two types of moving images (expansional and contractional optical flow) under the conditions of with and without scents. During the moving image presentation, we measured the latency and duration of vection. After finishing the stimuli presentation, the subjects evaluated the strength of the vection that they experienced using subjective values. This allowed for the examination of the relationship between vection and olfactory stimuli.

2 METHODOLOGY

2.1 Olfactory Display

We used the Fragrance Jet 2 olfactory display (Figure 1). This display uses the techniques of an ink-jet printer to produce a jet, which is broken into droplets with a small hole in the ink tank. Bubbles are formed in the ink by instantaneous heating, and ink is ejected due to the pressure created by the bubbles. The display can set up one ejection head, which can store three small tanks and one large tank; thus, this display can contain a maximum of four kinds of scents. There are 127 minute holes that are connected to the small tank and 256 minute holes in the head that are connected to the large tank. Since the display can emit scent from multiple holes at the same time, the ejection quantity can be set from 0 to 127 (in the small tank) or 0 to 256 (in the large tank). We denote the average ejection quantity at one time from each hole as the “unit average ejection quantity” (UAEQ), and the number of minute holes that are emitting at one time as “the number of simultaneous

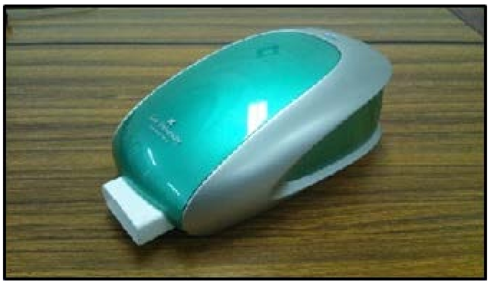


Figure 1: Olfactory Display: Fragrance Jet 2

ejections" (NSE), which we refer to as the ejection level. The UAEQ from one minute hole in the small tank is 4.7 pl and 7.3 pl in the large tank. The reproducibility of these values was confirmed without depending on the residual quantity of ink on examination. As the emission occurs 150 times per unit time (100 ms), the ejection quantity (EQ) in the small tank was calculated as follows:

EQ (pl) = 4.7 (pl)(UAEQ) × (from 0 to 127)(NSE) × 150 (times)

Additionally, the display was equipped with a fan and 9 phases of wind velocity control in the range of 0.8-1.8 m/sec were used.

2.2 Scent Stimuli

Two kinds of scents were used in this experiment: lavender (oil of lavender) and banana (iso-amyl acetate). These scents ware diluted to 5% with ethanol and water, and the component ratios of each perfume are shown in Table 1.

Table 1: Scent stimuli

Scents		Lavender	Banana
Component ratios	Scent (%)	5	5
	Ethanol (%)	65	75
	Water (%)	30	20

To determine the ejection level at which the subjects could sense the scent, we measured the detection threshold, which is the minimum detectable concentration of scent. This method is based on the two-point comparison method that was proposed by the Japanese association of odor environment. The initial ejection level was set to five. If the subject correctly identified the scent twice at the initial level, we reduced the level by two according to the descent method and ended the measurement when the subject could not identify the scent. When the subject was unable to correctly identify the scent at the initial level, we adopted the rising method and increased the level by three. The measurement was complete when the subject identified the scent correctly twice. Based on this result, the ejection level to be used was determined. Table 2 shows the measurement values of six participants (between the age of 20 and 30 years, male).

Table 2: Result of olfactory detection threshold measurement

	Lavender	Banana
Average	2.67	3.67
SD	1.97	2.07
Maximum	5	7
Minimum	1	1

2.3 Vection Stimuli

To induce vection, two kinds of moving images were used. The expansion stimulus was used to induce a feeling of forward movement which is frequently experienced in everyday life and VR games. When a dot reached the edge of the screen and disappeared, a continuous stimulus presentation could be obtained by rearranging the dot from the center of the plane. The contraction stimulus was used to induce a feeling of backward movement. A dot was repositioned from the edge of the plane until it reached the center of the plane and disappeared. In each stimulus, a total of 2400 dots were displayed on the screen. The size of the dot on the screen was changed to be physically constant according to the distance change simulation. As the dots did not have a density gradient and did not give a static depth cue, the depth cues were only provided by motion. The speed of each dot was approximately 3.7° per second at the viewing angle. Figure 2 shows a still image of the vection-inducing visual stimulus, where arrows indicate movement of points.

The visual stimuli were presented using Oculus Rift DK2. The viewing angle was set at 110° in the diagonal direction and 90° in the horizontal direction and the motion stimuli that induced vection was displayed throughout the screen.

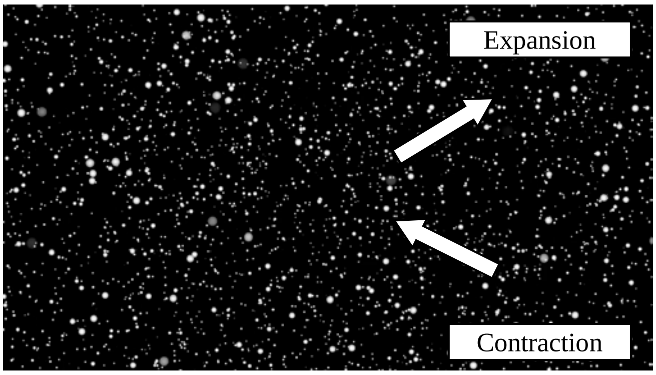


Figure 2: Still image of vection stimulus

2.4 Sound Stimuli

As a control condition, sound stimuli using a pure tone at the frequency of 440 Hz and 0.1amplitude were presented to investigate the influence of olfactory stimuli on vection more clearly. For the sound presentation, a set of sealed dynamic headphones (SE-MJ 522, Pioneer) was used, and the playback volume was set to the sound that would normally be heard through a speaker.

2.5 Experimental Conditions

All experiments were performed while the subject was seated on a chair with their jaw positioned on a chin rest. The subjects wore HMDs and headphones throughout the experiments. The distance from the ejection exit of the head of the olfactory display to the nose was fixed at 225 mm. The olfactory display was set up with the presentation port facing diagonally upwards, and the height of the jaws was adjusted by each subject so that the wind from the display hit their face firmly. Figure 3 shows the actual experimental condition. The wind speed was set to level 9, which is the maximum setting (on a scale of 1 to 9). The measured wind speed under these conditions was approximately 1.8 m/sec. During the experiment, a fan was constantly in operation even when no scents were being emitted. Thus, the risk of a change in perception due to the presence or absence of wind was eliminated.

2.6 Qualification of Subjects

Prior to performing the olfactory experiment, we conducted a test to confirm that the subjects had a general olfactory sense. In this test, we used odor panel analyses to qualify five standard odor solutions and tested whether each odor could be identified. The standard odor for panel selection was prepared on the basis of the T & T Olfactometer, which is used in national examinations to determine olfactory measurement operator [5]. Using odor solutions at the concentration determined by the Ministry of the Environment, the test was conducted according to the 5-2 method. Only subjects who passed the test were taken as subjects.

It is suggested that approximately one in 20 people do not experience vection (vection blind). Therefore, subjects were screened before the experiment was conducted. We asked subjects to observe expansional and contractional moving images. Subjects were then asked to evaluate the vection they felt. In cases where subjects answered that they did not feel vection (intensity 0) in more than half of these tests, they were excluded from the experiment.

3 EXPERIMENT 1: PRESENTATION METHOD OF OLFACTORY STIMULI

It is necessary to examine the ejection method so that the subjects can continue to feel the scent without adaptation for 70 s, including 30 s before the presentation of the vection stimuli. Therefore, a preliminary experiment was conducted with reference to a previous study [6].

3.1 Experimental method

Two kinds of scents were used in this experiment: lavender and banana. Four times the average value obtained in the above detection threshold measurement experiment was set as a reference value 1, and eight times as the average value was set as a reference value 2.

Table 3 shows the respective ejection levels, and as shown in Figure 4, scent ejection with a duration of 0.3 s and an interval of 1.3 s was repeated for 70 s.

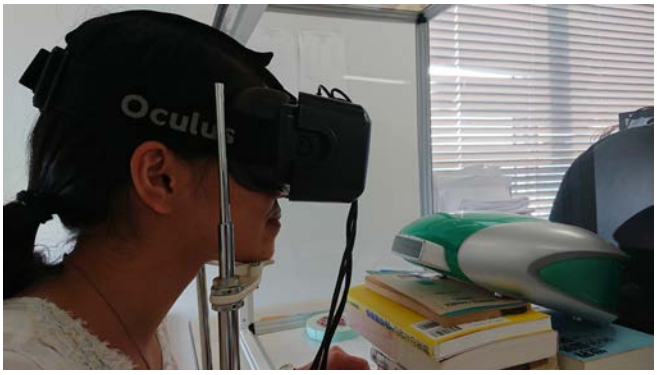


Figure 3: The experimental condition

Table 3: Ejection levels

	Detection threshold	Reference value 1	Reference value 2
Lavender	2.67	11	21
Banana	3.67	15	29

After the stimuli presentation, subjects evaluated their perception of the continuity of scent ejection at the four stages outlined in Table 4 and their perception of the intensity change described in Table 5. We determined a presentation condition that was suitable for this experiment. The selection of the condition was based on obtaining an average value of two or more, regarding the continuity, and at a level where no subjects selected zero regarding the strength.

Measurements were made with a 1-min break every 70 s and the presentation was repeated four times for each of the scents. Taking the order effect into consideration, we determined the first and second scents randomly and adjusted them, so each order was presented approximately an equal number of times. Between the presentations of each scent, there was a 5-min break. Due to the nature of the experiment, it was assumed that the subjects could easily feel the scents; therefore, prior to the first measurement, we confirmed verbally whether each subject could feel the ejection of scents twice. Next, we conducted the experiment with five subjects (between 20 and 25 years of age, male) using the reference value 1, and with five subjects (between 20 and 30 years of age, male) using the reference value 2.

3.2 Experimental Results

Table 6 shows the average evaluation values, standard deviation, and minimum values in the case of reference values 1 and 2 for each scent. As the average value of each cell was 2 or more, it can be stated that the subjects were able to smell the scents without adaptation for 70 s with an ejection duration of 0.3 s and an ejection interval of 1.3 s. However, with reference value 1, the minimum value of the continuity value of lavender and banana was 0, whereas in reference value 2, the minimum value of the continuity and intensity of lavender was 0. The results of this experiment revealed that some of the subjects could not sense the scent for 70 s.

When we interviewed the subjects who rated the continuity as 0, they made the following comments: “the

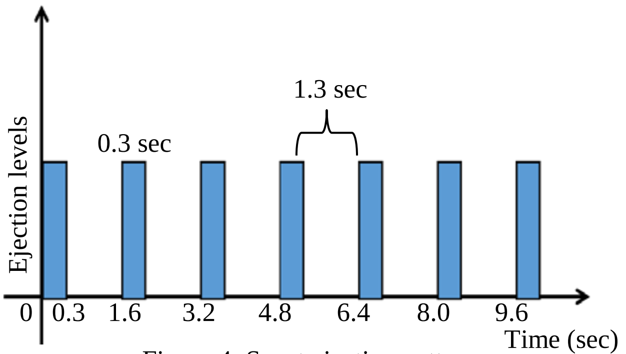


Figure 4: Scent ejection pattern

Table 4: Four-stage evaluation for continuity

0	I did not feel the scent midway through the exposure period
1	I felt it fragmentally
2	I felt it every breath
3	I felt it continuously

Table 5: Five-stage evaluation for the strength

0	I did not feel the scent midway through the exposure period
1	I felt that the concentration gradually decreased
2	I felt that the concentration got increased or decreased
3	I felt that the concentration gradually increased
4	I felt no change in the concentration and felt the scent uniformly

Table 6: Evaluation values for the continuity and the strength of the scent stimuli

		Reference value 1		Reference value 2	
		Lavender	Banana	Lavender	Banana
Continuity	Ave.	2.10	2.00	2.30	2.75
	SD.	0.76	0.52	1.31	0.46
	Min.	0	0	0	1
Strength	Ave.	2.55	2.06	2.20	2.80
	SD.	1.16	1.31	1.85	1.04
	Min.	1	0	0	1

ejection amount in reference value 2 was so large that my nose became paralyzed.” Thus, we asked three subjects (between 20 and 25 years of age, male) to rate the scent intensity with both reference values, using the six-level odor intensity indication method [7] (see Table 7) consisting of odorless (0) to intense odor (5). This technique was devised in the field of odor control in Japan.

Table 7: Six-level odor intensity indication

0	Odorless
1	A faint smell that you can hardly perceive
2	Weak smell you can recognize
3	Easily perceptible smell
4	Strong smell
5	Intense smell

Table 8 shows the average and standard deviation (SD) of the evaluation values. Since the values are between 2 and 3, except for the banana reference value 1, we set the reference value 2 as the ejection level in this experiment so that the subject would feel it in the experiment.

Table 8: Scent intensity

	Reference value 1		Reference value 2	
	Lavender	Banana	Lavender	Banana
Average	2.42	1.92	2..83	2.58
SD	0.79	1.44	0.93	1.08

4 EXPERIMENT 2: EXAMINATION OF THE EFFECTS OF SCENT ON VECTION

In experiment 1, we confirmed the olfactory stimuli presentation method, which enabled the subjects to continue to perceive scent without adaption for 70 s. Using this method, we investigated the influence of scent on vection perception in experiment 2. The subjects wore an HMD and were presented with a moving image that induced vection under scent or sound presentation or movie only condition and evaluated its strength.

4.1 Experimental Method

The vection stimuli presentation time was set to 40 s. The scent and sound stimuli started 30 s before the presentation of the vection stimuli and continued to be presented for 70 s. The flow of one trial is shown in Figure 5. During the period in which moving stimuli were presented, the subjects were asked to report the duration of time for which they were experiencing vection (Report 1). This was reported by pressing the left button of the mouse. At this point, the time required for the first button press was recorded as the latency of vection. The total time during which the button was pressed in the 40 s period was recorded as the duration of vection. After the stimuli presentation was completed, the subjects were asked to report the strength of the vection by rating 0 when they did not feel vection and 100 when they felt vection very strongly (Report 2). These three variables have been repeatedly used in previous vection experiments (e.g., Seno et al., 2013) [8]. Furthermore, in cases where an olfactory stimulus was presented, subjects were also asked to report the subjective intensity value of the scent by selecting from the six-level odor intensity indicator that is displayed in Table 8 after the stimuli presentation was completed (Report 2). Under each condition, the flow of one trial, as shown in Figure 5, was repeated four times with a 1-min break between each trial. The experiment consisted of eight conditions, and each condition was carried out as one block with four consecutive trials.

The order of eight block experiments was randomized for each subject, with a 5-min break between the blocks.

4.2 Experimental Results

The results under expansion and contraction stimuli conditions are shown in Figure 6 and 7, respectively. The

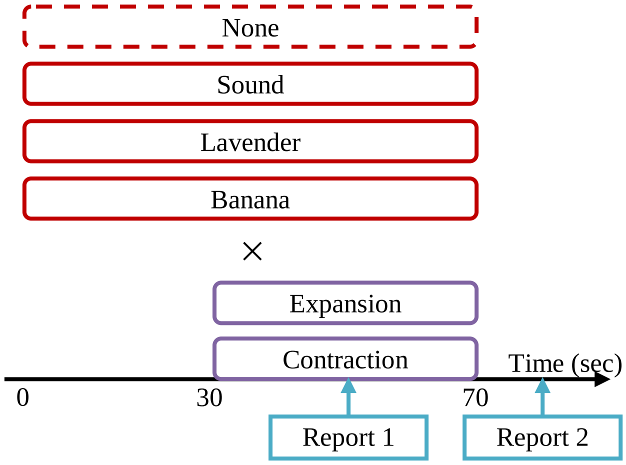


Figure 5: Flow of presenting stimuli

subjective vection strength in each condition was compared. It is known that when vection is strong, the latency is short, the vection duration is long, and the magnitude is high.

In the expansion stimulus condition, the order of the duration of latency was as follows, from shortest to longest: sound, lavender, none, and banana. The order of the duration of vection was as follows, from longest to shortest: sound, none, lavender, and banana. The order of the magnitude of vection, from the highest to the lowest, was as follows: lavender, sound, banana, and none. Therefore, it can be inferred that the sound condition had the greatest effect on the time of vection perceived by subjects, and the lavender condition had the greatest effect on the magnitude of vection. A nonparametric test, using the Friedman method, showed a significant difference only in the latency ($P < 0.01$). Therefore, when multiple comparisons were performed using the Bonferroni correction, the results showed no significant statistical difference ($P > 0.05$).

In the contraction stimulus condition, the order of the duration of latency was as follows, from shortest to longest: lavender, none, banana, and sound. The order of the duration of vection was as follows, from longest to shortest: lavender, none, banana, and sound. The order of the magnitude of vection, from the highest to the lowest, was as follows: lavender, sound, banana, and none. Therefore, it can be inferred that vection was perceived most strongly by subjects in the lavender condition. In addition, the sound stimulation that tended to promote vection perception in the expansion stimulus was the weakest vection promotor in the contraction stimulus condition. Similar to the case of the expansion stimulus, we examined a nonparametric test using the Friedman method, and the results showed that there was no significant difference in the responses to any of the three variables ($P > 0.05$). However, when the significance level was set to 10%, there was a tendency towards a difference in latency and magnitude (both $P < 0.10$). Therefore, while there was no statistical difference, the results indicate that the lavender stimulus tended to promote the perception of vection.

Thus, we could say that there was an interaction between the direction of vection and the types of scent.

5 CONSIDERATION

We examined how olfactory stimuli impact the perception of vection by presenting subjects with visual stimuli while also presenting scents using an olfactory display. No significant effects of scent presentation were observed in the three variables that represented vection intensity. However, there was a difference in the trends between two vection stimuli. In the expansion stimulus, although significant difference was observed at latency ($P < 0.01$), it was not possible to specify which condition was different. The condition that perceived the strongest vection was when sound stimulus was presented. It was lavender at the next point. In the contraction stimulus, there was a tendency towards a difference in the latency and magnitude (both $P < 0.10$). In each variable the condition that perceived the strongest vection was when lavender was presented. In the magnitude, the next point was a sound stimulus, but at the latency, the sound stimulus was the lowest among the four conditions. From the above, it turned out that there is a tendency that there is a difference in the stimulus affected depending on the vection direction. Furthermore, it appears that lavender may promote the perception of vection. There was no effect of scent on vection perception possibly because olfactory stimuli were not sensed as consciously as vision and auditory stimuli. We believe that the visual stimuli were too strong for the scents to be consciously sensed. Alternatively, it was also found that vection stimuli reduced the scent intensity of lavender. There was a difference in the trends between two vection stimuli possibly because the difference in the direction of the obtained self-motion sensation. The expansion stimulus induces a feeling of forward movement which is frequently experienced in everyday life, but the movement in the backward direction obtained from the contraction stimulus less experiences. Therefore, it is thought that under the contraction stimulus, it became more cautious and it became a different perception method than under the expansion stimulus. It is necessary to conduct a more detailed investigation on the possibility that lavender may promote the perception of vection.

6 CONCLUSION AND FUTURE WORK

We investigate the effect of scent on vection perception by presenting subjects with visual stimuli while also presenting them with scents using an olfactory display. The findings were as follows:

(1) Using a scent ejection time of 0.3 s and an ejection interval of 1.3 s, it is possible to present a subject with the scent at a level of about eight times the detection threshold for 70 s without adaptation.

(2) There was no significant difference between the perceived vection in the presence or absence of scent presentation in the three variables representing vection intensity. However, in the contraction stimulus, there was a tendency towards a difference. It seems that there is a difference in the stimulus affected depending on the vection direction.

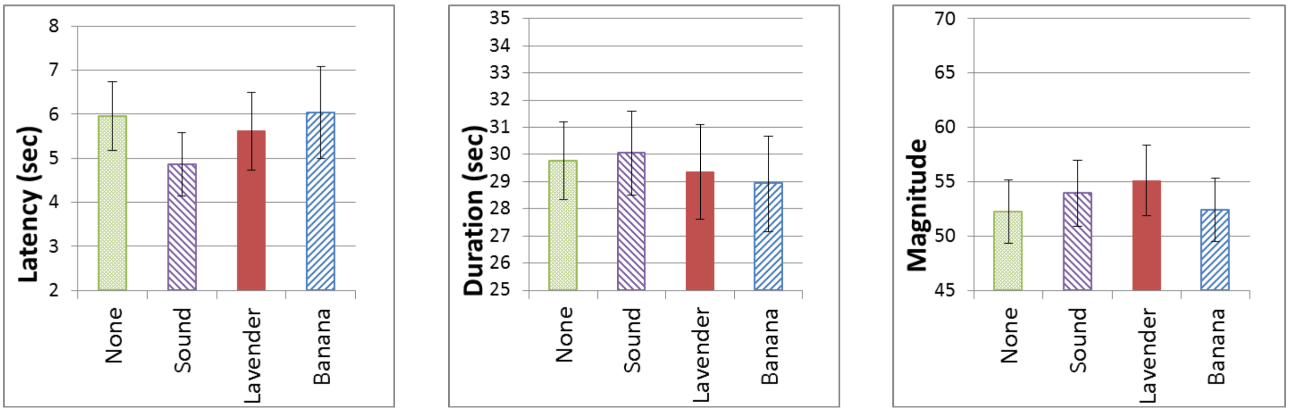


Figure 6: Expansion stimulus results

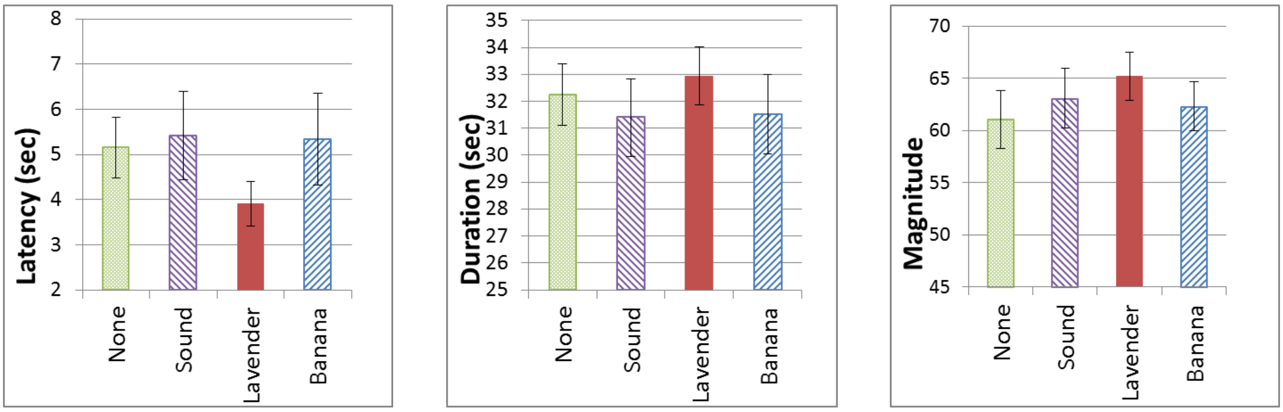


Figure 7: Contraction stimulus results

(3) Many indicators resulted in the strongest perception of the vection when presenting lavender. Lavender scent may promote vection perception.

In the future, the effect of the scent stimuli on vection perception should be further examined. We plan to conduct these experiments using weaker vection stimuli or stronger scent stimuli than was used in the current study. Furthermore, we will investigate the effects of vection stimulation on scent perception, and we would like to clarify the causal relation between the vection and scent perception. We are also interested in exploring the transition of consciousness between visual, auditory, and olfactory senses and the mutual influences between these senses.

REFERENCES

[1] S. Tachi, M. Sto, M. Hirose, Virtual Reality Studies, The Virtual Reality Society of Japan, (2011) (in Japanese)

[2] S. Sakamoto, Y. Osada, Y. Suzuki, & J. Gyoba, The effects of linearly moving sound images on self-motion perception, Acoust. Sci. & tech, Vol.25, No.1, pp.100-102 (2004).

[3] K. Murata, T. Seno, Y. Ozawa, & S. Ichihara. Self-Motion Perception Induced by Cutaneous Sensation Caused by Constant Wind, Psychology, Vol.5, pp.1777-1782 (2014).

[4] Y. Sakamoto, S. Ebihara, T. Ebihara, N. Tomita, K. Toba, S. Freeman, H. Arai, & M. Khzuki, Fall prevention using olfactory stimulation with lavender odor in elderly nursing home residents: A Randomized Controlled Trial, J. Am. Geriatr. Soc., Vol.60, pp1005-11 (2012).

[5] JAOE: Japan Association on Odor Environment, http://orea.or.jp/about/kyukakukensa.html, (2018).

[6] K. Ohtsu, A. Kadowaki, J. Sato, Y. Bannai, & K. Okada, Scent Presentation Method of Pulse Ejection Synchronized with the User’s Breathing, Information Processing Society of Japan Research Report Groupware and Network Services, 2008 (7(2008-GN-066)), pp.77-84 (2008) (in Japanese).

[7] S. Saito, J. Inouchi, S. Ayabe, F. Yoshii, & S. Nakano, Introduction to Olfactory-Foundations of Evaluation of Odorous Environment, Japan Association on Odor Environment, (2014) (in Japanese).

[8] T., Seno, K. Abe, & S. Kiyokawa, Wearing heavy iron clogs can inhibit vection, Multisensory Research, Vol.26, pp. 569-580 (2013).

A Distributed Multi-Viewpoint Internet Live Broadcasting System with Video Effects

Satoru Matsumoto[*] Tomoki Yoshihisa[*] Tomoya Kawakami[**] Yuuichi Teranishi [***]

[*]Cybermediacenter, Osaka University, Japan
[**] Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[***]National Institute of Information and Communications Technology, Japan
{smatsumoto, yoshihisa}@cmc.osaka-u.ac.jp

Abstract - Due to the recent popularization of omnidirectional cameras, multi-viewpoint live videos are often broadcast through the Internet. In multi-viewpoint Internet live broadcasting services, viewers can arbitrarily change their viewpoints. To reduce computational loads for video processing, many researchers have been developed some distributed Internet live broadcasting systems. These systems are designed for single-viewpoint live videos and screen images (images to be watched by viewers) are the same for all viewers. However, in the multi-viewpoint Internet live broadcasting services, screen images differ among viewers since they can change their viewpoints. Here, one of the main research challenges for multi-viewpoint Internet live broadcasting is how to reduce computational loads for adding effects under different screen images.

In this paper, focusing on this challenge, we propose and develop a distributed multi-viewpoint Internet live broadcasting system. It is difficult to determine which is better to add effects on the server side or player side. To determine this to reduce computational loads effectively, we use hierarchical rules.

Keywords: Streaming Delivery; Internet Live Broadcasting; Multi-viewpoint Camera.

1 INTRODUCTION

Due to the recent popularization of omnidirectional cameras, multi-viewpoint live videos are often broadcast through the Internet. In multi-viewpoint Internet live broadcasting services, viewers can arbitrarily change their viewpoints. For example, major live broadcasting services such as YouTube Live or Facebook provide 360 degrees videos, in which we can change our viewpoints. Also in recent Internet live broadcasting services, viewers or broadcasters often add the video or audio effects to broadcast videos. To reduce computational loads for adding effects, many researchers have been developed some distributed Internet live broadcasting systems. ([1] [2]).

These systems are designed for single-viewpoint live videos and screen images (images to be watched by viewers) are the same for all viewers. Therefore, screen images can be shared among processing servers and computational loads can be reduced by exploiting distributed computing systems. However, in the multi-viewpoint Internet live broadcasting services, screen images differ among viewers since they can change their viewpoints. Thus, screen images cannot be shared among processing servers. Here, one of the main research challenges for multi-viewpoint Internet live broadcasting systems is how to reduce computational loads for adding effects under different screen images.

In this paper, focusing on this challenge, we propose and develop a distributed multi-viewpoint Internet live broadcasting system. In our proposed system, video effects that can be shared among viewers are added by some distributed processing servers (on the server side). Video effects that cannot be shared among viewers are added by video players (on the player side). In such systems, it is difficult to determine which is better to add effects on the server side or player side. To determine this to effectively reduce computational loads, we use hierarchical rules. Also, we develop a distributed multi-viewpoint Internet live broadcasting system extending our previously developed system.

The paper is organized as follows. In Section 2, we introduce related work. We design and explain our proposed system in Section 3. We show evaluation results in Section 4 and discuss the results in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORK

Some systems for distributing video processing loads have been proposed. Most of them fix load distribution procedures in advance. However, starting Internet live broadcasting is easy in recent years, and it is difficult to grasp which machines start Internet live broadcastings. Therefore, conventional systems establish load distributions at server sides.

MediaPaaS encodes, re-encodes, and deliver videos using a server machine provided by cloud computing services [2]. Different from MediaPaas, our proposed system establishes load distributions using PIAX [9], a P2P agent platform. In [1], we have confirmed that video processing time such as encoding, distributing videos can be reduced by distributing the processing load to many different world broadcasting servers.

An Internet live broadcasting system that allows viewing recently recorded videos (playback) was developed in [3]. Several methods have been proposed for reducing the delay time to distribute videos of live Internet broadcasting. In live broadcasting using SmoothCache 2.0 [4], by caching video data from other peers and distributing them among cached peers using a P2P network, the communication load and delay times are reduced. Dai, et. al. proposed a distributed video broadcasting system using P2P networks to reduce delay times in [5]. In the HD method proposed in [6], communication traffic is reduced by simultaneously transmitting image data to many viewers by using one-to-

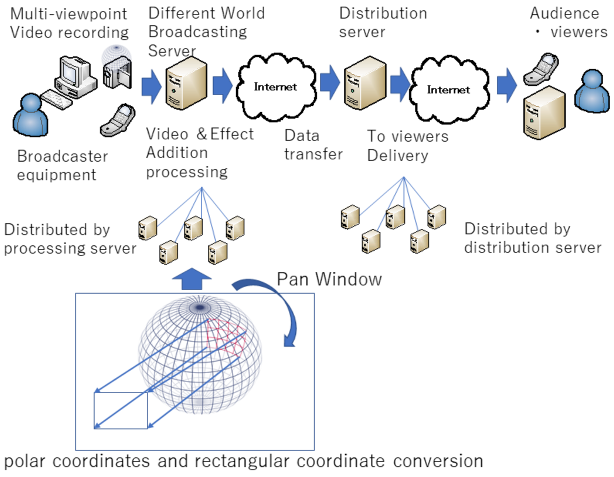

Multi-viewpoint Video recording
Different World Broadcasting Server
Distribution server
Audience viewers
Internet
Internet
Broadcaster equipment
Video ＆Effect Addition processing
Data transfer
To viewers Delivery
Distributed by processing server
Distributed by distribution server
Pan Window
polar coordinates and rectangular coordinate conversion

Figure 1: Our designed distributed multi-viewpoint Internet live broadcasting system.

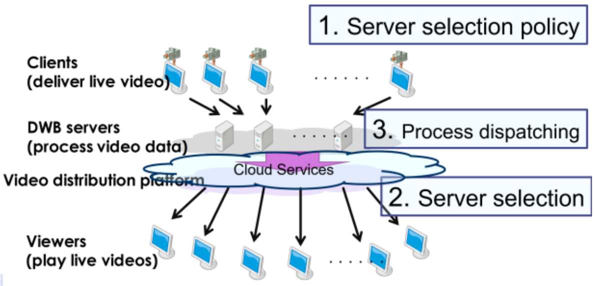


Figure 2: The system architecture of the Different World Broadcasting System

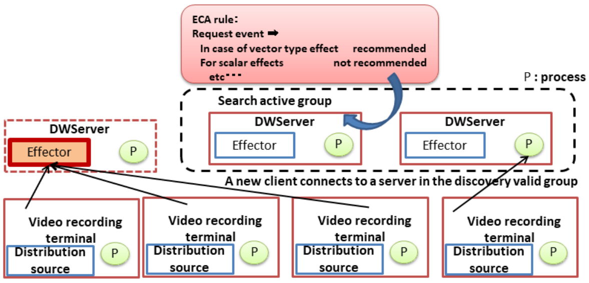


Figure 3: A load distribution mechanism by PIAX

many broadcasting with one-to-one communication. Even in our proposed system, these delay reduction methods can be applied when delivering videos, but our research considers video or audio effect additions.

Gibbon, et. al. proposed a system that performs video processing by transferring video data captured by a camera to a computer with high processing capability in [7]. Ting, et. al. proposed a system that directly stores images captured by computers with low processing power in external storage devices such as cloud storage in [8]. However, these systems target stored video data and can not be applied for live Internet broadcasting.

3 DISTRIBUTED INTERNET LIVE BROADCASTING SYSTEM

In this section, first, we explain our previously developed cloud-based live broadcasting system using ECA rules. After that, we explain the multi-viewpoint Internet live broadcasting system proposed in this paper.

3.1 Different World Broadcasting System

3.1.1. Summary of Different World Broadcasting System

In our previous research [1], we constructed a different-world broadcasting system using virtual machines provided by cloud service. These machines work as the different world broadcasting servers that add video effects. In general, many virtual machines can be easily used in a cloud service. Therefore, the use of multiple virtual machines as different-world broadcasting servers should enable high-speed effect-addition, while distributing the load among different-world broadcasting servers. Therefore, we implemented a distributed live Internet broadcasting system using the cloud service and evaluated its performance. In our developed system, video effect addition is executed on the virtual machines provided by the cloud service shown in Figure 1.

The clients select a server considering the load distribution. In conventional systems, load distribution is established by connecting processing servers via a load balancing mechanism such as a load balancer. In this method, when the load distribution mechanism needs to switch to another server while the video is being transmitted, the connection is interrupted. For this reason, it is difficult to smoothly switch servers while continuing the video distribution. Therefore, in our system, the load balancing mechanism selects a different world broadcasting server based on the requests.

3.1.2. System Architecture

The system architecture of the different world broadcasting system is shown in Figure 2. There are three types of machines. One is the clients which have cameras and record live videos. Another one is the different world broadcasting servers which execute processes for videos such as encoding, decoding, or video effect additions. The final one is the viewers which play live videos. Each client selects a different world broadcasting server that executes the desired video effect and transmits the video effect library and the recorded video to the different world broadcasting server. The different world broadcasting server is a virtual machine of the cloud service executes video processing on the video transmitted from the clients according to their requests. The video processed by the different world broadcasting server is delivered to the viewers via the video distribution service. In the system, the viewers receive the processed video after selecting the server or the channel of the video distribution service.

3.1.3. Load Distribution Mechanism

Figure 3 shows the load distribution mechanism of our developed system. Clients software and the client side PIAX

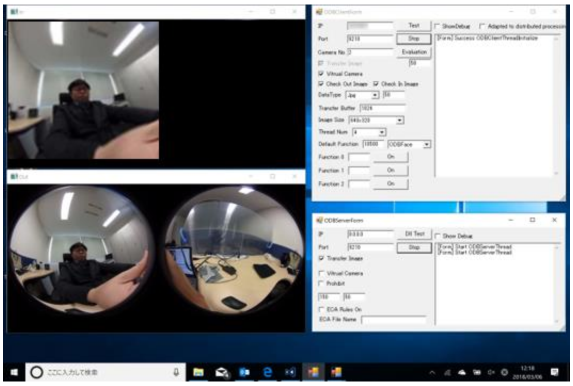
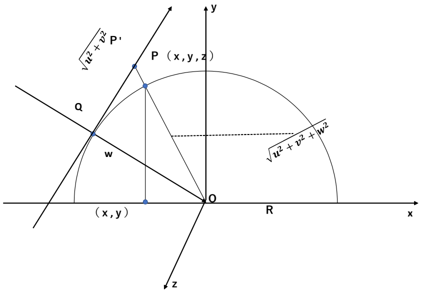


Figure 4: Server software and Client software



$$|OP'|^2 = |QP'|^2 + |QO|^2 = (u^2 + v^2) + w^2$$

$$OP'/OP = \sqrt{u^2+v^2+w^2}\,/\,R$$

$$P'(x',y',w') = \frac{\sqrt{u^2+v^2+w^2}}{R}(x,y,z) \qquad (1)$$

Figure 5: An Image of coordinate conversion

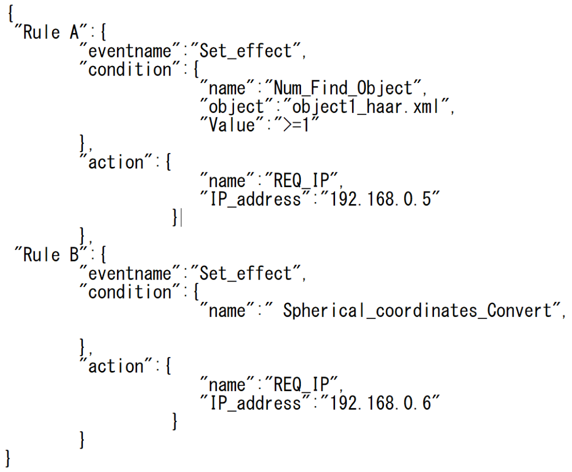
system are installed in the client. The different world broadcasting server software and the server side PIAX system are installed on the different world broadcasting servers. PIAX [9] is a Java-based platform middleware that enables efficient server resource search using resource search function of the overlay network. PIAX is provided as open source software. The PIAX system of the clients and the different world broadcasting servers connect with each other via the overlay network.

3.2 Extension of Different World Broadcasting System

In this section, first, we explain our proposed design of ECA rules for multi-viewpoint videos. After that, we explain the examples of hierarchical ECA rules.

3.2.1. Image Conversion of Multi-Viewpoint Videos

With omnidirectional cameras, it is not necessary to take dozens of omnidirectional images from a certain viewpoint and synthesize them on a computer to create a panoramic image. Instead, we can create multi-viewpoint videos from panoramic images. The lower left part of Figure 4 shows two panoramic images (front and back) for a multi-viewpoint video. These panoramic images were obtained from cameras using fisheye lenses. It is necessary to convert these images into a planar image. There are many methods that obtain wide images from car-mounted fisheye lenses and correct the distortion [11]. Figure 5 shows how to obtain a wide image from a panoramic image in our proposed system. As shown in this figure, the wide images are obtained by assuming an imaginary hemispherical border for the panoramic images. The converted wide image is shown in the upper left part of Figure 4. The conversion transforms virtual hemispherical polar coordinates into rectangular coordinates using equation (1). In our proposed system, the distributed processing of polar/rectangular coordinates.



```
{
 ″Rule A″:{
        ″eventname″:″Set_effect″,
        ″condition″:{
                    ″name″:″Num_Find_Object″,
                    ″object″:″object1_haar.xml″,
                    ″Value″:″>=1″
        },
        ″action″:{
                    ″name″:″REQ_IP″,
                    ″IP_address″:″192.168.0.5″
                    }|
        },
 ″Rule B″:{
        ″eventname″:″Set_effect″,
        ″condition″:{
                    ″name″:″ Spherical_coordinates_Convert″,

        },
        ″action″:{
                    ″name″:″REQ_IP″,
                    ″IP_address″:″192.168.0.6″
                    }
        }
}
```

Figure 6: Examples of ECA rules

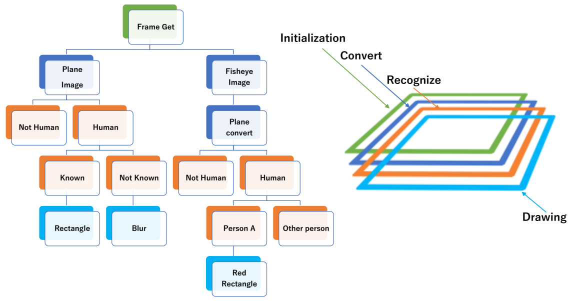


Figure 7: Examples of hierarchical ECA rules

3.2.2. Design of ECA Rules for Multi-Viewpoint Videos

Video effects have procedures, and for example, the face detection process is generally performed before the mosaic effect of the detected face. The "Timer" or "Message" function of ECA rules in the proposed system can define such a procedure. If the procedure is defined in the ECA rules and the ECA rules are order dependency, the system needs to execute the rules according to the sequence. Otherwise, in cases that the ECA rules do not depend on the processing request, the system can execute the rules concurrently, and the processing time can be shortened compared with the ECA rules having the order dependent relation. In the current system, it is impossible to process

ECA rules in parallel. Parallel processing for cloud computing services is a future task. Lists of events, conditions, and actions are described in the previous research [1]. Figure 6 shows an example of ECA rules. In this example, the servers with IP addresses of 192.168.0.5 and 6 are assigned as initial values to video processing requests from the video recording terminals named "Num_Find_Object" and "Spherical_coordinates_Convert".

In cases that processes of ECA rule have sequences, the system should execute the processes in the order of the sequence. Otherwise, the system can execute them in parallel. For example, Figure 7 shows an example of hierarchical ECA rules.

- 1. Is it a fisheye lens image? <- Perform full spherical coordinate transformation <- Human detection.
- 2. Are Humans in image <- Who? <- Match with a specific person <- Blur is applied.
- 3. Are Humans in image <- a Known person registered in DB? <- If you are an unregistered person, blur.

ECA rules are classified into hierarchies of detection, conversion, inquiry, and pixel processing.

3.2.3. Implementation of Proposed System

We developed a distributed live Internet broadcasting system using Microsoft Azure as a cloud service. The different world broadcasting servers run on the virtual machine provided by the Azure service. Each virtual machines is logically connected through a virtual network (VNet) which is one of the services provided by Microsoft Azure. Figure 4 shows a screenshot of server software and the client software. When starting the video effect adding process, it provides an interface of the different world broadcasting server software. Figure 4 shows a screenshot of client software for distributing video to a client. We can visually check the result of applying the selected.

In the client software holds the IP address of a different world broadcasting server to request video processing. If the "Apply distributed processing" checkbox in the client software dialog box is checked, the client software requests the different world broadcasting server to execute the video processing specified by the pull-down menu the initial IP address.

4 EXPERIMENTAL EVALUATION

We evaluated the performances of our implemented system built on the virtual machine provided by the Microsoft Azure service. The results follow.

4.1 Evaluation System

In the experiment, Theta S made by RICHO Co., Ltd. was used as an omnidirectional camera. Each video frame is encoded in jpeg format, transmitted and received, as USB virtual camera. Image conversions and rule processing are realized by different world broadcasting. In the evaluation, we measured the turnaround time from the time to generate original image data to the time to obtain processed image data. We will confirm the reduction of the turnaround times. In order to confirm the efficiency of the proposed system, the video processing time including the processing time of the ECA rule and the turnaround time was measured as evaluation indexes.

We experimented with the evaluation considering the following two points.

(1) The turnaround time was compared and evaluated when the video effect processing requests ware concentrated on a video effect processing server without using the ECA rule.

(2) The image effect processing requests do not concentrate, and the ECA rule was adopted. For evaluation, we assigned multiple computers with the same performance parameters.

To select an available different world broadcasting servers, we used the PIAX overlay network. When a different world broadcasting server becomes overloaded, the server sends a notification to the PIAX process on the server side and waits for decreasing the loads. The turnaround time of the evaluation was measured in two cases. One is a concentrated case where three clients request video processing to one of three different world broadcasting servers. The other case is a completely distributed case where each of the three client requests services to three different servers.

We perform live broadcasting in this evaluation. For the evaluation, as the video effect described in the ECA rule, a process in which face detection video processing has been applied after the above coordinate conversion is executed. The time required for sending and receiving frame data was defined as turnaround time.

4.2 Evaluation Environment

In this evaluation, a different world broadcasting server runs on a virtual machine provided by the Microsoft Azure service. Table 1 shows the specifications of the virtual machine and OS. We used five different virtual machines for different world broadcasting servers. Open CV parallelized by Intel 's Parallel Computing Library TBB [10] was used as a library for executing video processing on a different world broadcasting server. The clients are a PC installed at Osaka University. Table 2 shows the specifications of the client PC. We attached a full omnidirectional camera to only one. These PCs communicate with different world broadcasting servers via different home optical networks and avoid the congestion of the network.

Table 1: Specifications of Microsoft Azure Virtual Machines

OS	Microsoft Windows Server 2016
Microsoft Azure Plan	Standalone Server Microsoft Corporation Virtual Machine x64-based PC
CPU	Intel E5-2697 v3 Equivalent 2.4GHz
Main memory	3.584MB

Table 2: Specifications of Client PC

OS	Client PC 1～3	Microsoft Windows 10 Pro Version 1709,1511
CPU	Client PC1	Intel i7-7660U Equivalent 2.5GHz
	Client PC2	Intel i5-6300U Equivalent 2.4GHz
	Client PC3	Intel i3-4020Y Equivalent 1.5GHz
Main memory	Client PC1	8.00 MB
	Client PC2	8.00 MB
	Client PC3	4.00 MB

4.3 Evaluation Results

Processes were assigned based on ECA rules among different world broadcasting servers.

Figure [8, 9] show the evaluation results of the turnaround time under the evaluation environment described in Section 4.2. The horizontal axis shows the recorded frame the number and the vertical axis shows the turnaround time. In Figure 8 where the load is concentrated on a single different world broadcasting server, turnaround times increase gradually. In Figure 9, the image processing requests are distributed to three different world broadcasting servers. In this case, the turnaround time is under 1500 msec, and the processing delay is around 7500 msec during processing.

The different world broadcasting server that PC 2 requests image effect processing is a VM in the East Japan region in the real environment of Microsoft Azure. Therefore, there were variations in the route and the turnaround time changes.

We also measured the turnaround time required to contact the recommended different world broadcasting server by PIAX. The average time to process a query for getting recommended different world broadcasting server be 16.28 [msec].

As a result, we confirmed that the processing request is allocated to the different world broadcasting server based on the ECA rule, and the load is distributed. Moreover, we confirmed that the turnaround time might fluctuate even if the hardware performance of the virtual machine is equivalent due to the effect of communication delay, etc.

5 DISCUSSION

In past paper [1], the video processing was detecting the face of people in the video with the specified effect described in the ECA rule. Turnaround time fluctuation was confirmed among cloud computing service virtual machines. This is caused by factors such as actual server performance fluctuations due to differences in the cloud environment of the network distance. Therefore, when the user configures the system, it is recommended to consider these problems.

In this paper, we have proposed hierarchical rules of three stages. The stages follow: Prepare rules for selecting the location of three types of processing, processing by a local client, processing by edge computing, processing by cloud computing. An image of the hierarchical rules is shown in Figure [10, 11] shows an example of rules proposed in this

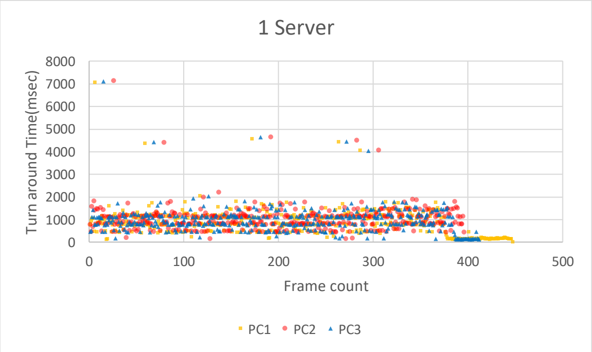


Figure 8: Turnaround times under one cloud server

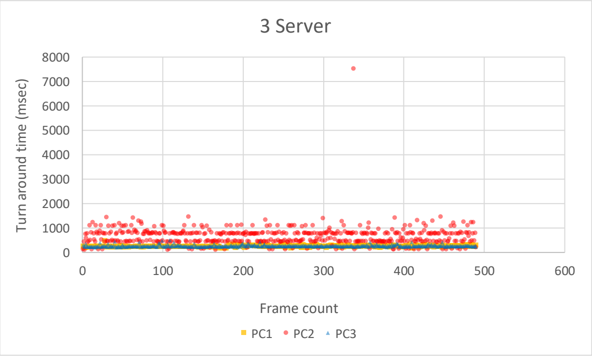


Figure 9: Turnaround times under three cloud servers

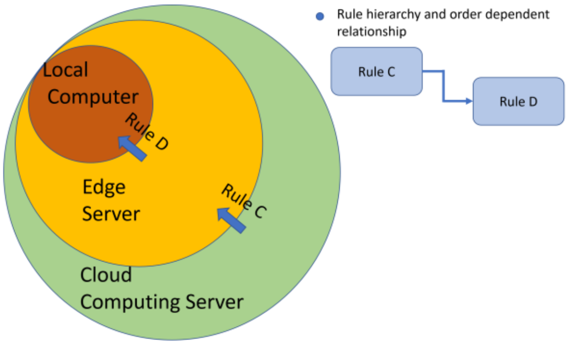
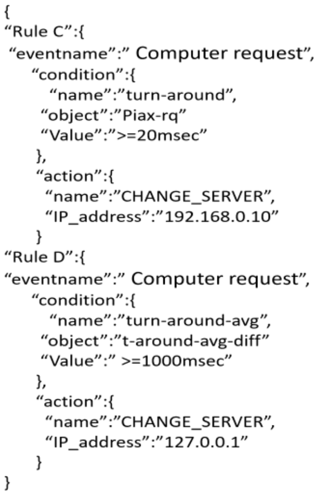


Figure 10: An image of the hierarchical rules



```
{
“Rule C”:{
 “eventname”:” Computer request”,
     “condition”:{
        “name”:”turn-around”,
      “object”:”Piax-rq”
      “Value”:”>=20msec”
     },
     “action”:{
       “name”:”CHANGE_SERVER”,
       “IP_address”:”192.168.0.10”
     }
“Rule D”:{
“eventname”:” Computer request”,
     “condition”:{
        “name”:”turn-around-avg”,
      “object”:”t-around-avg-diff”
       “Value”:” >=1000msec”
     },
     “action”:{
       “name”:”CHANGE_SERVER”,
       “IP_address”:”127.0.0.1”
     }
}
```

Figure 11: An example of our proposed rules

paper. In this rule, the different world broadcasting server that adds the video effects changes when the performance of the current server decreases.

6 CONCLUSION

In this research, we implemented and evaluated a multi-viewpoint distributed live Internet (different world) broadcasting system that adds various image processing using the computer provided by cloud service. By describing processing assignment with ECA rule, we can flexibly assign video processing to appropriate virtual machines. In the developed system, the PIAX platform was used to search and communicate among virtual machines. Even if the number of virtual machines changed, video processing could be continuously allocated during live Internet broadcast. System evaluation confirmed that the turnaround time of video processing could be shortened by using the proposed system.

In the future, we plan to exploit edge computing environments in which computers on the edge of the Internet can execute video processes. The processing time can be reduced since the turnaround times of edge computers are short.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers JP15H02702 and JP18K11316, and by Kayamori Foundation of Informational Science Advancement.
Also this work includes the result of NICT Osaka University joint research “research and development of advanced network platform technology for large scale distributed computing.”

REFERENCES

[1] Satoru Matsumoto, Yoshimasa Ishi, Tomoki Yoshihisa, Tomoya Kawakami, and Yuuichi Teranishi, “Different Worlds Broadcasting: A Distributed Internet Live Broadcasting System with Video and Audio Effects,” in Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA 2017), pp. 71-78 (2017).

[2] Satoru Matsumoto, Yoshimasa Ishi, Tomoki Yoshihisa, Tomoya Kawakami, and Yuuichi Teranishi, “A Design and Implementation of Distributed Internet Live Broadcasting Systems Enhanced by Cloud Computing Services,” in Proc. of the International Workshop on Informatics (IWIN 2017), pp. 111-118 (2017).

[3] Y. Gotoh, T. Yoshihisa, H. Taniguchi, and M. Kanazawa, “Brossom: a P2P streaming system for webcast,” Journal of Networking Technology, Vol. 2, No. 4, pp. 169-181 (2011).

[4] R. Roverso, R. Reale, S. El-Ansary, and S. Haridi, “Smooth-Cache 2.0: CDN-quality adaptive HTTP live streaming on peer-to-peer overlays,” Proceedings of the 6th ACM Multi-media Systems Conference (MMSys 2015), pp. 61-72 (2015).

[5] J. Dai, Z. Chang, and G.S.H. Chan, “Delay optimization for multi-source multi-channel overlay live streaming,” Pro-ceedings of the IEEE International Conference on Commu-nications (ICC 2015), pp. 6959-6964 (2015).

[6] T. Yoshihisa and S. Nishio, “A division-based broadcasting method considering channel bandwidths for NVoD services,” IEEE Transactions on Broadcasting, vol. 59, no. 1, pp. 62-71 (2013).

[7] D. Gibbon and L. Begaja, “Distributed processing for big data video analytics,” IEEE ComSoc MMTC E-Letter, vol. 9, no. 3, pp. 29-31 (2014).

[8] W.-C. Ting, K.-H. Lu, C.-W. Lo, S.-H. Chang, and P.C. Liu, “Smart video hosting and processing platform for Internet-of-Things,” Proceedings of the 2014 IEEE International Conference on Internet of Things (iThings 2014), pp. 169-176 (2014).

[9] M. Yoshida，T. Okuda，Y. Teranishi，K. Harumoto，S. Shimojyo，“PIAX: A P2P Platform for Integration of Multi-overlay and Distributed Agent Mechanisms，” Transactions of Information Processing Society of Japan / Information Processing Society of Japan，Vol. 49, No. 1, pp. 402-413, (2008).

[10] Thread Building Blocks, https://www.threadingbuildingblocks.org/, (referred October 1, 2017).

[11] J. Jeong, H. Kim, B. Kim, S. Cho, “Wide rear vehicle recognition using a fisheye lens camera image” 2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 691-693 (2016)

Session 4:
Network and Security

(Chair: Kozo Okano)

Detection of Malware-infected Hosts with Data Compression Algorithm

Yasushi Okano, Kazunori Kamiya, Atsutoshi Kumagai, Taishi Nishiyama,
Bo Hu, Masaki Tanikawa, Kazuhiko Ohkubo

NTT Secure Platform Laboratories, Japan
{okano.yasushi, kamiya.kazunori, kumagai.atsutoshi, nishiyama.taishi,
hu.bo, tanikawa.masaki, ohkubo.kazuhiko}@lab.ntt.co.jp

Abstract

Machine learning is being actively used to detect malware-infested hosts and their malicious communications. When applying machine learning, designing the right feature is the key for accurate detection. BoW (Bag of Words)-based feature extraction is widely applied in natural language processing and also utilized for malicious communication detection. However, BoW-based feature extraction does not always scale for handling network logs that often have new data sequences. By focusing on the fact that new data sequences in network logs are in many cases mostly similar but partly different, we propose a new detection method based on a data compression algorithm. Since the compression algorithm has a characteristic that data size after compressing is related to similarity of data, a compression algorithm based feature can be utilized for classification. According to our evaluation results with real-field proxy logs in an enterprise network, the proposed method has better at detection than a BoW-based detection method. In particular, its true positive rate (TPR) in a low false positive rate (FPR) area (0.5%) is over 30% higher than that for the BoW-based method. In addition, the results show that the proposed method effectively detects an infected host communicating with malicious URL that includes partially modified string from original malicious logs.

1 Introduction

Malware is becoming more sophisticated and has so many variants that anti-virus software does not detect all of them. In fact, it is reported that over 127.5 million pieces of malware were registered in 2016 [1]. To compliment the fact that detection at the endpoint is not always successful, network log analysis is one solution that monitors logs taken from network devices such as proxy and firewall and finds malicious communications derived from infected hosts.

In current log analysis, many monitoring rules have been deployed. Examples include network scans being detected once the number of different destination IP addresses from one source IP address exceeds a predefined threshold and a specific malware infection being determined if one host accesses a blacklisted URL. Many monitoring rules are based on operators' elaborations on creating rules, deciding thresholds, and maintaining blacklists. However, as malware has evolved to become able to change communication patterns easily, heuristic rule creation adds cost to operations and has difficulty catching up with malware modification. As a result, machine learning is gaining attention for automatically detecting evolving malware and for helping operations.

In applying machine learning for network log analysis, first, machine learning calculates feature values from network logs. For instance, feature values range from the length of a string, frequency of terms in device logs, and so on. Second, machine learning will classify the data into legitimate or malicious on the basis of feature values. In this process, infected hosts and malicious communications are detected.

There are many detection algorithms from LR (Logistic Regression), SVM (Support Vector Machine), RandomForest, and DNN (Deep Neural Network). However, the most critical factor for accurate detection is designing the right feature for the problem.

BoW (Bag of Words)-based features are widely applied in natural language processing and also utilized to detect malicious communications. However, BoW-based feature extraction does not always scale for handling network logs, which often have new data sequences.

By focusing on the fact that new data sequences in network logs are in many cases mostly similar but partly different, we propose a compression algorithm based feature and apply it to supervised learning for detecting malicious communications and infected hosts.

Simply put, a compression algorithm based feature is one form of the compression rate of data, which means how small the data becomes after being compressed. When the data sequence is similar to that in existing malicious data, the compression rate should be small because this data sequence is effectively compressed. On the other hand, if another data sequence is totally different from that in existing malicious data, the compression rate should be large. In this sense, the compression rate can be useful for finding if one data sequence is similar to that in malicious data. Consequently, the compression rate can contribute to detecting malicious communications.

We evaluated the proposed method with real proxy logs taken from an enterprise network. The results show that the proposed method is better at detection than the BoW-based method. In particular, the results show that the proposed method effectively detects an infected host communicating with a malicious URL that includes a partially modified string from original malicious logs.

Overall, our research makes three contributions.

1. We first apply the compression algorithm feature to supervised machine learning to detect malicious communications and infected hosts.

2. We evaluate the proposed method with real enterprise proxy logs and demonstrate that the proposed method performs better than a BoW-based classifier.

3. We analyze true positive and false negative use cases and clarify that the proposed method effectively detects partially modified strings from original malicious logs.

2 Related Work

2.1 Classification based on Compression Algorithm

Benedetto et al. [2] proposed relative entropy. Although patterns of the same consecutive code or similar repeated code are effectively compressed, patterns of different code are not. Relative information volume of data sequence x against data A is linked to how well data is compressed. Based on this observation, Benedetto et al. define relative entropy as how well new data x will be compressed with existing data A. Consequently, this is formulated as follows.

$$C_A(x) = Z(A\ cat\ x) - Z(A) \tag{1}$$

where Z is the function to output the data size after compression, and cat is the function to concatenate the first and second data sequences. Sometimes, normalized relative entropy is also used, which is defined to divide relative entropy by the size of data x.

Relative entropy has been applied to classification problems in several research areas [3–7]. To classify data x into group A and B, data x is normally classified into the more similar group. Relative entropy can be used as one index of expressing similarity; when relative entropy with group X is small, data x is similar to group X.

Bratko et al. [5] applied relative entropy to classify spam e-mails. They reported that it was more accurate than BoW based classification.

Nishida et al. [6] introduced a smoothing parameter and set the score in accordance with the following equation to classify malicious tweets from twitter logs.

$$Score = \frac{C_A(x) + \gamma}{C_B(x) + \gamma} \tag{2}$$

where γ is a smoothing parameter that should be set large to alleviate the impact of minor letters appearing a few times in a data string. Data x is classified as A if the score is small and B otherwise. This scoring technique enables us to apply a compression algorithm for a classification problem of comparably long data. Nishida et al. [6] also demonstrated that classification of twitter logs with this scoring mechanism has better accuracy than feature extraction with morphological analysis and classification with a CW(Confidence-Weighted linear classification) method [8].

Different compression algorithms are used depending on their purposes. It is reported that LZSS (gzip), LZW (compress), PMP (rar) are applicable for text data [3–6]. Adachi et al. [7] reported that bzip is applicable for music pieces.

2.2 Method of extracting feature from URL string

The BoW method is widely used to extract features from strings. BoW decomposes string text into words by separation of letters or morphological analysis and then generates each word as a one-dimensional feature. Since a URL is deemed as a one text string, BoW features can be extracted. Kumagai et al. [9] proposed BoW-based feature generation to apply LR supervised learning with L1 regularization and demonstrated that their method has better area under curve (AUC) than blacklist based detection.

Nelms et al. [10] proposed describing a URL attribute with a regular expression and applying unsupervised learning to generate a malware-specific URL access template. By comparing a target URL and the above template, the method successfully detects malware communication even when malware slightly modifies its access pattern.

In the security context, on the basis of knowledge on malware analysis, many kinds of statistical features have been proposed [11] such as the length of a URL and ratio of vowels in a URL.

3 Proposal

We propose applying a compression algorithm based feature to apply supervised learning to detect malicious URLs and infected hosts. Since a large part of malware uses HTTP as a communication protocol with C2 servers, it can be mixed with normal Web access and is hard for operators to distinguish. Thus, in our research, we focus on analyzing HTTP proxy logs and detecting malicious URLs to find infected hosts.

An important observation on malware communication in HTTP is that they tend to access C2 servers with a slightly modified URL string in order to slip through blacklist-based detection with minimum engineering effort. In this case, simple blacklist matching does not catch up with malicious URLs since malware may have various URL access patterns even if its modifications are small. In this sense, we expect the compression algorithm based feature to correctly describe the similarity between a slightly modified malicious URL and a known malicious URL.

To the best of our knowledge, our proposal is the first to apply a compression algorithm to detect malicious communication URLs and infected hosts. In addition,

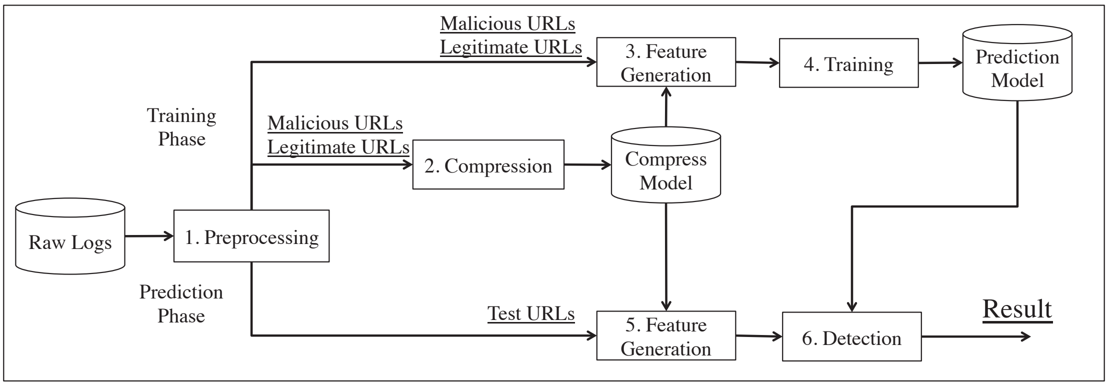


Figure 1: Overview of proposed method

Table 1: Dataset

	Number of hosts	Number of logs	Collection Period	Log Type
Malicious Logs	71,310	7,152,479	Feb. 2015 - Jul. 2015	Sandbox logs
Legitimate Logs	1,940	36,581,398	Feb. 2014 - Mar. 2014	Proxy logs in enterprise

our method is different from existing compression algorithm based methods in that we use a compression algorithm-based score as a feature in supervised learning and the feature can be combined with other features. Furthermore, our research considers a URL structure that has many kinds of attributes such as FQDN, PATH, and QueryString to generate a multi-vector compression algorithm feature for each attribute.

Figure 1 shows an overview of the proposed method. The flow of our proposal is as follows.

1. Input raw logs and execute preprocessing to obtain malicious URLs, legitimate URLs and test URLs

2. Compress malicious URLs and legitimate URLs to generate compress model

3. Input malicious URLs and legitimate URLs with application of compress model to generate compression algorithm features, namely Z_{pos}(Malicious Compression Rate) and Z_{neg}(Legitimate Compression Rate). Z_{pos} and Z_{neg} are defined in equation (3) and (4) respectively. Features are calculated for each attribute of a URL.

4. Train classifier with compression algorithm features and generate prediction model

5. Generate compression algorithm features from test URLs with application of compress model

6. Detect malicious URLs and infected hosts with application of prediction model

As with preprocessing, suitable data must be selected in machine learning for correctly estimating a classifier's performance. We execute two-phase cleansing in this process. First, we delete duplicate URLs in legitimate and malicious logs. This is because hosts may access the same URLs repeatedly. To correctly estimate a classifier, we leave first-to-appear logs in a dataset and eliminate duplicate logs.

Second, we eliminate URLs included in both malicious and legitimate logs, since having the same logs in both datasets may degrade the classifier's performance. In fact, there are many cases in which the same URLs are included in both logs. For instance, some service URLs are automatically accessed from specific applications installed in many environments. Search engine URLs are also often accessed from infected hosts for connectivity checks and included in malicious logs.

As for the compression algorithm features, we define Z_{pos} and Z_{neg} as follows.

$$Z_{pos}(x) = \frac{C_{pos}(x) + \gamma}{L(x) + \gamma} \tag{3}$$

$$Z_{neg}(x) = \frac{C_{neg}(x) + \gamma}{L(x) + \gamma} \tag{4}$$

where C_{pos} and C_{neg} are relative entropy between data x and malicious log(pos) or legitimate log(neg), L is data size of x, and γ is a smoothing parameter.

4 Evaluation Method

4.1 Dataset

The dataset used for all evaluations is shown in table 1.

Malicious logs are taken from an in-house sandbox [12] where we run over 70K malware downloaded on a daily basis from a malware-sharing site and collect pcaps to extract URL information. Legitimate logs are taken from real-environment proxy in an enterprise network.

4.2 Evaluation Indices

Evaluations are executed on the basis of a holdout test that uses previous data in time series as the training dataset and evaluates with later data in time. Evaluation indices are AUC, partial AUC ($pAUC$) [13], and true positive rate $(TPR)_{0.5\%}$ [14].

AUC is the area under the curve drawn on a 2D surface of a false positive rate (FPR) and TPR by changing the score threshold. $pAUC$ is the area under the curve of a limited range of a FPR $[p1, p2]$. Considering the TPR as a function having a FPR as a variable, AUC and $pAUC$ are defined as follows.

$$AUC = \int_0^1 TPR \; dFPR \tag{5}$$

$$pAUC = \int_{p1}^{p2} TPR \; dFPR \tag{6}$$

Through our evaluation, we set $[p1, p2] = [0, 0.1]$.

$TPR_{0.5\%}$ is the TPR value for a low FPR, specifically $FPR = 0.5\%$. In security operations, a low FPR is crucial since the final judgment is done by operators. $pAUC$ and $TPR_{0.5\%}$ are important indices to estimate detection capability with a low FPR.

4.3 Comparative Evaluation

We evaluate the proposed method in comparison with the conventional BoW-based detection method. First, we compared detection capabilities of the proposed and conventional methods. We also measured detection accuracy over time to find out how fast trained models deteriorate. Computational efforts are another important factor for practical use, so we measure CPU time and memory usage of the proposed and BoW method. BoW of a URL is extracted by setting $/, ?, =, \&$ as a separator and splitting the URL. In this evaluation we set the classifier as SVM, the compression algorithm as LZT [15], and TPR/FPR as a per host-basis calculation.

5 Evaluation Result

5.1 Comparative Evaluation

Table 2 shows the evaluation results for the proposed and BoW-based classification methods. From these results, the proposed method has better $TPR_{0.5\%}$ and $pAUC$ than the conventional BoW-based classification method.

Figure 2 shows the $TPR_{0.5\%}$ deterioration over time where the vertical axis is $TPR_{0.5\%}$ and the horizontal axis is time in weeks. This figure shows that $TPR_{0.5\%}$ gradually decreases over time. However, the proposed method always achieves a higher $TPR_{0.5\%}$ than the BoW method until 14 weeks have past.

Table 3 and 4 show comparison with BoW method on CPU time and memory usage, respectively. Proposed method consumes most of CPU time for compression process and its time is longer than any other process of BoW method. Still, once compression is completed, feature generation, training and detection are finished with less CPU time than BoW method. As for memory usage, proposed method consumes small memory for compression process and less memory for feature generation, training and detection compared with BoW method. Although BoW method generates one-hot vector for every single word appeared in URL so that memory usage tends to increase, proposed method generates compression algorithm feature vector in several dimensions so that memory usage does not steeply increase.

Table 2: Evaluation with conventional BoW-based classification method

Method	AUC	pAUC	TPR$_{0.5\%}$
Proposed	0.9306	0.0825	65.30%
BoW	0.9030	0.0657	32.00%

Table 3: CPU Time Comparison with Conventional Method (seconds)

Method	Compress	Generate Feature	Train	Detect
Proposed	13,809	3,752	22	438
BoW	-	7,145	950	408

Table 4: Memory Usage Comparison with Conventional Method (MB)

Method	Compress	Generate Feature	Train	Detect
Proposed	2,401.0	15,999.8	4.7	424.4
BoW	-	56,988.2	1,241.2	529.4

6 Consideration

We consider the reason the compression algorithm feature contributes to better classifying malicious and legitimate logs. Figure 3 shows the histogram of Z_{pos} of URL attributes for both malicious and legitimate logs, where the red and blue zone are histograms of malicious and legitimate logs, respectively. The histogram of malicious logs contains three peaks: A) the compression rate is very small, B) the compression rate is as high as that for legitimate logs, and C) the compression rate is very high.

A sample URL that belongs to pattern A is shown in table 5. For security reason, FQDN is masked with 'www.example.com' and QueryString values are masked with meta words. The first row shows the original URL string and its length, the second row shows the LZT compressed state and relative entropy with malicious logs, and the third row shows that with legitimate logs, where

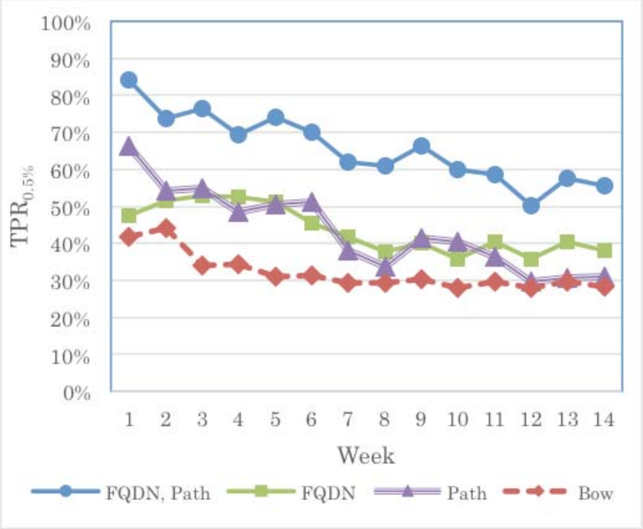


Figure 2: *TPR* deterioration over time

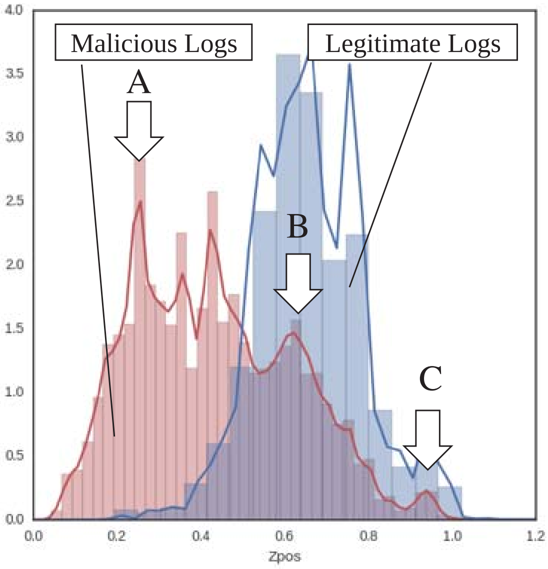


Figure 3: Histogram of Z_{pos}.

’|’ shows that data sequences between ’|’ marks are expressed in 1 code. In compressing with malicious logs, the table shows that a 1,352-bit-long URL is compressed to 220 bits and many data sequences are expressed as 1 code. Especially in QueryString of URL, almost one key (e.g. ”dstid=1”) or one combination of a key and value (e.g. ”countryid=...”) is compressed as 1 code. This observation suggests that a QueryString key and value combination that exists in the training dataset is automatically recognized and compressed as 1 code. In contrast, a key and value combination that does not exist in the training dataset is automatically split. This is one use case that QueryString key exists but its value is modified in malware communication.

Other examples of pattern A for the FQDN attribute are FQDNs having sequential numbers in host names such as host1.example.com and host2.example.com. These FQDNs are recognized as totally different strings by exact matching, but in the compression algorithm that has the characteristic of longest matching, two FQDNs are recognized as similar strings. In fact, host2.example.com is compressed as |host|2.|example.com| after training data host1.example.com. This is another use case that FQDN is partially modified to similar FQDN.

Table 5: Effectively classified URL and compression state

<table>
<tr><td>Uncompressed URL: 1352bit:</td></tr>
<tr><td>http://www.example.com//offers/DynamicOfferScreen?offerid=foo&distid=bar&leadp=baz&countryid=qux&sysbit=quux&dfb=corge&hb=grault&isagg=garply&version=waldo&external=fred&external=plugh&</td></tr>
<tr><td>Relative Entropy for Legitimate Logs: 900bit</td></tr>
<tr><td>|http://www.example.com/|/of|fers|/D|yna|mic|Off|erS|cre|en|?o|ffe|rid=foo|&dis|tid=b|ar|&le|adp|=ba|z&c|ountry|id=qu|x&s|ys|bit|=quux&|dfb|=corge&|hb=|grault&is|agg|=garply&ver|sion=|waldo|&ex|ternal|=fred&e|xter|nal|=plugh&|</td></tr>
<tr><td>Relative Entropy for Malicious Logs: 220bit</td></tr>
<tr><td>|http://www.example.com//offers/DynamicOfferScreen?offerid=foo|&distid=b|ar|&leadp=baz&|countryid=qux&sysbit=quux&df|b=corge&hb=grault|&isagg=garply&versio|n=wal|do&e|xternal=fred&e|xternal=plugh&|</td></tr>
</table>

Table 6: Poorly classified URL and compression state

<table>
<tr><td>Uncompressed URL: 4688 bits</td></tr>
<tr><td>http://www.example.com/api/vp/1?clk=gLg_PHWA9aSyioXkt-F4b3J9cI1ybf-t-x7VxWH5dmAWXwln-z...(omit)</td></tr>
<tr><td>Relative Entropy for Legitimate Logs: 4420bit</td></tr>
<tr><td>|http://www.example.com/api/|vp|/1|?cl|k=|gLg|_PH|WA9|aS|yio|Xkt|-F4|b3J|9cI|1y|bf-|t-|x7|VxW|H5d|mAW|Xwl|n-|z···(omit)</td></tr>
<tr><td>Relative Entropy for Malicious Logs: 4980bit</td></tr>
<tr><td>A|http://www.example.com/api/v|p/1|?cl|k=|gL|g_|PH|WA|9a|Sy|io|Xk|t-|F4|b3|J9|cI|1yb|f-|t-x|7V|xWH|5d|mA|WX|wl|n-|z···(omit)</td></tr>
</table>

URLs belongings to pattern B tend to have same FQDNs existing both in malicious URLs and legitimate URLs. Since Path of these URLs are different, compression rate does not get so small against both malicious logs and legitimate logs and makes detection difficult.

A typical URL belongings to pattern C is shown in table 6. This URL has an encoded or encrypted string. The second row shows compression results of the URL. It shows that the compression rate becomes large for both malicious and legitimate logs and makes classification difficult.

7 Conclusion

We proposed a novel method for detecting malicious communication of infected hosts by generating a compression algorithm feature of URL attributes and classifying with supervised learning. Through evaluation, we demonstrated that the proposed method has higher detection capability than the conventional BoW-based detection method. In particular, its *TPR* in a low *FPR* area (0.5%) is over 30% higher than that of the BoW-based method. In addition, we clarified how the compression algorithm works in classification and demonstrated a real use case in which the proposed method detected malicious URL strings that are similar to but slightly different from existing malicious URLs.

REFERENCES

[1] `https://www.av-test.org/fileadmin/pdf/security_report/AV-TEST_Security_Report_2016-2017.pdf`, "SECURITY REPORT 2016/17", AV-TEST, 2017.

[2] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto, "Language trees and zipping", Phys. Rev. Lett., vol.88, 2002.

[3] Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana, "Towards Parameter-Free DataMining", KDD, pp. 206-215, 2004.

[4] Yuval Marton, Ning Wu, and Lisa Hellerstein, "On compression-based text classification", European Conference on Information Retrieval, pp. 300-314, 2005.

[5] Andrej Bratko, Gordon V. Cormack, Bogdan Filipič, Thomas R. Lynam, and Blaž Zupan, "Spam filtering using statistical data compression", Journal of Machine Learning Research, vol.7, pp.2673-2698, 2006.

[6] Kyosuke Nishida, Ryohei Banno, Ko Fujimura, and Takashide Hoshide, "Tweet-Topic Classification using Data Compression", DBSJ Journal, Vol.10, No.1, pp.1-6, 2011.

[7] Hanae Adachi, Masayuki Okabe, and Kyoji Umemura, "Recognition of Music Composer with Compression-based Dissimilarity Measure", DEIM2013.

[8] Mark Dredze, Koby Crammer, and Fernando Pereira, “Confidence-weighted linear classification”, Proceedings of 25th International Conference on Machine Learning, pp.264 - 271, 2008.

[9] Atsutoshi Kumagai, Yasushi Okano, Kazunori Kamiya, and Masaki Tanikawa, "Supervised Classification for Detecting Malware Infected Host in HTTP Traffic and Long-time Evaluation for Detection Performance using Mixed Data", IEICE-ICSS, Vol.116, No.522, pp.43-48, 2016.

[10] Terry Nelms, Roberto Perdisci, and Mustaque Ahamad, “ExecScent: mining for new C&C domains in live networks with adaptive control protocol templates”, 22nd USENIX Conf., pp.589 - 604, Aug. 2013.

[11] Karel Bartos, Michal Sofka, and Vojtech Franc, “Optimized invariant representation of network traffic for detecting unseen malware variants", in USENIX Security Symposium, pp. 807 - 822, 2016.

[12] Kazufumi Aoki, Takeshi Yagi, Makoto Iwamura, and Mitsutaka Itoh, “Controlling malware http communications in dynamic analysis system using search engine”, Cyberspace Safety and Security (CSS), 2011 Third International Workshop onIEEE, pp.1 - 6 2011.

[13] Yasushi Okano, Atsutoshi Kumagai, Masaki Tanikawa, Yoshito Oshima, Kenji Aiko, Kazumitsu Umehashi, and Junichi Murakami, "Proposal of selection of training data using misdetected goodware for preventing misdetection of a static detector of malware", IEICE-PRMU, Vol.115, No.224, pp.163-170, 2015.

[14] L. E. Dodd and M. S. Pepe. “Partial AUC estimation and regression”, Biometrics, 59(3), pp.614 - 623, 2003.

[15] Peter Tischer, “A modified Lempel-Ziv-Welch data compression scheme”, Aust. Comp. Sci. Commun. 9, 1, pp.262-272, 1987.

Detection of Malware Infection using Traffic Models based on the Similarity between Malware Samples

Masatsugu Ichino[†], Yuuki Mori[†], Mitsuhiro Hatada[‡], and Hiroshi Yoshiura[†]

[†]Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan
[‡]NTT Communications, Japan
ichino@inf.uec.ac.jp, y-mori@uec.ac.jp, m.hatada@ntt.com, yoshiura@uec.ac.jp

Abstract - New types of malware are appearing every day, and malware attacks have become an urgent problem. Current methods of detecting malware use malware signatures, which need to be identified and registered in advance. However, the daily appearance of new types of malware makes such identification and registration impractical. A more practical approach is to identify malware on the basis of traffic behavior since each malware type displays a unique behavior. We have developed a method for detecting malware infection using traffic models based on the similarity between traffic of malware samples. Malware-infected traffic is divided into clusters on the basis of traffic behavior, and a model representing each cluster is created. These models are used to identify target traffic samples as infected or normal. This method should enable the detection of infection caused by a new type of malware if the malware's traffic behavior is similar to that represented by one of the models. Simulation evaluation demonstrated that the proposed method can effectively identify malware-infect traffic with high accuracy.

Keywords: security, malware, malware detection, traffic, clustering

1 Introduction

New types of malware are appearing every day, and malware attacks have become an urgent problem. Current methods of detecting malware use malware signatures, which need to be identified and registered in advance. However, the daily appearance of new types of malware make such identification and registration impractical. A more practical approach is to identify malware on the basis of traffic behavior. We have developed a method for detecting malware infection using traffic models based on the similarity between traffic of malware samples.

Since each malware type displays a unique behavior, we propose malware infection detection using the traffic models based on the similarity between malware samples. In the method we developed, there are three basic steps: create feature values by considering time series of the traffic data, cluster malware samples by considering similarities between them, and create a representative model for each malware cluster. Malware-infected traffic is divided into clusters on the basis of traffic behavior. A model of each cluster is created, and the models are used to identify target traffic as infected or normal. With this method, it should be possible to detect infection caused by a new type of malware if the resulting traffic behavior is similar to that represented by one of the models.

This paper is organized as follows. Section 2 introduces related work. Section 3 describes the proposed method. Section 4 describes the evaluation method. Section 5 presents and discusses the key results. Section 6 concludes the paper with a brief summary of the key points and a mention of future work.

2 Related work

There have been various studies of malware detection using traffic data. Some used the definitions provided by security vendors for detecting malware infection, and some did not.

Studies in the first group (e.g., [1][2][3]) classified malware traffic samples into groups on the basis of the definitions and created models of infected traffic for each group. However, security vendor definitions are not always based on the characteristics of infected traffic. It is thus better to create models on the basis of the characteristics of infected traffic.

Some studies in the second group ([4][5]) created models of infected traffic on the basis of the characteristics of infected traffic but did not consider the time series of the traffic data and the similarities between malware samples. Other studies ([6][7]) created models of infected traffic on the basis of the characteristics of infected traffic considering similarities between malware samples but did not consider the time series of the traffic data. Still other studies ([8][9]) created models of infected traffic on the basis of the characteristics of infected traffic considering the time series of the traffic data but did not consider the similarities between malware samples.

Traffic data is a stream of network information, and previous studies have demonstrated the effectiveness of considering the time series of the traffic data. Consideration of the similarities between malware samples is also necessary for representing common characteristics of infected traffic.

Therefore, in our study, we create feature values by considering the time series of each malware traffic sample. Next, we divided the malware samples into clusters on the basis of their similarities. Then, we created models representing the common characteristics of the infected traffic for each cluster.

3 Proposed method

3.1 Create feature values by considering time series of traffic data

To create a feature vector representing the time series of the traffic data, the time series is divided into 1-s time slots, as shown in Figure 1. The time slots are then grouped into intervals lasting a defined number of seconds for analysis.

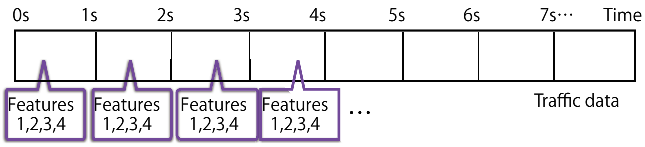


Figure 1: Division of time series of traffic data into 1-s time slots

Dividing time-varying traffic into time slots with a fixed duration and monitoring that traffic in units of time slots enables normal and infected traffic to be distinguished by focusing on the overall temporal variation in that traffic. In this work, we set the time-slot width to 1 s and determined the features for every time slot. The feature values are extracted from each time slot, and a vector concatenating the feature values is created for each time slot. In this study, we used minimum packet size per time slot, number of SYN packets per time slot, ratio of SYN packets to TCP packets per time slot, and number of ACK packets per time slot as the feature values.

3.2 Cluster malware samples by considering similarities between malware samples

To represent the traffic data as a code sequence, extracted features are clustered (in this study, we used the LBG + splitting vector quantization algorithm [10] to do this), and code is created for each cluster. The distances between the feature vector of target time slot and the code for each cluster is calculated, and a search is made for the nearest code. The time slot is then shifted, and a search is again made for the nearest code. This series of nearest codes is called a "transition pattern." An example transition pattern is shown in Figure 2.

The number of occurrences of each transition pattern per time interval (for example 40 s) is counted, and the ratio of each target transition pattern to all types of transition patterns is calculated, as shown in Figure 3. The time interval is then shifted, and the number of occurrences of each transition pattern per time interval is again counted, and the ratio of each target transition pattern to all types of transition patterns is calculated.

To evaluate the similarities between two malware traffic samples, their correlation coefficient is calculated using the occurrence frequency ratios, like those shown in Figure 3. The correlation coefficient represents the correlation between each sample's digital sequence of the number of transition patterns × the number of time intervals. The coefficient is calculated using

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2},} \tag{1}$$

where x and y are the probability variables, $\bar{x}$ is the mean value of x, $\bar{y}$ is the mean value of y, n is number of transition patterns × the number of time intervals.

Calculation of the correlation coefficient requires that the n of x equals the n of y. However, each malware traffic sample has a variable number of time intervals because each malware sample has a variable number of infected time intervals. The number of time intervals is thus adjusted by applying dynamic programming matching (DP matching) to the digital sequences of the two samples. The correlation coefficient is calculated using the adjusted digital sequences. DP matching adjusts the time lengths of the two samples by considering the time-series information and stretching the parts that are similar between the samples.

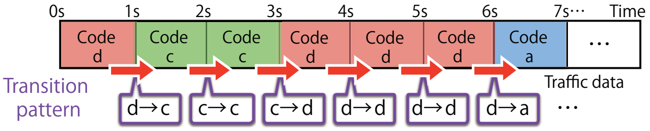
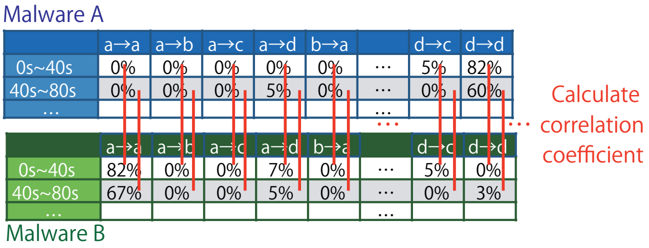


Figure 2: Example transition pattern


Malware A
Calculate correlation coefficient
…
Malware B

	a→a	a→b	a→c	a→d	b→a		d→c	d→d
0s~40s	0%	0%	0%	0%	0%	…	5%	82%
40s~80s	0%	0%	0%	5%	0%	…	0%	60%
…						…		

	a→a	a→b	a→c	a→d	b→a		d→c	d→d
0s~40s	82%	0%	0%	7%	0%	…	5%	0%
40s~80s	67%	0%	0%	5%	0%	…	0%	3%
…						…		

Figure 3: Example occurrence frequency ratios

3.3 Create representative model for each malware cluster

A representative model is created for each malware cluster by extracting a representative malware sample.

The malware samples are classified using hierarchical clustering based on correlation coefficients. Since the optimal number of clusters is unknown in advance, hierarchical clustering is used as it does not require advance setting of the number of clusters. We extracted the malware sample that had the most malware's traffic samples with a correlation coefficient greater than the threshold(upper threshold) as the initial representative malware sample for the cluster. By the same token, the malware samples for which the correlation coefficient between the two malware traffic samples is less than the threshold (lower threshold) are deselected in each cluster to remove the samples for which the correlation is weak. This clustering is repeated until all training samples are divided into clusters.

The extracted malware traffic sample for a cluster is used as a representative model for that cluster in order to model sequential traffic data that actually occurred.

4 Evaluation

4.1 Experimental datasets

We used the anti-Malware engineering WorkShop (MWS) Datasets [11] for our evaluation. In particular, we used the CCC DATAset and the D3M Dataset for training. As malware-infected traffic data, we used 317 malware samples (151 from CCC DATAset 2010, 156 from CCC DATAset 2011, and 10 from D3M 2012) for the training data such that the data used for training were older than the data used for testing. The normal traffic data used for training were captured between 2011 and 2015.

We also used the CCC DATAset and D3M Dataset for testing. As malware-infected traffic data, we used 200 malware samples (177 from CCC DATAset 2011, 15 from D3M 2013, 5 from 2014, and 3 from D3M 2015) for the testing data such that the data used for testing were newer the data used for training.

The CCC 2010 and CCC 2011 attack data include communications prior to malware infection. Thus, given the purpose of our evaluation, we extracted from this attack traffic only the traffic following malware infection using the method described by our group et al. [6].

4.2 Experimental methods

To evaluate the effectiveness of the proposed method, we compared its detection performance with that of three reference methods.

4.2.1 Detection using proposed method

We performed six basic steps .

(Step 1) Extract features of training data

Step 1-1 We divided the traffic data into 1-s time slots. We used the packet header information because the payload information was often encrypted.

Step 1-2 From each time slot, we extracted four features that we had determined to be effective for infection detection: minimum packet size per time slot, number of SYN packets per time slot, ratio of SYN packets to TCP packets per time slot, and number of ACK packets per time slot. The four features are evaluated as effective features for infection detection in [6].

Step 1-3 We normalized the extracted feature values by using the min-max method.

Step 1-4 We represented the time-slot information as a four-dimension feature vector by concatenating the normalized values.

(Step 2) Create codebook for training data

Step 2-1 We applied the LBG+splitting vector quantization algorithm to the vectors with the cluster number set to four.

Step 2-2 We calculated a representative vector (cluster center) for each malware cluster and collected the vectors into a codebook representing the characteristics of each cluster.

(Step 3) Create time-series representation of training data

Step 3-1 We calculated the distances between the feature vector of time slot and each code. And we assigned code of minimum distance to the time slot. The time slot is then shifted, and a search is again made for the nearest code.

Step 3-2 We assigned a code to all time slots.

Step 3-3 We counted the frequency of each transition pattern in each time interval and represented the ratio of the frequencies as time-series information. There were 16 transition patterns (a→a, a→b, $\cdots$, d→c, d→d) because we used four codes. We set the time interval to 10, 20, 30, 40, or 60 s. For example, when we set the time interval to 10 s, we calculated the frequency of each transition pattern in each 10-s interval (comprising ten time slots) and calculated the ratio of the frequencies of each transition pattern. We then shifted the time interval and calculated the frequency of each transition pattern per interval and calculated the ratio of each target transition pattern to all types of transition patterns.

(Step 4) Calculate similarity (correlation coefficient) between each pair of samples in training data

We calculated the correlation coefficient between each pair of malware samples. A total of 50,086 ($={}_{317}C_2$) correlation coefficients were calculated for each interval. We adjusted the time length (number of transition pattern × number of time interval) of each pair of malware samples by using DP matching.

(Step 5) Create representative model for each malware cluster for training data

We performed hierarchical clustering using the correlation coefficients calculated in step 4. A coefficient greater than 0.7 (upper threshold) generally means that the correlation is high. A coefficient less than 0.2 (lower threshold) generally means that the correlation is low. Given these criteria, we selected the malware sample that had the most traffic samples with a correlation coefficient greater than 0.7 as the initial representative malware sample for the cluster. To keep a somewhat high correlation between each pair of malware traffic samples in the cluster, we deselected the malware samples that did not correspond to more than 70% of samples in the cluster; that is, the correlation coefficient was more than 0.2.

(Step 6) Calculate similarity between two samples in testing data

Step 6-1 We created the time series features of the testing data using steps 1 and 3.

Step 6-2 We created a model of normal traffic using steps 1 to 5.

Step 6-3 We calculated the cumulative minimum distance between each representative cluster model and the target malware traffic sample and calculated the cumulative minimum distance between the model of normal traffic and the target malware traffic sample.

Step 6-4 We compared the two distances for each sample. If the distance between the representative cluster model and the sample was greater than that between the model of normal traffic and the sample, the sample was identified as normal. Otherwise it was identified as infected.

4.2.2 Detection using one representative model

For detection using one representative model, we used the time-series information. We did not use the similarity between pairs of malware samples. The average malware traffic sample of the training data was treated as the representative model of malware-infected traffic.

We created the time-series information for the target malware traffic samples using steps 1 to 3 above. We calculated the mean ratio of the frequencies of each transition pattern for all malware traffic samples and selected the sample that was closest to the mean as the representative model of malware-infected traffic.

For testing, we created time-series information for the malware traffic samples using steps 1 to 3 above. We calculated the cumulative minimum distance between the target sample and the model of infected traffic. We also calculated the cumulative minimum distance between the sample and the model of normal traffic and identified the sample as normal or malware-infected on the basis of the two distances.

4.2.3 Detection using models based on security vendor's definitions

For detection using models based on a security vendor's definitions, we used the time-series information and clusters for classification. We did not use the similarity between pairs of malware samples.

We created time-series information for the target malware traffic sample using steps 1 to 3 above. Next, we divided the training malware traffic samples into clusters defined by the security vendor: BKDR, PE, Mal, TROJ, andWORM. We calculated the mean ratio of the frequencies of the transition patterns of the malware samples in each cluster. We selected the sample in each cluster with the frequency closest to the mean as the representative model of malware-infected traffic.

We calculated the cumulative minimum distance between the model of malware-infected traffic and target traffic sample and calculated the cumulative minimum distance between the model of normal traffic and target sample. We identified the sample as normal or infected on the basis of the distances.

5 Results

5.1 Identification rate of proposed method

The identification rate of the proposed method by changing the time interval is summarized in Table 1. The time interval is the duration during which the code transitions were counted, as described in section 3.2. The number of patterns of infected traffic is the number of hierarchical clusters, as described in section 3.3. The identification rate is the number of correctly identified malware-infected traffic samples divided by the total number of such samples in the testing data.

Table 1: Identification rate of proposed method

Time interval (s)	No. of patterns of infected traffic	Identification rate (%)
10	17	100
20	15	100
30	12	100
40	12	100
60	11	99.0

The identification rate was 100% for time intervals of 10, 20, 30, and 40 s, meaning that it is robust against the time interval.

5.2 Identification rate of other methods

To evaluate the effectiveness of proposed method, we compared its identification rate with those of the three reference methods. The identification rate of the one-representative-model method (section 4.2.2) is shown in Table 2. The identification rate of the security-vendor-definition-based method (section 4.2.3) is shown in Table 3. The number of patterns of infected traffic is five because the data used included five malware families.

The identification rate of proposed method is better than those of other methods.

Table 2: Identification rate of one-representative-model method

Time interval (s)	No. of patterns of infected traffic	Identification rate (%)
10	1	12.5
20	1	14.5
30	1	25.5
40	1	32.5
60	1	52.5

Table 3: Identification rate of security-vendor-definition-based method

Time interval (s)	No. of patterns of infected traffic	Identification rate (%)
10	5	47.5
20	5	78.0
30	5	92.0
40	5	87.5
60	5	98.5

6 Conclusion

Our method for detecting malware-infected traffic samples is based on the similarity between the pair of malware samples in this paper. Simulation evaluation demonstrated that the proposed method can effectively identify malware-infect traffic with high accuracy.

Future work includes conducting a large-scale experiment to better evaluate the effectiveness of the proposed method.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15H01684.

REFERENCES

[1] Y. Otsuki, M. Ichino, S. Kimura, M. Hatada, H. Yoshiura, Evaluating payload features for malware infection detection, Journal of Information Processing (JIP), Vol. 22, No. 2, pp. 376-387 (2014).

[2] K. Kuwabara, H. Kikuchi, M. Terada, M. Fujiwara, Heuristics for Detecting Types of Infections from Captured Packets, Computer Security Symposium, pp.1-6 (2009)(in Japanese).

[3] A. Mohaisen, A. G. West, A. Mankin, O. Alrawi, Chatter, Classifying Malware Families Using System Event Ordering, IEEE Conference on Communications and Network Security, pp.283-291 (2014).

[4] D. Chiba, T. Yagi, M. Akiyama, K. Aoki, T. Hariu, Design and Evaluation of a Profiling Method to Detect Post-infection Communications, Computer Security Symposium, pp. 960-967 (2014)(in Japanese).

[5] Gautam Thatte, Urbashi Mitra and John Heidemann, Parametric Methods for Anomaly Detection in Aggregate Traffic, IEEE/ACM Transactions on Networking, vol. 19, issue 2, pp.512-525 (2011).

[6] M. Ichino, K. Kawamoto, T. Iwano, M. Hatada, H. Yoshiura, Evaluating header information features for malware infection detection, Journal of Information Processing, vol.23, no. 5, pp. 603-612 (2015).

[7] T.-F. Yen and M. K. Reiter, Traffic Aggregation for Malware Detection, 5th International Conference, Detection of Intrusions and Malware, and Vulnerability Assessment, pp.207-227 (2008).

[8] T. Ichida, M. Ichino, M. Hatada, N. Komatsu, H. Yoshiura, A study on malware detection method using feature's state transition, Symposium on Cryptography and Information Security, 1E1-2 (2012)(in Japanese).

[9] G. Nychis, V. Sekar, D. G. Andersen, H. Kim, H. Zhang, An Empirical Evaluation of Entropy-based Traffic Anomaly Detection, IMC '08 Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, pp.151-156 (2008).

[10] Y Linde, A. Buzo, R. Gray, An Algorithm for Vector Quantizer Design, IEEE Trans. on Commun., vol. 28 no. 1, pp.84-95 (1980).

[11] M. Hatada, M. Akiyama, T. Matsuki, T. Kasama, Empowering Anti-malware Research in Japan by Sharing the MWS Datasets, Journal of Information Processing, pp.579-588 (2015)

A Method of Lightweight Secure Communication Considering Reliability in IPv6 Wireless Sensor Network

Shunya Koyama*, Yoshitaka Nakamura**, and Hiroshi Inamura**

*Graduate School of Systems Information Science, Future University Hakodate, Japan
** School of Systems Information Science, Future University Hakodate, Japan
{g2117020, y-nakamr, inamura}@fun.ac.jp

Abstract - Recently, IPv6 wireless sensor networks are widely spread in various fields including IoT environments. However, on these sensor networks, it is difficult to use secure communication technologies that can become large overhead, due to power saving of the wireless nodes is important. As one approach to deal with this problem, a method of focusing on Nonce which is one element of security and separating it from secure communication is proposed. However, it remains a problem that not suit in environments such as wireless sensor networks where reliability of communication is not ensured. In this paper, we propose a Nonce truncation method that can deal with such environments. Our method transfer information of about several bits that can estimate the Nonce associated the ciphertext as the truncated Nonce value. We evaluated the effectiveness of our method by comparing the lifetime of the nodes between the method and previous methods.

Keywords: IoT, Reliability of Communication, Secure Communication, Nonce

1 INTRODUCTION

IPv6 wireless sensor networks are widely spread in various fields including IoT environments lately, because of the development of low-power sensor devices and wireless communication technologies. The penetration rate of these device has been increased, and about 50 billion devices will be interconnected in 2020 [1]. It is also expected to be utilized in various fields.

On the other hand, there are constrained networks that impose strict restrictions on the computing power and the communication quality of the sensor devices. They are called LLNs (Low power and Lossy Networks) [2], are composed of communication devices with limited computing resources. Also, the reliability of communication is not guaranteed due to high packet loss rate and so on. These strictly constrained networks are needed to meet the demand of IoT services in various fields.

As one of the proposals for dealing with such constrained networks, IETF (Internet Engineering Task Force) has established a policy to expand part of it based on communication standards used in conventional wireless sensor networks representing Zigbee [3]. As a typical example of this, there is a method of providing an adaptation layer for using IPv6 (Internet Protocol v6) technology on IEEE 802.15.4 which is a data link layer technology for power saving of communication equipment. Specifically, there are 6LoWPAN (IPv6 over Low-Power Wireless Personal Area Networks) which compresses IPv6 header or UDP header, RPL (IPv6 Routing Protocol for Low Power and Lossy Networks) which is a routing protocol to support the above-mentioned unstable network path reliability, and so on.

While it is expected that such communication scheme targeting LLNs based on IEEE 802.15.4 will become widespread, security problems aimed at these constrained networks are also becoming apparent [4]. Also, there are proposals for the lightweight secure communication methods for sensor networks not using IP which was a major before the spread of IoT service, but it has a background which is very different from recent sensor network. For this reason, it is difficult to apply conventional security technology for the LLNs environment. Especially, the problems that be not able to support IEEE802.15.4 small frame size, and unstable communication quality are left.

In this paper, we discuss the unstable communication quality and the resource constraints of sensor devices which are the features of LLNs. Then, we design the secure communication method that can deal with these features. To this end, we focus on Nonce which is one of the security elements and address a method to truncate this. Also, in this method, we design a lightweight secure communication scheme that can operate without applying excessive overhead to sensor devices at any frame loss rate.

2 RELATED WORK

As previous method, a lightweight secure communication system has been proposed, which is focusing on Nonce (Number used once) that is a part of security elements. In the following, we describe the mechanism of the previous method and its applicability to LLNs, based on the basic secure communication technology.

2.1 Overview of basic secure communication

In this section, we describe the general secure communication establishment method, and the secure design when applying it to the LLNs based on IEEE 802.15.4, in consideration of the data frame structure.

2.1.1 Establishing general secure communication

Strictly speaking, the establishment method of secure communication differs depending on the Block Cipher Modes of Operation selected. As famous example of the

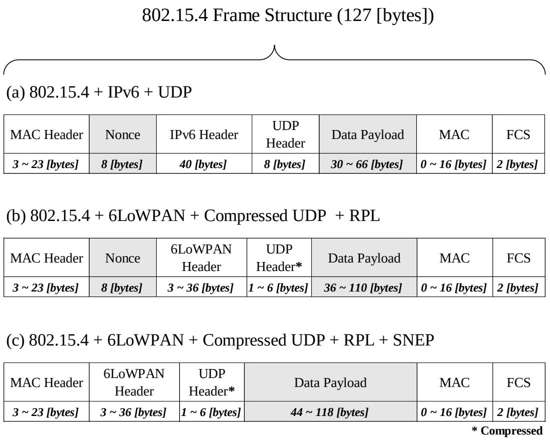
Modes of Operation, there is CCM (Counter with CBC-MAC) mode combines confidentiality and authenticity in an efficient way as authenticated encryption mode. The CCM mode coincides exactly with the design concept of the communication scheme for LLNs from the viewpoint of its versatility, resource constraints, frame size limitation. Therefore, we focus on the CCM mode and explain the operation outline.

CCM mode provides security using secret key and Nonce, and MAC (Message Authentication Code). Here, Nonce is a security element to make it possible to use the same Key multiple times without security risk, and MAC is an arbitrary security element to provide integrity or authenticity.

At this time, in particular with respect to the calculation method of Nonce, the value corresponding to each encrypted frame must be unique from the viewpoint of security risk. In the NIST (National Institute of Standards and Technology), they have listed several recommended specifications and calculation methods of Nonce, and the size should be 8 bytes or more [5]. Further, as one of the calculation methods, a method using a counter value starting from an arbitrary value is recommended. The value is incremented and shared every time different ciphertexts are generated. The methods are described later are based on this calculation method.

2.1.2 Secure communication design in LLNs

Fig.1 shows an example of a simple data frame structures when the above described secure communication is applied to LLNs.



802.15.4 Frame Structure (127 [bytes])

(a) 802.15.4 + IPv6 + UDP

MAC Header	Nonce	IPv6 Header	UDP Header	Data Payload	MAC	FCS
3 ~ 23 [bytes]	*8 [bytes]*	*40 [bytes]*	*8 [bytes]*	*30 ~ 66 [bytes]*	*0 ~ 16 [bytes]*	*2 [bytes]*

(b) 802.15.4 + 6LoWPAN + Compressed UDP + RPL

MAC Header	Nonce	6LoWPAN Header	UDP Header*	Data Payload	MAC	FCS
3 ~ 23 [bytes]	*8 [bytes]*	*3 ~ 36 [bytes]*	*1 ~ 6 [bytes]*	*36 ~ 110 [bytes]*	*0 ~ 16 [bytes]*	*2 [bytes]*

(c) 802.15.4 + 6LoWPAN + Compressed UDP + RPL + SNEP

MAC Header	6LoWPAN Header	UDP Header*	Data Payload	MAC	FCS
3 ~ 23 [bytes]	*3 ~ 36 [bytes]*	*1 ~ 6 [bytes]*	*44 ~ 118 [bytes]*	*0 ~ 16 [bytes]*	*2 [bytes]*

*** Compressed**

Figure 1: Structure pattern of encrypted frames in LLNs

In Fig.1, (a)(b)(c) commonly indicate the frame structure when IP technology is introduced on IEEE 802.15.4 and encrypted using CCM mode. In addition, for each frame structure, (a) is introduced UDP into IPv6, (b) is introduced 6LoWPAN and RPL over (a), and (c) is introduced SNEP described later in section 2.2 of previous method over (b). As can be seen from the figure, the security elements communicated can be large overhead and suppress MAC payload in the environment with limited frame size as LLNs. Therefore, there is a possibility of increasing the processing load of sensor devices through fragment processing, it is desirable to make the size as small as possible. In particular, in each frame structure excluding (c), the ratio of Nonce to MAC payload occupies so large that if Nonce can be completely eliminated, on average about 16% and about 12% of the payload can be expanded.

In order to properly operate the Block Cipher Modes of Operation, there is no strict restriction that each security element must secure a certain size or more. However, if you select the smallest value among the simply selectable sizes, there is also the possibility of impairing the safety of secure communication. From that point of view, NIST recommends size of Nonce is 8 bytes or more. Thus, a method of reducing the size without losing the safety of secure communication is ideal.

2.2 SNEP (Secure Encryption Network Protocol)

Following the previous section, a method of separating Nonce from communication and reducing its size to zero without reducing the safety of secure communication called SNEP (Secure Network Encryption Protocol) [6] has been proposed. Specifically, this method shares only the initial value of Nonce using communication, and thereafter incrementing Nonce value stored in the sensor devices according to the number of received encrypted frames. If an encrypted frame is lost in the middle due to interruption of communication, resynchronization process is performed to transmit the entire value of Nonce. It is shown that in an environment with the stable communication quality, the communication overhead by the secure communication is reduced. On the other hand, in the environment such as LLNs which the communication quality is unstable, the resynchronization process frequently occurs. Therefore, this means secure communication overhead is actually increase, and network congestion problem may occur.

3 PROPOSAL METHOD

3.1 Research tasks

In the previous method, if an encrypted frame is lost in the middle due to interruption of communication or the like, it is necessary to repeat the resynchronization process of Nonce for recovery secure communication. Therefore, it is not support to environment where the frame loss rate can be high. Also, according to a general secure communication method, the ratio of encrypted frames occupied by Nonce is large, and there is a possibility that a heavy load is applied to the sensor devices and the network itself due to inefficient fragment processing. For this reason, any method is difficult to adopt to LLNs where communication quality is unstable and frame size is limited, and a method capable of dealing with these problems is required.

In this paper, we propose a method to deal with the above problem by estimating Nonce only by sensor devices itself from the truncated value that the size changes according to the frame loss rate.

3.2 Basic operation

In this section, we describe the basic mechanism for truncating a Nonce. As the block cipher mode of operation for establishing secure communication, CCM mode are used. Also, as a calculation method of Nonce corresponding to each encrypted frame, a counter value that increment the value according to the frame is used. First step, the initial value of Nonce is shared between Sender and Receiver, and the whole value is stored in the sensor device as in the previous method. Thereafter, in the sharing of Nonce, only the N of least significant bits (N LSBs) are assigned on communication. Hereinafter, this N bit is called a truncated Nonce length. Fig.2 shows the operation flow in the case where the truncated Nonce length is 1 as the specific example.

In Fig.2, (a)(b) commonly begin already synchronized entire Nonce between Sender and Receiver devices and estimate the entire Nonce value from the truncated value while the devices communicate several encrypted frames. First, (a) shows that the receiver succeeds in receiving the third encrypted frame after losing only the second encrypted frame. At this time, since there is a difference in the entire value of Nonce internally stored between the two sensor devices, receiver fails in decryption the third encrypted frame. At this stage, move on to step 2 of (a). Since the value of the received truncated Nonce is 1, receiver can decrypt the third encrypted frame by estimating entire value of Nonce that is greater than internal Nonce value and 1 LSB equals 1. On the other hand, in the case of (b), receiver receives the fourth encrypted frame after losing the second and third encrypted frames, hence fails in decryption even if estimates entire value of Nonce like (a). In the case (b) shown in this figure, although correct Nonce is "*...0010 0110b*", in fact it is estimated "*...0010 0100b*" by mistake. In such a case, recovery secure communication by performing resynchronization process sharing the entire value of Nonce. Generally, such resynchronization process occurs only when the truncated Nonce length is x bits and the frame is lost consecutively for 2^x times or more. For example, in the case of (b), this process occurs because the frame has been lost 2^1 times that is twice consecutively.

Therefore, depending on the selection of the truncated Nonce length, the same problem as SNEP may still occur. For this reason, it is necessary to select the truncated Nonce length flexibly so as to minimize the number of re-synchronous process according to the frame loss rate of the communication environment. Table 1 shows the occurrence probability of resynchronization process according to x bits of truncated Nonce length and frame loss rate.

Table 1: Probability of resynchronization process occurrence according to the truncated Nonce length and frame loss rate

Frame Loss Rate / Truncated Nonce Length	80%	60%	40%	20%	E_{pt}
1	64%	36%	16%	4%	$E_{pt}^{2^1}$
2	40%	13%	2.5%	0%	$E_{pt}^{2^2}$
4	2.8%	0%	0%	0%	$E_{pt}^{2^4}$
x	$80\%^{2^x}$	$60\%^{2^x}$	$40\%^{2^x}$	$20\%^{2^x}$	$E_{pt}^{2^x}$

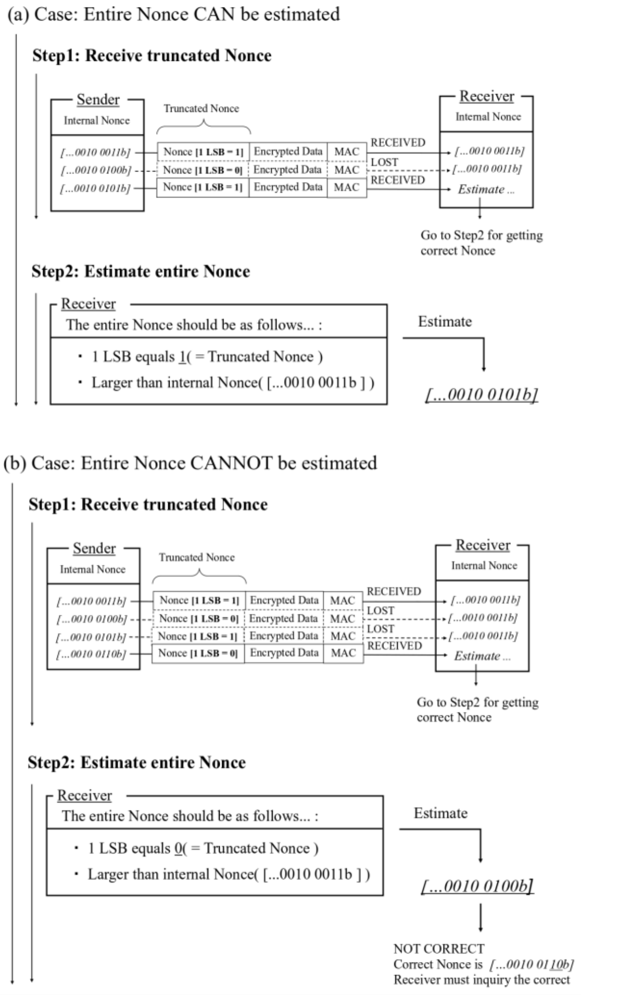

(a) Case: Entire Nonce CAN be estimated
Step1: Receive truncated Nonce
Sender
Internal Nonce
[...0010 0011b]
[...0010 0100b]
[...0010 0101b]
Truncated Nonce
Nonce [1 LSB = 1] | Encrypted Data | MAC
Nonce [1 LSB = 0] | Encrypted Data | MAC
Nonce [1 LSB = 1] | Encrypted Data | MAC
RECEIVED
LOST
RECEIVED
Receiver
Internal Nonce
[...0010 0011b]
[...0010 0011b]
Estimate ...
Go to Step2 for getting correct Nonce
Step2: Estimate entire Nonce
Receiver
The entire Nonce should be as follows... :
・ 1 LSB equals 1(= Truncated Nonce)
・ Larger than internal Nonce([...0010 0011b])
Estimate
[...0010 0101b]
(b) Case: Entire Nonce CANNOT be estimated
Step1: Receive truncated Nonce
Sender
Internal Nonce
[...0010 0011b]
[...0010 0100b]
[...0010 0101b]
[...0010 0110b]
Truncated Nonce
Nonce [1 LSB = 1] | Encrypted Data | MAC
Nonce [1 LSB = 0] | Encrypted Data | MAC
Nonce [1 LSB = 1] | Encrypted Data | MAC
Nonce [1 LSB = 0] | Encrypted Data | MAC
RECEIVED
LOST
LOST
RECEIVED
Receiver
Internal Nonce
[...0010 0011b]
[...0010 0011b]
[...0010 0011b]
Estimate ...
Go to Step2 for getting correct Nonce
Step2: Estimate entire Nonce
Receiver
The entire Nonce should be as follows... :
・ 1 LSB equals 0(= Truncated Nonce)
・ Larger than internal Nonce([...0010 0011b])
Estimate
[...0010 0100b]
NOT CORRECT
Correct Nonce is [...0010 0110b]
Receiver must inquiry the correct

Figure 2: Operation flow in the case where the truncated Nonce length is 1

3.3 Optimization of resynchronization process occurrence count

Considering the characteristics of LLNs, it is necessary to minimize the occurrence probability of resynchronization process as much as possible so that the same problem as SNEP does not occur. For that purpose, it is ideal to flexibly select the truncated Nonce length as short as possible according to the frame loss rate.

In order to deal with this problem, we use ETX (Expected Transmission Count) adopted in routing protocols used in many wireless sensor networks. ETX is a metric index using link quality, and its value is defined as the reciprocal of the frame arrival rate. Therefore, the sensor devices having information corresponding to Table 1, can select the truncated Nonce length dynamically to minimize the occurrence probability of the resynchronization process to any value or less by using this value.

4 EVALUATIONS

About the proposed method and previous methods in this research, we performed evaluation experiments after implementing these on the network simulator. Hereinafter, we describe the experimental environment, evaluation method, experiment method and the detail of these.

4.1 Experiment environment

We implemented the proposed method and (b)(c) in Fig.1 as the previous methods on devices and operated on the network simulator. Specifically, we created a simple small-scale model that established secure communication between two sensor devices such as Sender and Receiver, operated each method in this model. At this time, we emulated all sensor devices as Zolertia Z1 hardware [7]. For simplicity, unidirectional communication is performed from the Sender to the Receiver, and encrypted frames are transmitted and received in this scenario.

Detailed simulation parameters in the experimental environment are shown in Table 2. The communication standard conforms to (b) in Fig.1 as general standard. In addition, a frame loss rate is used as an index representing communication quality in LLNs. Also, only the length of Nonce is selected from among 0 to 8 bits or 8 bytes different according to the frame loss rate. Moreover, considering resynchronization process due to frame loss, experiments were performed until all data arrives at Receiver and completely decrypted after establishing secure communication.

Table 2: Simulation parameters

Parameter	Value
Data Link Protocol	IEEE802.15.4
Network Protocol	6LoWPAN + RPL
Transport Protocol	Compressed UDP
Frame Size	127[bytes]
Transfer Data	1000[Kbytes] * 10
Cipher Mode	AES-CCM*
Key Length	128[bits]
Block Length	128[bits]
MAC Length	8[bytes]
Frame Loss Rate	0%～90%
Nonce Length	0, 1, 2, 4, 8[bits], 8[bytes]

4.2 Evaluation method

In each experimental method, we measured the lifetime of the sensor device from the power consumption of the Sender emulated as Zolertia Z1 hardware. We evaluate the effectiveness by calculating and comparing the lifetime ratio of each method where general method (b) in Fig.1 as 1 value.

4.3 Experimental method

One experiment for each combination of frame loss rate and truncated Nonce length and the other experiment in the case of continuing to select the optimal truncated Nonce length to minimize the number of synchronization process. We describe the details of each experiment method below.

4.3.1 Experiment for each combination of frame loss rate and truncated Nonce length

In this experiment, we evaluate whether the length of each Nonce can correspond to any communication quality assuming LLNs environment as proposed method and previous method. First, about the lifetime ratio of each sensor devices, we calculated from the power consumption until the Receiver took 1,000 KB of data from the Sender 10 times and decrypted all the data. At this time, to evaluate the performance for each length of Nonce in accordance with the frame loss rate, the experiment was proposed at intervals of 10% to 20% in the frame loss rate in Table 2.

4.3.2 Experiment for continuing to select the optimal truncated Nonce length

In this experiment, we evaluate whether the effectiveness can be shown compared with the previous method when dynamically selecting optimum Nonce length using the proposed method. The basic simulation parameters were as shown in Table 2, but the frame loss rate was changed randomly between 20% and 80%, and the occurrence probability of resynchronization process was always 5% or less using ETX. Also, it was assumed that 1000 KB of data was transmitted ten times a day. In such an environment, we calculated the average lifetime ratio of sensor devices in each method.

5 RESULTS AND DISCUSSION

5.1 Results

The result obtained in each experimental method are shown in the following section.

5.1.1 Experimental results for each combination of frame loss rate and truncated Nonce length

The result obtained by the experiment according to the combination of frame loss rate and truncated Nonce length is shown below.

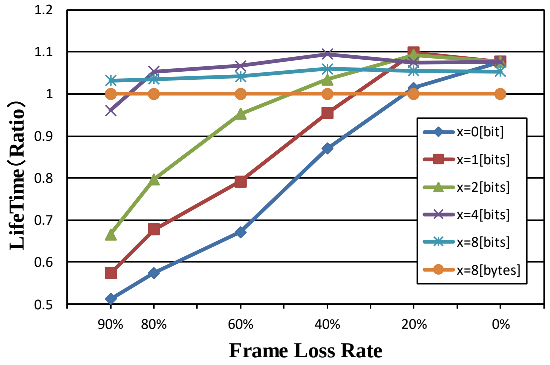


Figure 3: Lifetime ratio according to truncated Nonce length and frame loss rate by simulation

Fig.3 shows Sender's lifetime ratio measured for each frame loss rate and truncated Nonce length (hereinafter referred to as *x*) in the simulation parameters shown in Table 2. In the case where the frame loss rate was 20% or less, all the proposed method and the previous method had improved the lifetime compared with the general method of transmitting 8 bytes of Nonce. On the other hand, when the frame loss rate exceeded 20%, the lifetime sharply decreased according to the length of Nonce. In particular, the rate of decreased was remarkable when the truncated Nonce length was 4 bits, but in the case of 8 bits, any frame loss rate was improved. Also, it cloud be seen that the truncated Nonce length at which the lifetime improves most was different depending on the frame loss rate except 0%.

5.1.2 Experimental results for continuing to select the optimal truncated Nonce length

The result obtained by the experiment for continuing to select the optimal truncated Nonce length so that the occurrence probability of the re-synchronous process within 5% is shown in following Table 3.

Table 3: Lifetime ratio obtained for each method by simulation

Method	Lifetime Ratio
General(Nonce Length: 8[bytes])	1
Previous(Nonce Length: 0[byte])	0.625
Proposal	1.058

The effectiveness of the proposed method is clear because the proposed method was improved the lifetime by about 6%, while SNEP that previous method dropped the lifetime about 37% when compared with the lifetime of general method that transmitted 8 bytes of Nonce.

5.2 Discussion

From the results shown in Fig.3 and Table 3, the lifetime of the proposed method is better than previous methods. Also, if the truncated Nonce length is about 4 to 8 bits, the lifetime is roughly improved in any frame loss rate, except when the loss rate is extremely high. This means that truncated Nonce length is enough size to operate in the LLNs environment. On the other hand, depending on the select of the truncated Nonce length, it is also clear that the possibility of greatly decreasing the lifetime also remains. So, it is effective to select continually the optimum truncated Nonce length.

However, in the experimental environment, since evaluation is limited to a simple secure communication model between two devices, in the future it is necessary to verify the effectiveness from many aspects according to the real environment. Particularly, there are many problems such as dealing with frame delay, handling burst loss caused by network congestion problem. It is also necessary to consider approaches to deal with these problems.

6 CONCLUSION

In this paper, we discussed the unstable communication quality and the resource constraints of sensor devices which are the features of LLNs. Then, we designed the secure communication method that could deal with these features. To this purpose, we focused on Nonce which is one of the security elements and proposed the method to truncate this. As a result, we confirmed the proposed method improved the lifetime as lightweight secure communication method that deal with unstable communication quality which was the problem of the previous method. As the evaluation method, we implemented the proposed method and conventional method on sensor terminal which emulated, measured its lifetime ratio and compared it.

As future prospects, there are we should address examine experiments and evaluation methods considering various more real environments. And, it is also necessary to deal with the response to burst loss and the delay problem of the encrypted frames in connectionless network. Furthermore, in order to improve the proposed method, we will adjust the number of times to estimate the entire Nonce value according to the truncated Nonce length by measuring and comparing the processing load in the decryption process and resynchronization process.

REFERENCES

[1] D.Evans (Cisco Internet Bussiness Solutions Group), “The Internet of Things - How the Next Evolution of the Internet Is Changing Everything, ” <https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf > [Accessed May 20, 2018].

[2] C. Bormann, M. Ersue, and A. Keranen, “Terminology for Constrained-Node Networks,” Internet Engineering Task Force RFC7228, 2014.

[3] ZigBee Alliance., “Zigbee specification. Technical Report Document 053474r20,” Zigbee Alliance, 2014.

[4] D.Airehrour, J.Gutierrez, and S.K.Ray, “Secure Routing for Internet of Things: A survey, ” Journal of Network and Computer Applications, Vol.66, pp.404-412,2016.

[5] National Institute of Standards and Technology, “FIPS PUB 140-2 Security Requirements for Cryptographic Modules, ” 2002.

[6] A.Perrig, R.Szewczyk, J.D.Tygar, V.Web, and D.E.Culler.,“SPINS: security protocols for sensor networks,” Wireless Networks Journal, Vol.8, pp.521-534, 2002.

[7] Zolertia, “Zolertia Z1 Datasheet,” <http://zolertia.sourceforge.net/wiki/images/e/e8/Z1_RevC_Datasheet.pdf> [Accessed May 20, 2018].

Effective Countermeasure Against Biometric Spoofing Attacks using Unsupervised One-class Learning

Vishu Gupta[†], Masakatsu Nishigaki[†], Tetsushi Ohki[†]

[†]Faculty of Informatics, Shizuoka University, Japan
vishu@sec.inf.shizuoka.ac.jp, {nisigaki, ohki}@inf.shizuoka.ac.jp

Abstract - With the advent of new technologies, the methods of presentation attacks as well as the security measures taken against it is diversifying with each passing day and are competing with each other. Yet the imposter is able to make an access to a system illegally by deceiving the security through the use of material containing artificial biometrics traits like printed photo, display etc. Therefore, we propose a novel presentation attack detection algorithm which can ensure security against unknown presentation attacks without any prior knowledge of fake samples. Moreover, our proposed algorithm can detect presentation attack with single static image only. The key tasks are divided into two parts, creating a smooth manifold of live samples and determining whether the query image is included in the manifold. In this paper, we utilize one class system such as SVM(Support Vector Machine) and DCGAN(Deep Convolutional Generative Adversarial Network) to learn the manifold of live samples. For DCGAN we propose a liveness scoring scheme based on the AnoGAN(Anomaly Generative Adversarial Network) framework. Based on these, we utilize the proposed method to face presentation attack detection. Through our experiment we were able to produce decent results by using palm live/fake image dataset.

Keywords: biometrics, spoofing, presentation attack detection, anomaly detection, generative adversarial networks

1 Introduction

In a society where not only simple tasks but decision making of people to computers is coming it will become an important requirement to guarantee that the outsourcing was performed by the user's own will, and also to correctly detect it when counterfeiting acts are forged or improperly tampered. It is essential to guarantee the authenticity of the terminal in addition to the authenticity of the terminal itself to satisfy these requirements. Biometric authentication system is drawing attention, which can guarantee the authenticity of the terminal user.

Biometric authentication system (BAS) registers preliminary collected biometric information as a template and verifies whether it belongs to a legitimate user by calculating the similarity with the biometric information acquired at the time of authentication. BAS uses a biometric feature of the person without fear of forgetting, losing, or theft compared to an authentication method using a password or a token.

On the other hand, biometric information such as faces, sounds, fingerprints, handwriting is difficult to keep secret in daily life. Biometric presentation attack is becoming a large threat since fake biometric information becomes more sophisticated along with the rapid development of sensors, printers, and manufacturing machines.

To develop a BAS that is secure against presentation attacks, demand for designing a robust presentation attack detection(PAD) algorithms which classify an input sample as live or fake is increasing.

Many previous approaches discussed the PAD features which can guarantee security against a specific impersonation attack such as frequency spectrum for printed photo [1], three dimensionalities of live face [2] and so on. However, the methods of presentation attacks are diversifying day by day. It is difficult to learn in advance PAD features that can detect all these attacks.

Regarding the problem, PAD algorithms have made it possible to detect various presentation attacks by combining multiple classifiers that solve binary classifications problem between live and fake samples such as [3, 4]. However, these methods still have some problems. At first, it is necessary to obtain not only biometric samples but also a large number of fake samples for each type of presentation attack. Second, the PAD algorithm does not guarantee whether an unknown sample is classified as a presentation attack. Here we define unknown sample as a sample that is not included in the samples for training. Note that unknown sample includes not only samples intended to resemble live sample but also any synthetic samples since it is sufficiently effective in the registration process.

The subject of this paper is to investigate the security against the presentation attack using an unknown sample. This solves the fundamental problem of making PAD difficult due to the diversification of attacks. The main contributions of this work are as follows:

1. We propose a novel Presentation Attack Detection (PAD) algorithm which can be learned only with live samples and guarantee security against unknown sample by utilizing GAN based anomaly detection algorithm.

2. Proposed PAD algorithm is evaluated with public database (Replay-Attack Database [5]) and achieved 3% of HTER (Half Total Error Rate) by using a model trained only with live samples.

2 Related Work

All the prior research that has been conducted on Anomaly detection is performed by having to train the system by using both live samples along with fake samples which are used for presentation attack. For all these conducted researches, the core difference lies in the method used to model the real and fake attempts. Prior methodologies based on the employed cues are being classified in a recent study [6] where they are divided into three major categories.

The first category is a method to detect face liveness which relies on image quality/distortion measures. Work in [7] which consists of detecting print attacks using the difference in the 2D Fourier spectra is an example of the method in this category. The work stated in [8] utilizes the Lambertian model which comprises of variational Retinex based approach and Gaussian filters difference as its two methods.

The second category uses methods which are based on detecting different signs of vitality which make use of characteristics corresponding to live faces. For example, presentation attack detection in [9] uses blinking which is used with others cues in other work. Such as, [10] recommend the use of all the dynamic information content of the video represented using dynamic mode decomposition method.

The last category consists of methods based on the difference in motion patterns between real and presentation attacks. It is assumed that presentation attack has rigid motion whereas real-access attempts has both rigid and non-rigid motion. These approach depends on the fact that real accesses correlate with 3D structures whereas presentation attack media are often at 2D planes. Eulerian motion magnification using two sets of features composed of LBPs(Local Binary Patterns) [11] to enhance facial expressions is a typical case of the method in this category. New liveness detection method is proposed in [12] which utilizes the difference in optical flow fields generated by the movements of 2D planes and 3D objects.

While most of the existing methods uses real access data to try and learn a general classifier to outline presentation attack attempts, work in [13] uses both texture and motion cues, the authors built two presentation attack detection methods, one being a generative approach while the other being a discriminative method to study the client-specific information embedded in the feature space and its effects on the performance of the system.

The current work in terms of detection mechanism share some similarities to the existing approach which utilizes image content representation, is distinct in the way we formulate the existing the detection problem. The common approach used to detect anomaly in an image uses two-class formulation where they separate the negative from the positive samples, our proposition uses one-class pattern classification methods, testing it in a modified as well as an existing method which yields good results to identify presentation attack attempts. Moreover, the evaluations are performed by using a self-made database which better reflects the difficulties of detection in realistic scenarios.

3 Presentation Attack Detection using Anomaly GAN

3.1 Generative Adversarial Networks

Goodfellow et al. introduced a concept of Generative Adversarial Network(GAN) [14] which learns a *generator* expression indistinguishable by a *discriminator* by training a *generator* model and *discriminator* model simultaneously. The aim of the *generator* is to fool the *discriminator* by learning the probability distribution of the input samples. Let $\boldsymbol{x}$ be an input sample whose true probability distribution is $p(\boldsymbol{x})$. G is a *generator* that takes a latent vector $\boldsymbol{z}$ randomly selected from the latent space $\mathcal{Z}$ and outputs a new sample $G(\boldsymbol{z})$. The *discriminator* D then outputs the probability that the given input is either the true input from $p(\boldsymbol{x})$ or the $G(\boldsymbol{z})$ from the *generator*. These two models are simultaneously trained using the min-max game of the formula:

$$\min_D \max_G V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \quad (1)$$

Radford et al. [15] introduced deep convolutional generative adversarial networks (DCGAN) for unsupervised learning of features by utilizing convolutional neural networks as the *generator* and *discriminator* network. More specifically, they replaced the pooling layer with stride convolution layer so that the network can learn its own spatial upsampling. Additionally, they removed the full connection layer at the top of the convolution feature to improve the model stability. Finally, batch normalization was utilized to suppress training problems caused by poor initialization and helps propagation of gradients in deep models by normalizing each unit to have zero mean and unit variance.

3.2 Proposed Anomaly GAN for PAD

To detect presentation attack using single image, we propose a unsupervised learning to identify anomalies in imaging data as candidates for fake sample. Figure 1 shows an overview of our proposal. Our proposed scheme is based on unsupervised anomaly detection scheme proposed in [16](hereafter, AnoGAN). AnoGAN uses DCGAN to learn a manifold of live sample variability, acompanying a anomaly scoring scheme based on the mapping from an image space to a latent space.

3.2.1 Palm Imaging Model

We learn the palm image manifold $\mathcal{X}$ on the image space with unsupervised learning using only the live palm images. When query image is not included in the learned manifold $\mathcal{X}$, it can be detected as an unknown input.In DCGAN [15], *generator* uses latent vector $\boldsymbol{z}$ chosen from latent space $\mathcal{Z}$ uniformly at random to obtain a smooth mapping $G(\boldsymbol{z})$ to palm image manifold $\mathcal{X}$.

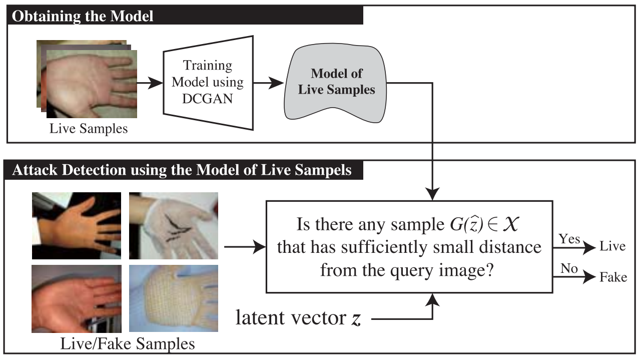

Obtaining the Model
Live Samples
Training Model using DCGAN
Model of Live Samples
Attack Detection using the Model of Live Sampels
Is there any sample $G(\hat{z}) \in \mathcal{X}$ that has sufficiently small distance from the query image?
Yes
Live
No
Fake
latent vector z
Live/Fake Samples

Figure 1: Overview of our proposal

3.2.2 Deriving latent vector

We can detect the Presentation Attack by checking whether query image $\boldsymbol{x}_q$ is included in the palm image manifold $\mathcal{X}$ learned in the clause 3.2.1. Since DCGAN calculates $G(\boldsymbol{z})$ using the randomly chosen latent vector $\boldsymbol{z}$, $G(\boldsymbol{z})$ corresponds to a random point on the palm image manifold $\mathcal{X}$. Consequently, the distance between $G(\boldsymbol{z})$ and the query image $\boldsymbol{x}_q$ does not necessarily become small even if it is a live sample. Therefore, to detect an unknown sample, we should confirm the existence of a latent vector $\widehat{\boldsymbol{z}}$ that has sufficiently small distance between the query image $\boldsymbol{x}_q$ and $G(\widehat{\boldsymbol{z}})$ on the manifold $\mathcal{X}$.

For finding the $\widehat{\boldsymbol{z}}$ from randomly chosen latent vector $\boldsymbol{z}$, we use the backpropagation approach proposed in [16]. The loss function $\mathcal{L}(\boldsymbol{z}_\gamma)$ for backpropagation is defined as follows:

$$\mathcal{L}(\boldsymbol{z}_\gamma) = (1-\lambda)\cdot\mathcal{L}_R(\boldsymbol{z}_\gamma) + \lambda\cdot\mathcal{L}_D(\boldsymbol{z}_\gamma) \tag{2}$$

where $\boldsymbol{z}_\gamma$ is an updated latent vector to fool *discriminator* D, $\mathcal{L}_R(\boldsymbol{z}_\gamma)$ is the generator loss, $\mathcal{L}_D(\boldsymbol{z}_\gamma)$ is the discriminator loss and λ is a fixed parameter for convex combination. The residual loss and the discriminator loss can be obtained as follows:

$$\mathcal{L}_R(\boldsymbol{z}_\gamma) = \sum |\boldsymbol{x}_q - G(\boldsymbol{z}_\gamma)| \tag{3}$$

$$\mathcal{L}_D(\boldsymbol{z}_\gamma) = \sum |\boldsymbol{f}(\boldsymbol{x}_q) - \boldsymbol{f}(G(\boldsymbol{z}_\gamma))| \tag{4}$$

where $\boldsymbol{f}(\cdot)$ is an output of the *discriminator* function. Only the coefficients of $\boldsymbol{z}$ are adapted via backpropagation. The trained parameters of the *generator* model and *discriminator* model are kept fixed. In our proposal, $\widehat{\boldsymbol{z}}$ is obtained by applying backpropagation process α times with query image $\boldsymbol{x}_q$ and randomly selected $\boldsymbol{z}$. The obtained $\widehat{\boldsymbol{z}}$ is used in classification process.

3.2.3 Classification

In classification process, we investigated the three types of score function, anomaly score $A(\boldsymbol{x})$, residual score $R(\boldsymbol{x})$, and discriminator score $D(\boldsymbol{x})$, respectively. The relationship between each score is defined as follows

$$A(\boldsymbol{x}_q) = (1-\lambda)\cdot R(\boldsymbol{x}_q) + \lambda\cdot D(\boldsymbol{x}_q) \tag{5}$$

where the residucal score $R(\boldsymbol{x}_q)$ and discrimination score $D(\boldsymbol{x}_q)$ are defined by the residual loss $\mathcal{L}_R(\widehat{\boldsymbol{z}})$ and discriminator loss $\mathcal{L}_D(\widehat{\boldsymbol{z}})$ using at the α update iteration of the mapping procedure to the latent space, respectively. All score functions output a large score for an anomaly image. In our experiments, we use λ = 0.9 in equations (2) and (5) which was found empirically due to preceding experiments on our palm dataset.

4 Experiment

In this section, first a description of the custom made database and the evaluation protocols used in this experiment is provided, following by experimental results obtained from the database used. All the experiments were carried out using Python with the tensorflow and pytorch library on a machine with configuration (Intel i7-5930K, 64GB RAM, 12x Intel(R) Core(TM), Ubuntu 64bit) environment.

4.1 Database

In our experiment, we constructed a custom made database to make sure that the system is being able to make clear distinction between live and fake samples even when the system encounters unexpected inputs which has no direct relation with the hand as though many previous works have used public live/fake dataset such as Replay-Attack Database [5], it contains only a specific type of fake photo and video samples making it inadequate in terms of unknown samples.

The custom made database used in the experiment consists of 8748 live samples and 6648 fake samples of palm with an image resolution of 160x120 pixels taken directly from approximately 2000 people with the camera of resolution 1280×720pixels. For training, total of 8000 samples were selected to train the model. The test set comprised of 748 live palm samples and 6648 fake palm samples from cases not include in the training set. The training that we are performing in this experiment is totally completely unsupervised. Example of true samples and different variety of fake samples that were used while training the system is given as below in Figure 2. In order to include as many variety of unexpected fake samples as possible to check the accuracy of the system, we included photos such as (b)printed photo, (c)hand wearing synthetic glove, (d) hand wearing cotton glove etc.

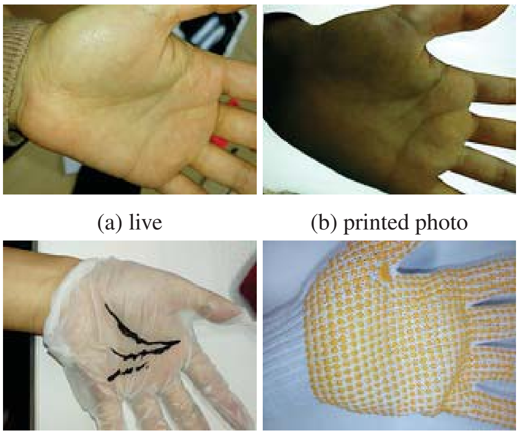
4.2 Evaluation Protocol

The manifold of live images was solely learned on image data of 8000 live cases from the database with the aim to model the variety of live appearance. For performance evaluation in anomaly detection we ran various different protocols exploited by researchers.

4.2.1 Proposed Protocol

All the training and test conducted for the anomaly detection in this work are based on the one class system where only the live samples are used to develop the model. In particular, the following systems are used for the development and evaluation:

- AnoGAN+RAW: The AnoGAN which uses one class system trained using the original image



(a) live (b) printed photo

(c) synthetic glove (d) cotton glove

Figure 2: Example samples used for training the model for 1 class system. (a) is an example of true sample used for training the model and (b) to (d) are the different variety of fake samples that were used for testing the model

- SVM1+RAW: The one-class SVM with a Gaussian kernel trained using the original image
- SVM1+LBP: The one-class SVM with a Gaussian kernel trained using the LBP feature

For each of these protocols, the model was trained using 8000 live samples and the test was conducted by using 100 fake samples along with 100 live samples which were not included in 8000 live samples that were used for training the model in order to check the accuracy of the model in order to distinguish the fake samples from the live samples. Residual score $R(\boldsymbol{x}_q)$ is taken into account in order to differentiate between live samples and fake samples. The purpose of this unsupervised one class training is to find the epochs whose training accuracy as well as prediction accuracy are good and which does not cause overfitting. For this dataset the epochs which showed the best result is 50. As we try to increase the epochs the accuracy of the model decreased. In this experiment the unsupervised learning was conducted by changing the epochs as 20, 25, 50,$\cdots$,100. In the result section of this experiment, we used the result of 25 epochs as the representative example and the result of 50 epochs as it shows the best result. The residual score can vary each time the test is conducted even if the image used for testing is the same because the residual loss measures the visual dissimilarity between query image $\boldsymbol{x}_q$ and generated image $G(\hat{\boldsymbol{z}})$ in the image space by finding a point $\hat{\boldsymbol{z}}$ in the latent space that corresponds to an image $G(\hat{\boldsymbol{z}})$ that is visually most similar to query image $\boldsymbol{x}_q$ and that is located on the manifold $\mathcal{X}$. We ran 100 back-propagation steps (α = 100) for the mapping of new images to the latent space $\mathcal{Z}$.

4.2.2 Evaluation Metric

For evaluating the result obtained, we consider the Area Under Curve (AUC) obtained from Receiver Operating Characteristic (ROC) curves. The ROC curve was made using residual loss as the parameter which yields good results as shown in [16]. Vertical axis and horizontal axis of ROC curves usually present True Positive and False Positive Rate respectively. It indicates that the plots top left corner is the optimal point. Preferable TPR for the ROC curve is equal to one which makes the preferable AUC's values approaching to one.

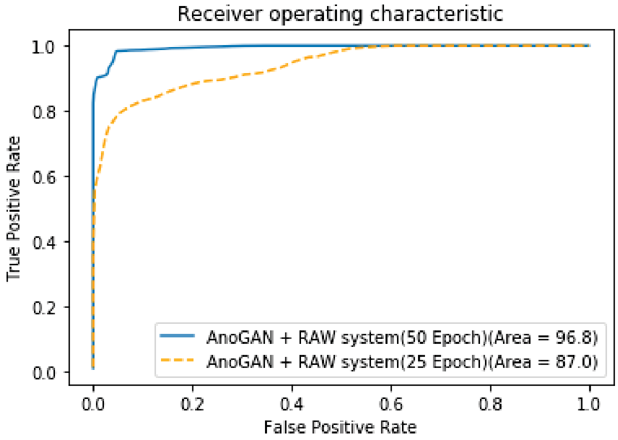


Figure 3: The above figure represents the ROC graph of the AnoGAN model trained for 25 (orange) and 50 (blue) epochs respectively by using live samples as an input image for training.

System	AUC(%)
AnoGAN+RAW (25 Epoch)	87.0
AnoGAN+RAW (50 Epoch)	96.8
SVM1+RAW	34.3
SVM1+LBP	83.5

Table 1: Area under the ROC (AUC) (%) for different systems obtained by using custom database

4.3 Evaluation Results

The one-class systems introduced earlier are evaluated on the custom made database which used 8000 live samples to develop the model. In order to make sure that there is no bias in the result obtained after testing each of the model we took out a total of 200 samples randomly from the palm database, 100 samples each from live samples and fake samples. For the fake samples, even though the 100 images taken out were selected at random, it was made sure that it contained all the variety of samples that were taken into account while creating the fake samples. By doing so, we can see to what extent the trained model produces the desired result even if it encounters unexpected input which is fake but has no direct relation with hand.

Figure 3 represents the ROC graph of the results obtained from different models where the Y-axis shows True Positive Rate and the X-axis shows False Positive Rate. Additionally, Table 1 and 2 show the AUC and HTER (Half Total Error Rate) respectively. Note that HTER can be calculated by $\min(TN + FP)/2$.

Table 1 shows that the best performing one-class system in terms of average performance is Ano-Gan+RAW with an av-

System	HTER(%)
AnoGAN+RAW (25 Epoch)	13
AnoGAN+RAW (50 Epoch)	3
SVM1+RAW	34
SVM1+LBP	17

Table 2: Half Total Error Rate(HTER)(%) for different systems obtained by using custom database

erage AUC of 96.8%. The result obtained in terms of AUC by using one class SVM system [17] as a model for training and testing the dataset as used for training and testing AnoGAN is also shown in Table 1 and 2. It is clearly visible that the proposed AnoGAN system is far more better than the conventional one class SVM system. As far as the one class SVM system are concerned, SVM1+LBP performed better as compared to SVM1+RAW. It is because as stated in [18] by using LBP feature they were able to perform their experiment in a robust way which was computationally fast and didn't required any user-cooperation. Moreover, the extensive experimental analysis done by them on a publicly available database showed excellent results compared to existing works which proves clarifies that SVM1+LBP will show better results as compared to SVM1+RAW.

5 Discussion

While we believe that an unsupervised model such as AnoGAN could have many benefits, we also see some research problems. First, we have to find a way to determine the number of epochs for which you have to train the system by relating it with the number of images used for training. Second, while testing the one-class SVM method the SVM1 + LBP produced better results as compared to SVM1 + RAW. So if the code of AnoGAN is designed in such a way that it calculates the loss function while taking LBP (histogram, cosine similarity) into consideration from the point of training then there is a chance that it might produces better result. Third, we have only examined anomaly detection systems based on 1 class SVM and AnoGAN, it would be better if we study other anomaly detection approach also. So, it can be concluded that even if you train the system by using the true samples only, it does not perform well enough and more modification and research should be conducted to improve the performance of this type of system. In future experiments we would like to take the cost included in the making of the data set as well as the time it takes to detect the fake samples from a group of samples into consideration which would serve as a checkpoint from where we can strive for further improvement.

6 Conclusion

In this study, we investigated a face presentation attack detection method based on an anomaly detection using Generative Adversarial Network. Our remarkable result is that the proposed PAD scheme achieved 96.8% AUC and 3% of HTER by using a model trained only with real samples. It is clearly visible that our proposal can achieve far more better result than conventional one-class SVM systems. Additionally, it should be noted that our method can detect presentation attack by using single static image. Therefore, this method can also be directly applied to deal with video presentation attack or be integrated with a video-based palm liveness detection method for better performance. It is left to investigate about loss function suitable for presentation attack and reduce the number of backpropagation α to improve our method more secure and convenient.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number JP18K11294. We thank Eizaburo Iwata and Hiroki Kamanaka for useful discussions. The Palm Presentaton Attack Dataset is provided by Normee co, ltd.

References

[1] A Pacut and A Czajka. Aliveness Detection for IRIS Biometrics. In *Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International*, pages 122–129. IEEE, 2006.

[2] Maria De Marsico, Michele Nappi, Daniel Riccio, and Jean-Luc Dugelay. Moving face spoofing detection via 3d projective invariants. In *Biometrics (ICB), 2012 5th IAPR International Conference on*, pages 73–78. IEEE, 2012.

[3] Heeseung Choi, Raechoong Kang, Kyungtaek Choi, and Jaihie Kim. Aliveness Detection of Fingerprints using Multiple Static Features. In *Proc. of World Academy of Science, Engineering and Technology*, pages 201–205, 2007.

[4] Ricardo N Rodrigues, Niranjan Kamat, and Venu Govindaraju. Evaluation of biometric spoofing in a multimodal system. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–5, 2010.

[5] André Anjos and Sébastien Marcel. Counter-measures to photo attacks in face recognition - A public database and a baseline. In *Biometrics (IJCB), 2011 International Joint Conference on*, pages 1–7. IEEE, 2011.

[6] Tiago de Freitas Pereira, Jukka Komulainen, André Anjos, José Mario De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. *EURASIP Journal on Image and Video Processing*, 2014(1):2, 2014.

[7] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. In *Biometric Technology for Human Identification*, volume 5404, pages 296–304. International Society for Optics and Photonics, 2004.

[8] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010.

[9] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[10] Santosh Tirunagari, Norman Poh, David Windridge, Aamo Iorliam, Nik Suki, and Anthony TS Ho. Detection of face spoofing using visual dynamics. *IEEE transactions on information forensics and security*, 10(4):762–777, 2015.

[11] D.Harwood T.Ojala, M.Pietikainen. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Image Processing. Proceedings of the 12th IAPR International Conference on*, pages vol. 1, pp. 582–585. IEEE, 1994.

[12] Wei Bao, Hong Li, Nan Li, and Wei Jiang. A liveness detection method for face recognition based on optical flow field. In *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*, pages 233–236. IEEE, 2009.

[13] Ivana Chingovska and André Rabello Dos Anjos. On the use of client identity information for face antispoofing. *IEEE Transactions on Information Forensics and Security*, 10(4):787–796, 2015.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[16] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Proceedings of the 25th International Conference of the Information Processing in Medical Imaging IPMI 2017, Boone, NC, USA*, pages 146–157, June 2017.

[17] Alex Smola John Shawe-Taylor John Platt Bernhard Schlkopf, Robert Williamson. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pages 582–588, 1999.

[18] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *Biometrics (IJCB), 2011 international joint conference on*, pages 1–7. IEEE, 2011.

Session 5:
Data Models

(Chair: Yoshia Saito)

A Proposal of Deep Learning Training Data Generator for Image Recognition

Tsukasa Kudo[†], and Ren Takimoto[†]

[†]Faculty of Informatics, Shizuoka Institute of Science and Technology, Japan
kudo.tsukasa@sist.ac.jp

Abstract - At present, image recognition technologies utilizing the deep learning are developing rapidly, and some cases have been reported in which their recognition accuracy has exceeded the human vision. On the other hand, in order to conduct deep learning, it is necessary to accumulate a large amount of training data. And, this often becomes the obstacles to applying the deep learning to actual business systems. So, in this study, we propose a method to generate the images of training data automatically by using computer graphics (CG). For example, in inventory management of parts, target objects are composed only of inventory shelves and parts. That is, although it is difficult to accumulate a large number of their actual images, it is expected that these images can be easily generated by using CG. However, while it becomes easy to prepare a large amount of the training data by this method, there may be a deviation between the actual images and generated CG images. Therefore, firstly, we evaluate the influence of this deviation on the deep learning by using plural kinds of CG images. Next, we conduct experiments using training data in which a part of the CG images are replaced with the actual images, and we evaluate the improvement of the accuracy corresponding to the ratio of actual images. As a result, we show that by mixing a certain ratio of actual images, effective training data for the deep learning can be generated with CG images.

Keywords: Deep learning, Computer Grapahics, Inventory management system, Stocktaking, Convolutional neural network

1 INTRODUCTION

Currently, the accuracy of image recognition utilizing the deep learning is rapidly improving, and the case of achieving accuracy exceeding human vision has also been reported [4], [6], [14]. As a result, its applications are rapidly spreading to various fields [3], [5]. On the other hand, since a large amount of training data is necessary in the deep learning, a large amount of image data has to be accumulated for image recognition, too. Therefore, in several application fields, there is a problem that it takes a long time to accumulate such an image data.

For example, in the manufacturing factory of mechanical products which our laboratory supports its production management system improvement, the stocktaking of parts inventory is a heavy workload. So, we proposed a method to automatically discriminate the inventory satisfaction by image recognition utilizing deep learning, and constructed and evaluated its prototype. Here, the number of image data used for this evaluation is 1,600 per part. However, in the actual factory, the parts are delivered to the assembly field from parts shelves collectively: by each product lot, namely product manufacturing unit, or by each order composed of several products. So, the number of times of variation of the shelves is comparatively small. For example, in the case where its variation occurs once a day, about 250 images are obtained a year. That is, to accumulate 1,600 images, it takes more than 6 years. In addition, parts are stored in each inventory shelf, and their kinds extend to thousands. And, since most parts are heavy, it takes large man-hours to deliberately change the situation of the inventory shelves by hand so many times. For these reasons, to prepare the training data efficiently has become the problem of applying the deep learning to the actual inventory satisfaction discrimination by the image recognition.

Here, each image of an inventory shelf is composed of only two types of objects: the inventory shelf itself and the parts. And, the parts are placed in the shelf according to a certain rule. For example, in the case of storing relatively small parts in a bulk container, the state of the inventory shelf can be composed by piling up the parts randomly from the bottom of the container. This suggests that a large amount of various image data of each inventory shelf can be generated by the computer graphics (CG) automatically.

The motivation of this study is to show there are fields as follows: it is difficult to accumulate a large amount of training data composed of photographs of the actual objects (hereinafter, "actual images") for the deep learning; but, it is easy to accumulate the training data by utilizing CG. In addition, it includes showing the method to obtain higher image recognition accuracy in the case of utilizing the CG images. Therefore, the final goal of this study is to develop a system for image recognition of actual object by using deep learning which training data is generated by utilizing CG tool efficiently.In addition, it is expected to be effective in the case of the abovementioned stocktaking and so on.

In this paper, we perform a feasibility study to develop such a system, namely the evaluation of the image recognition accuracy in the case of using CG data for the deep learning. Concretely, presupposing the inventory shelf, we classify the number of stored parts from 5 to 80 every 5 and evaluate the accuracy of their estimated quantities

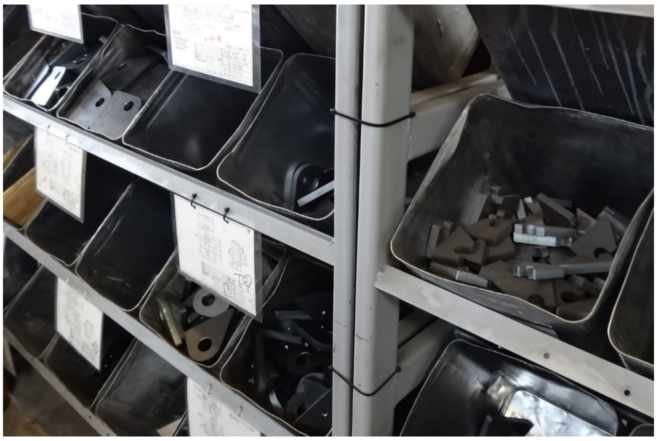


Figure 1: Inventory shelves of bulk container

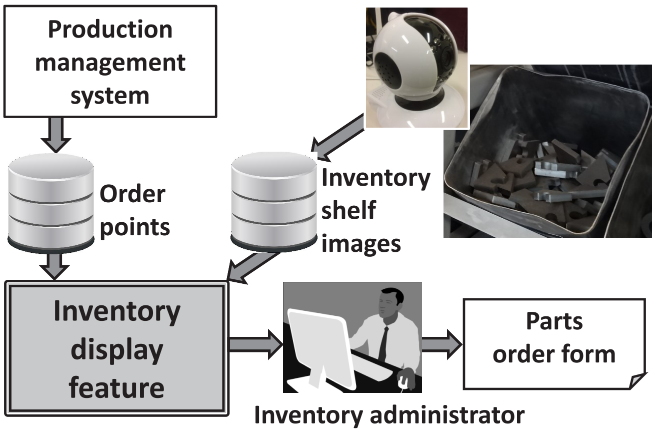


Figure 2: Inventory management utilizing images

by using the supervised learning.

On the other hand, the creation of high-precision CG image similar to the actual images also requires a large number of man-hours. Conversely, the actual images can be obtained even in the case of the above-mentioned inventory shelves in the factory, if the number of the images is small.

So, we conduct this evaluation in the following two stages. Firstly, we extract the factors in the CG image creation and evaluate the influence of each factor on the accuracy of the image recognition. Next, we evaluate the change of this accuracy in the case where the CG images are replaced with the actual image by each ratio. As a result, we show that by replacing just a part of the CG images with the actual images, it is possible to achieve the same accuracy in the case of using the training data constructed of only the actual images.

The remainder of this paper is organized as follows. Section 2 shows the related works and the problem to create the training data, and we propose the training data generation method with CG for the stocktaking in Section 3. Section 4 shows the implementation of this method, and Section 5 shows the experiments and evaluations with the training data generated by this method. We discuss the evaluation results in Section 6, and conclude this paper in Section 7.

2 RELATED WORKS AND PROBLEM

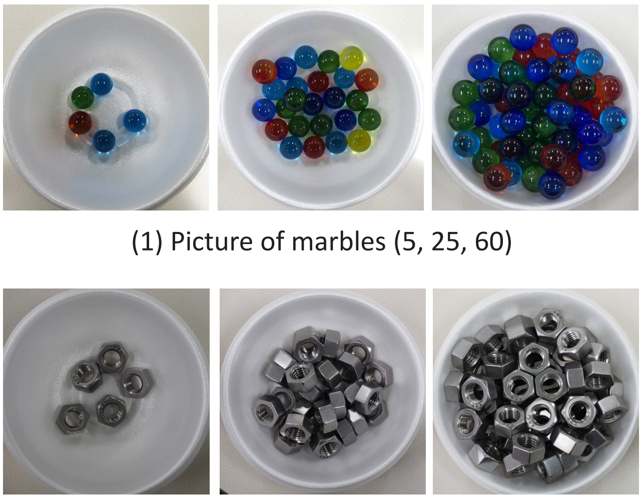
In this section, we explain the background and related works of our idea described in this paper. Our laboratory supports the factory to introduce and operate its production management system, which manufactures mechanical products. Since thousands of parts in various shapes are stored in each inventory shelf in the factory, the workload required for stock-taking of inventory is a serious problem. In particular, in the case where parts are stored in the bulk container as shown in Fig. 1, they cannot be counted from the outside. So, it is necessary to take out the parts in the container and count them up. So, it is a major factor increasing man-hours.

On the other hand, with the progress of the Internet of Things (IoT), various sensors such as surveillance cameras are controlled remotely, and their data is accumulated and analyzed in the server. As a result, since so various and enormous data has been stored in the database, it has become difficult to deal with such a data with conventional relational databases. So, various NoSQL databases have been put to practical use to manipulate such a data efficiently [11]. For example, MongoDB is a kind of document-oriented NoSQL databases and provides the GridFS interface to manipulate such a data efficiently in the distributed environment [1]. That is, at present, an environment has been developed, in which large capacity of image and video data can be easily handled.

Due to such a technical background, we proposed an inventory satisfaction discrimination method using the inventory shelf images shown in Fig. 2 [9]. Here, the ordering point quantity is determined for each part in the inventory shelves by the production management system, and inventory is replenished when its inventory quantity falls below this ordering point [12]. Therefore, to discriminate visually that the inventory of each part satisfies this ordering point quantity is more efficient than performing the stocktaking of the inventory. For example, in the case of the inventory shelf on the upper right of Fig. 1, while it is difficult to grasp its exact quantity, it is relatively easy to discriminate that there are ten or more parts.

And, in the case where the discrimination is difficult, by replenishing the part from the viewpoint of safety, it becomes not necessary to count the inventory quantity. As a result, we showed that the efficiency of the inventory management could be achieved by showing both the production plan data and current images of the inventory shelves to the inventory administrator as shown in Fig. 2. However, even by this method, some problems about the workload of the inventory manager remain: he must check many inventory shelves one by one; especially, in the case where a large number of parts are stored in the bulk container, it takes time for the discrimination.

On the other hand, currently, the accuracy of image



(1) Picture of marbles (5, 25, 60)

(2) Picture of nuts (5, 25, 60)

Figure 3: Picture example of experimental objects

recognition is rapidly improving by utilizing deep learning. For example, in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), algorithms compete for object detection and image classification of large-scale. And, after the deep learning was applied in 2012, the recognition rate has been greatly improved. Moreover, it exceeded 5.1% of human recognition rate in 2015 [14]. Along with the improvement in recognition rate, the application of deep learning to image recognition is spreading in various fields such as face authentication, medical image diagnosis, plant disease detection, and so on [3], [5], [10].

So, we conceived to apply the deep learning to the inventory satisfaction discrimination utilizing image recognition. And, as a feasibility study, we conducted the experiment to recognize the classified quantity of the objects by the supervised deep learning. Then, we constructed the deep convolution neural network (deep CNN) and performed the deep learning; we evaluated its accuracy of quantity estimation by using the test data of images prepared separately[7], [8]. For the target objects, we used the marbles and nuts, which is easy to create training data. And, we classified each of them from 5 to 80 every 5 with the label of their quantity.

Figures 3 shows image examples of training data in the case of 5, 25 and 60. We prepared 100 image data for each class, that is, the total is 1,600 for each of marble and nut. Then, we padded 400 images by up/down and right/left inversion, and 90 and 270 degree rotation. As a result, we created 500 image data for each class, that is, the total was 8,000 for each of marble and nut. Next, we separated them into 6,000 training data and 2,000 test data, then we separated 600 data from the training data as the verification data. Ultimately, with 5,400 training data, we conducted supervised deep learning with the above-mentioned labels. Fig. 4 shows the composition of deep CNN for this deep learning.

After that, we evaluated the accuracy of quantity estimation in this deep CNN by using the test data of images prepared separately. Firstly, we conducted the comparative evaluations between the deep CNN and human vision by using marbles. As for human vision, we showed the images of the training data of each class to let the humans learn its quantity. Then, the humans were shown the test image for 10 seconds to answer its quantity. As a result, we found that although the human vision was able to accurately estimate the quantity in the case where it was small, the deep CNN's accuracy was higher in the case where the quantity is equal or more than 20. This experiment corresponded to the inventory shelf of the bulk container shown in Fig. 2. And, as shown the image of 60 in Fig. 3, it was difficult for humans to estimate the quantity in a short time in the case of many objects.

Furthermore, we evaluated the distribution of the estimated quantity. And, we confirmed that there was the tendency that they are gathered in the vicinity of the actual quantity, although some of them may disperse rough. As a result, we concluded that it could be applied to actual inventory management systems by taking the following countermeasures: the safety stock should be increase to permit the error; the estimated quantity should be compared with the logical inventory quantity calculated by the production management system to detect the error.

Here, in this experiment, as we treated small marbles and nuts shown in Fig. 3, it was easy to obtain a large amount of photo to create the training data by shuffling these objects in the bowl at every time. However, in the actual factory, the bulky and heavy parts shown in Fig. 1 are treated. In addition, since it is in operation, we cannot hinder the workers. That is, there is a problem that it is difficult to accumulate a large amount of training data for applying the deep learning.

Generally, to apply the deep learning, accumulating a large amount of training data is an important factor to improve its performance, and training data is collected in various ways in each application field. On the other hand, there are application fields where collecting the training data is difficult like this case. Therefore, it is considered effective to develop an efficient preparing method of the training data in such a field.

3 PROPOSAL OF TRAINING DATA GENERATION METHOD UTILIZING CG

As a solution to the problem of the fields where the accumulation of a large amount of training data is difficult, we propose a training data generation method by utilizing CG. Considering the characteristics of the inventory shelf, usually only one kind of parts is stored in one shelf. That is, from the viewpoint of CG, it is possible to construct the state of the inventory shelf with only two objects, one inventory shelf and one part. Furthermore, it is not necessary to consider the deformation

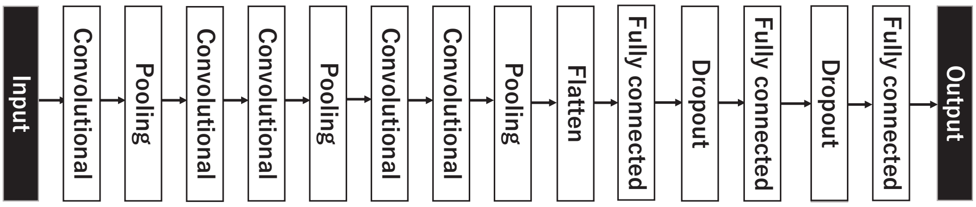


Figure 4: Construction of deep convolutional neural network

of each object. That is, by piling up the part objects randomly in the inventory shelf object, the state of the inventory shelf can be constructed virtually.

Basically, to construct such an inventory shelf by CG, the four factors should be considered: the shape of the objects, the material of the objects, the illumination as the environment, and the placement of the part objects. Firstly, the shapes of objects can be realistically created by using CG modeling tools, due to the measuring data of the actual inventory shelf and part. Secondly, the material expresses the textures of these objects, and it can be also added by the CG modeling tools. However, it is necessary to adjust the material by not only human vision, but also checking the influence to the deep learning. Thirdly, the illumination needs to be determined based on the actual factory illumination environment, and it can be added in the same way as the material. These three factors are static with respect to each inventory shelf. So, it can be used repeatedly after once created.

Fourthly, the placement of the part objects is the most important factor to create a large amount of training data, and the following three requirements should be considered. The first is a physical requirement, for example, it is necessary that the placement in CG does not collapse even if it is actually placed. The second is the realization of random placement, that is, in order to accumulate a large number of training data, it is necessary to construct different placement states even in the case of the same number of the same parts. The third is the rule of placement. For example, the parts in the second container from the left of the uppermost shelf of figure 1 are the one like plates, and they piled for every 4.

Regarding such a placement, various kinds of CG tools are provided. And, many of them can be controlled by the programming language automatically and provide the feature of physical simulation such as free-fall by gravity. Therefore, it is possible to create the various placement state similar to the real world by executing the physics simulation after randomly placing parts by using programming languages.

For example, we show the case to generate the marble training data shown in Fig. 3 by CG tool. Firstly, as shown in (a) of Fig. 5, we create the bowl and marble objects, then place the marble objects randomly above the bowl. Next, the marble objects are dropped in the direction of the arrow shown in (a) of Fig. 5 by using the physical simulation of free-fall by gravity. After falling in the bowl, it converges to the natural state shown in (b) by the control of the physical simulation. And, by saving the rendering image of this result, the image shown in Fig. 3 can be obtained. By repeating these processes automatically by using the programming language, various large amount of training data can be created.

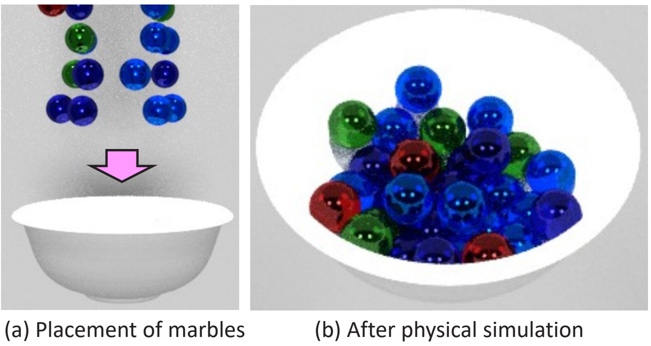

(a) Placement of marbles (b) After physical simulation

Figure 5: Image example of experimental objects

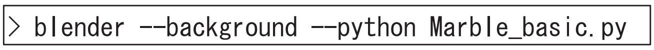

```
> blender --background --python Marble_basic.py
```

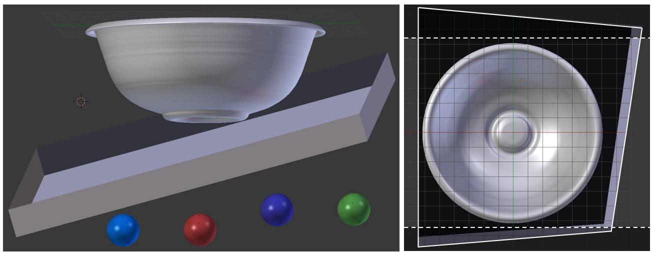
Figure 6: Batch file to control Blender by Python

As described above, it is expected that a large amount of training data can be generated efficiently by using CG tools in the following case: the target images are composed of a small number of objects; and, a large amount of different image data can be created according to a certain procedure. As a result, it is considered that the training data can be generated with CG images as substitute for actual images in a certain field, where it is difficult to accumulate the training data by actual images.

4 IMPLEMENTATION OF TRAINING DATA GENERATOR

In this study, we used 3DCD creation software Blender 2.79 [2] to generate the training data. Blender can be controlled by Python script, and Python 3.5.3 is shipped with the above-mentioned Blender. So, after we place the necessary objects and lights in Blender, the arbitrary number of training data can be generated automatically by using Python program according to the processes shown in Section 3. We created Blender objects and Python programs separately and executed them with the batch file shown in Figure 6.

(a) of Fig. 7 shows the placement of objects in Blender,



(a) Initial placement of objects in Blender

(b) Camera view

Figure 7: Initial placement of objects in Blender

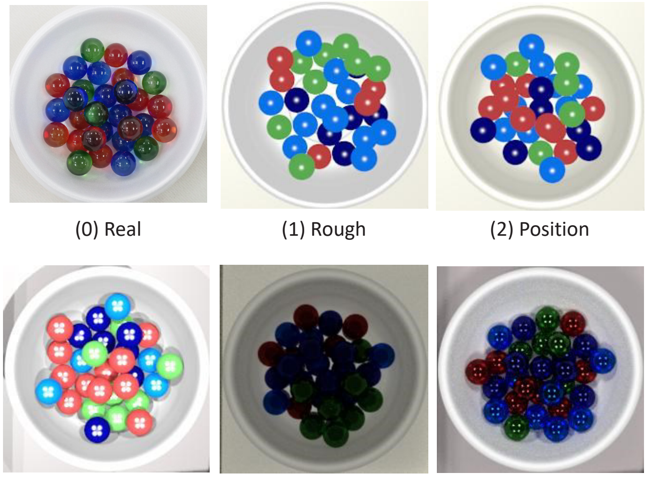
which is composed of a bowl, four marbles, and a square tray. In this experiment, same as the previous study, we used four types of marbles. So, we placed each one under the tray. And, the tray is used to detect the spillage of marble objects from the bowl object. That is since there is a possibility that the marble objects spill out in the physical simulation in the case where the number of them is large, such an image must be excluded. (b) of Fig. 7 shows the view from the camera for the rendering. The white solid quadrangle is the outline of the tray, and the lower right is the lowest position. So, the spilled marble objects gather here. And, the spillage can be detected by comparing the image before and after the simulation with trimming the bowl object. Incidentally, as for the illumination, we placed Hemi lamp, which lights up the whole area equally, and a directional lamp on the side of the camera; the range between white dashed lines in (b) of Fig. 7 was the rendering area.

The procedure to generate the images for the training data by using the batch file in Fig. 6 was as follows. Firstly, the Blender file was opened and the marble objects in (a) of Fig. 7 were randomly selected, then their copies were placed hierarchically as shown in (a) of Fig. 5. Here, to shorten the falling time, each hierarchy was divided into four quadrants, and marble objects were placed randomly within the range of each quadrant. Next, the state of (b) in Fig. 5 was generated by physical simulation. Then, rendering was performed by the camera placed right above the bowl object as shown in (b) of Fig. 7, and its image was saved in a file. After that, the Blender file was reopened to return to the initial state, then the same procedure was repeated.

After all the image data generated, they are converted to the training data by another Python program by the following procedures. Firstly, images with spilled marble objects were excluded, then the outside of the bowl object in the remained images was trimmed to create images similar to Fig. 3. Then, the designated number of images were saved as the training data. After that, by the same procedure as the previous experiment as shown in Section 2, the images were padded 4 times to create the final training data.

Table 1: Experiment cases on marbles

No	Case	Position	Illumination	Material
(1)	Rough			
(2)	Position	○		
(3)	Illumination	○	○	
(4)	Matrrial	○		○
(5)	All	○	○	○



(0) Real

(1) Rough

(2) Position

(3) Illumination

(4) Material

(5) All

Figure 8: Images of marbles by photography and CG

5 EXPERIMENTS AND EVALUATIONS

In this experiment, firstly we evaluate each factor of creating CG images from the viewpoint of the influence on the accuracy of image recognition. Next, we evaluate the change of this accuracy according to the ratio of replacing a part of the CG images with the actual images. Firstly, images with spilled marble objects were excluded, then the outside of the bowl object in the remained images was trimmed to create images similar to Fig. 3.

5.1 Evaluations of CG images for marbles

Table 1 shows the factors of creating the CG image, which is evaluated in these experiments: (2) Position, (3) Illumination and (4) Material. We evaluated the influence of the improvement of each factor, and the case of combining them. Incidentally, in Table 1, (1) Rough is the state before improvement, and (5) All is the combination case. Figure 8 shows the examples of the images of the training data of 30 marbles: (0) Real shows the actual image and the others show the CG images created with each case of Table 1. Since marble objects were randomly placed in Blender in each case, the placements of the marbles were not the same.

In (1) Rough of Fig. 8, since bowl modeling accuracy was low, the placement of marbles expanded and

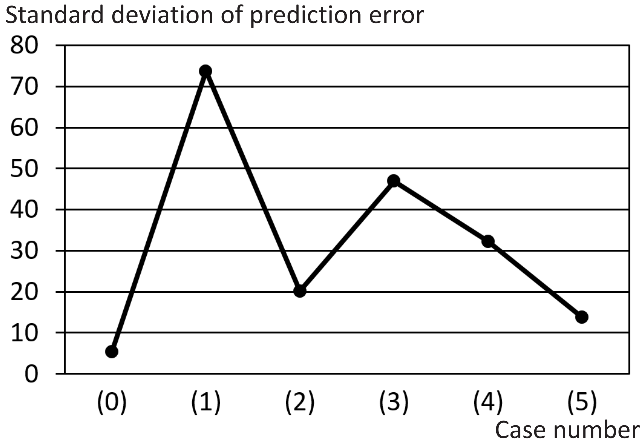


Figure 9: Change in standard deviation of errors

there were more gaps between marbles than the one in (0) Real. So, in (2) Position, we improved the shape of the bowl, then the placement of marbles became closer to (0). Next, since the actual images were taken in a room where multiple fluorescent lamps were installed on the ceiling, we placed multiple lamps in Blender to make the CG images close to the actual environment in (3) Illumination. In (4) Material, we improved only the material of the marbles from (2). In Blender, since it was necessary to change the rendering engine from "Blender rendering" to "Cycles rendering" in order to produce the marble material, the brightness of the whole image was changed in (4). In (5) All, we added the lamps of (3) to the CG image of (4).

Next, we trained the deep CNN shown in Fig. 4 by these training data. And, we prepared the test data composed of 50 actual images for each case of 5, 20, 40, 60 and 80 marbles. Then, we obtained the estimated quantity of each test data by the deep CNN. Figure 9 shows the standard deviation of the estimate error of each case of Table 1. This error is the difference between the actual quantity in each image and the estimated quantity using the deep CNN. In this case, since we classified the number of marbles from 5 to 80 every 5 and trained by the supervised learning, the error also changed in units of 5.

Here, "Case number" corresponds to "No." in the Table 1. Although the standard deviation of (0) Real was 5.3, the one of (1) Rough worsened to 73.7. And, it was bettered to 20.1 by improving the position as shown in (2). On the other hand, each improvement of (3) Illumination and (4) Material worsened rather in the case of executing separately. However, in the case of applying both shown in (5) All, it bettered to 13.8, that is, it was necessary to improve the illumination and material together.

Also, Fig. 10 shows the distribution of errors of estimated quantities with respect to each actual number of marbles in the case of (0) Real and (5) All. In (0), the error is distributed in the range of ±10 around the correct "0"; while in (5), many peaks of distribution appeared before and after "0". However, even in the latter case, though the error "over" namely 15 or more was occurred about 20% in the case of 40 marbles, the error was within the range of roughly ±10 in the case of other numbers.

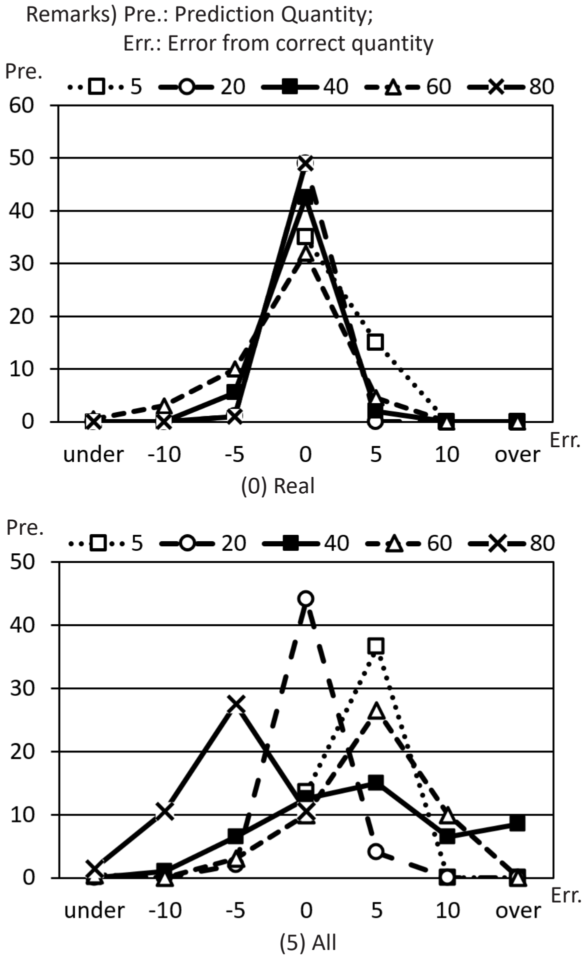


Figure 10: Distribution of errors of estimated quantities

5.2 Evaluations of CG and actual mixed images

Next, we evaluated the change of the accuracy in the case where the CG images are replaced with the actual images by each ratio. We used CG images of (5) All in Table 1 in this experiment, and the designated number of images from the end were replaced with the actual images. For example, in the case of "5%", we prepared 500 images using 425 CG images and 75 actual images. In addition, the images for evaluation are the actual images same as Section 5.1.

Figure. 11 shows the change of the standard deviation of errors according to the mixed ratio of the actual data from 5% to 20% for every 5%. The standard deviation improved to about 2/3 when 5% of the CG images were replaced with the actual images, which was about 55% improvement as for the difference between (5) and (1). Similarly, as of 10%, it improved to about half, which was about 80% as for the difference between (5) and (1). Furthermore, as of 20%, the standard deviation became

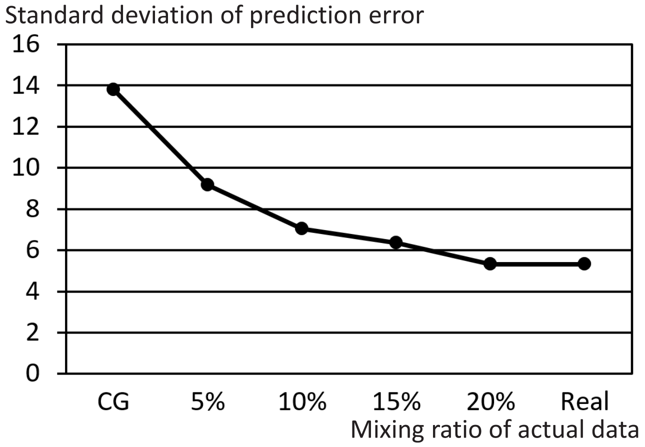


Figure 11: Standard deviation of errors on mixing ratio

almost the same as of the actual images.

That is, in the case of replacing the CG images with the actual images, the improvement of the accuracy was relatively larger than the replacement ratio.

6 DISCUSSION

In this study, we evaluated the accuracy in the case of using the CG images instead of the actual images in the training data of the deep learning. As a result, as shown in Fig. 10, we found that a certain degree of accuracy could be achieved, while the errors spread larger compared to the case of the actual images. On the other hand, as shown in the relation between the CG images of Fig. 8 and the standard deviations of the error in Fig. 9, the accuracy was greatly affected by the factors to create the CG images.

In other words, it is expected that the accuracy can be improved by refining the CG image shown in (5) of Fig. 8 to make it closer to the actual images shown in (0). However, the more we refine the CG images to look like the actual images, the more the workload becomes big. That is, it is necessary to repeat the several works accompanied by trial and error: the first is the adjustments of the CG images such as material and illumination; the second is the deep learning of the deep CNN. And since each takes a long time, the ratio of improvement and cost is expected to decline.

Therefore, we showed that the accuracy can be improved efficiently by using the training data, in which a part of the CG images were replaced with the actual images, as shown in Fig. 11. That is, we found that in the case where the training data was generated with the CG images are used, it is effective to prepare the actual images in possible and to use them as a part of the training data. In addition, at the factory like the target of this study, the operation to estimate the inventory quantities by using the deep learning as follows is possible: initially, the training data is composed of the CG images and a few actual images; while increasing the actual images, the training of the deep CNN is repeated by using these images and the accuracy can be gradually improved.

As a result, we found that even the fields where it is difficult to accumulate a large number of actual images for the deep learning, there is the case which target is composed of the simple objects from the viewpoint of CG. In such a case, we consider that to use the CG images generated automatically for the training data is effective.

However, in this study, we have verified the effectiveness of the proposed method by using only the marbles at just the laboratory. So, the verification in the actual factory remains as the future study: the first is the verification with the actual parts and in the factory environment; the second is the development of the method to collect the actual images efficiently in the factory.

7 CONCLUSION

Currently, the accuracy of image recognition utilizing the deep learning is rapidly improving, and such an image recognition is applied to various fields. On the other hand, in order to improve the accuracy by the deep learning, it is necessary to accumulate a large amount of training data. And, the preparation of the training data often becomes the obstacle to apply the deep learning.

For this problem, we proposed a method to generate the training data automatically with CG in this study. And, we confirmed that the training data can be generated by using the CG tool automatically; a certain degree of the accuracy could be achieved by using such a training data. Furthermore, we showed that the same recognition accuracy could be obtained in both cases: in one case, the training data was created by a large number of the actual images; in another case, the training data was created by a large number of the CG images and less of the actual images. As a result, we found it is effective to generate the training data by using the CG images, in some fields where it is difficult to accumulate a large number of the actual images for the deep learning.

Future studies will focus on the confirmation that this method can be used generally such as various materials and environments.

Acknowledgments

In this study, we utilized the results of graduation research of Mr. Tenma Kawanaka for the actual data of marbles and the experimental environment of the deep learning.

REFERENCES

[1] K. Banker, “MongoDB in Action,” Manning Pubns Co. (2011).

[2] Blender Foundation, “Blender,” https://www.blender.org/ (reffered, May 2018).

[3] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, “PCANet: A simple deep learning baseline for image classification?,” IEEE Transactions on Image Processing, Vol. 24, No. 12, pp. 5017–5032 (2015).

[4] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” 2012 IEEE conference on computer vision and pattern recognition (CVPR), pp. 3642–3649 (2012).

[5] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique,” IEEE Transactions on Medical Imaging, Vol. 35, No. 5, pp. 1153–1159 (2016).

[6] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, “Deep learning,” MIT press (2016).

[7] T. Kawanaka, and T. Kudo, “Proposal of Required Amount Satisfaction Determination Method Using Deep Learning,” Proc. 80th national convention of IPSJ, pp. 4-543–4-544 (2018) (in Japanese).

[8] T. Kawanaka, and T. Kudo, “Inventory Satisfaction Discrimination Method Utilizing Images and Deep Learning,” Procedia Computer Science (2018) (in press).

[9] T. Kudo, Y. Ito, and Y. Serizawa, “An Application of MongoDB to Enterprise System Manipulating Enormous Data,” Int. J. Informatics Society, Vol. 9, No. 3, pp. 97–108 (2017).

[10] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” Frontiers in plant science, Vol. 7, Art. 1419 (2016).

[11] E. Redmond, and J.R. Wilson, “Seven Databases in Seven Weeks: A guide to Modern Databases and the NoSQL Movement,” Pragmatic Bookshelf (2012).

[12] S. P. Singh, “Production and Operation Management,” Vikas Publishing House Pvt Ltd (2014).

[13] E. A. Silver, D. F. Pyke, and R. Peterson, “Inventory management and production planning and scheduling,” John Wiley & Sons (1998).

[14] M. Verhelst, and B. Moons, “Embedded Deep Neural Network Processing: Algorithmic and Processor Techniques Bring Deep Learning to IoT and Edge Devices,” IEEE Solid-State Circuits Magazine, Vol. 9, No. 4, pp. 55–65 (2017).

Applying SAW to Regression Verification for C functions with Recursive Data Structure

Kozo Okano[†], Satoshi Harauchi[‡], Shin Maruyama[†], and Shinpei Ogata[†]

[†]Faculty of Engineering, Shinshu University, Japan

[‡]Mitsubishi Electric, Japan

okano@cs.shibshu-u.ac.jp, ogata@cs.shinshu-u.ac.jp

Abstract - Programs are usually revised due to performance improvements. In such cases, programmers have to ensure that the revised program also preserves the behavior of the previous version of the program. Regression testing is performed to check whether the both of the revised version and the previous version have the same behavior. It, however, needs much time and a lot of test-cases. Tools based on formal method will reduce the costs. They ensure that given two programs output the same results for the same inputs based on logical analysis of their source code and perform effective path search using SAT/SMT solvers. SAW, a novel tool based on formal methods, can check whether given two functions in C act in the same behavior (conformance verification). It, however, has a limitation that it cannot check function dealing with data structure. This paper proposes a new technique for conformance verification on C functions with data structures using SAW. The technique is based on a kind of bounded model checking. It also reports results on performance evaluation.

1 INTRODUCTION

C Programs are widely used and they are usually revised due to performance improvements. In such cases, programmers have to ensure that the revised program also preserves the behavior of the previous version of the program. Usually, regression testing is performed to check whether the both of the revised version and the previous version have the same behavior. It, however, needs much time and a lot of test-cases.

Formal Approach Techniques might help to reduce such costs. Based the approach several tools are developed. Such techniques exhaustively check whether the given two programs have always the same output for every same input. Thus, these tools will find potential bugs or firm confidence on the conformance with adequate efficiency. We call this kind of verification formal conformance verification (FCV).

Recent tools, however, don't fully support program dealing dynamic data structures especially recursive data structures. Such a program sometimes suffers halting problem in computability theory.

In this paper, we firstly propose a method for FVC for a program with recursive data structures. In order to avoid the halting problem, the method is based on bounded model verification technique [1], [2]. We also perform experimental evaluation using SAW (Software Analysis Workbench)[3]. SAW is a recent formal verification tool. We use SeaHorn[4] to compare with SAW. SeaHorn is also a recent formal verification tool for C language.

The rest of this paper organized as follows. Section 2 gives preliminaries. Section 3 describes the proposed method. Section 4 gives experimental evaluation. Section 5 discusses the results. Finally, Section 6 summarizes this paper.

2 PRELIMINARY AND RELATED WORK

In general, testing is the historical and popular method to check the quality of source code. For conformance testing, regression testing is widely used. Regression testing, however, costs time and workload. It also has the disadvantage that the verification becomes incomplete.

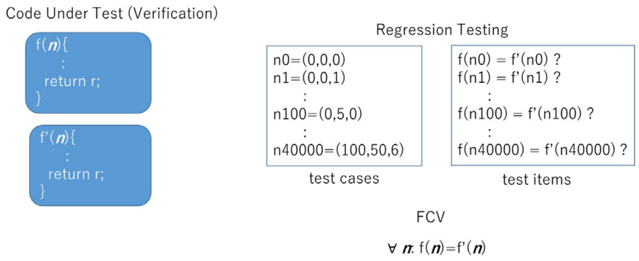


Figure 1: Regression Testing versus FCV

Figure 1 summarizes regression testing and FCV. As a given Code Under Test, or CUV Code Under Verification (CUT), let assume that two functions f and f' exist. We want to check that for any input $\boldsymbol{n}$, $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ holds.

In regression testing, we have to prepare enough size of test cases (in Figure 1, we have 40000 cases) and check all cases by executing a test driver. As we see, regression testing costs a lot of time and it does not cover whole of input cases.

2.1 Formal Conformance Verification

The following two theorems are well-known results in Computation Theory[5].

Theorem 1 (Termination Problem (Halting Problem)) The Termination Problem is undecidable.

In other words, $\forall f$ and $\forall \boldsymbol{n} \in \mathbb{Z}^{|\boldsymbol{n}|}$, whether $f(\boldsymbol{n})$ always terminates or not, is undecidable.

Theorem 2
General conformance checking problem is undecidable.

In other words, $\forall f, f'$ and $\forall \boldsymbol{n} \in \mathbb{Z}^{|\boldsymbol{n}|}$, whether $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ always holds or not, is undecidable.

If we restricted the condition, General conformance checking problem can be decidable. Thus, Restricted conformance checking problem is decidable.

Theorem 3
Restricted conformance checking problem is decidable.

$\forall f, f'$ and $\forall \boldsymbol{n} \in \mathbb{Z}_t^{|\boldsymbol{n}|}$, whether $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ always holds or not, is decidable, provided that both f and f' terminate for any $\boldsymbol{n} \in \mathbb{Z}_t^{|\boldsymbol{n}|}$, where $\mathbb{Z}_t$ is a whole set of t-bit integers for some fixed parameter t.

Proof 3
The size of $\mathbb{Z}_t$ is 2^t. Therefore, the size of $\mathbb{Z}_t^{|\boldsymbol{n}|}$ is at most $2^{t|\boldsymbol{n}|}$. The assumption grantees that we think only functions f and f' that always terminate for any $bmn \in \mathbb{Z}_t^{|\boldsymbol{n}|}$. Thus, we can compute the result of $f(\boldsymbol{n})$ and $f'(\boldsymbol{n})$ in finite steps $\alpha(\boldsymbol{n})$ and $\alpha'(\boldsymbol{n})$, respectively. In conclusion, we can deside if $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ always holds, in finite steps $2^{t|\boldsymbol{n}|} \cdot \max(\alpha(\boldsymbol{n}), \alpha(\boldsymbol{n}))$. □

In FCV, we check logical expression $\forall \boldsymbol{n} : f(\boldsymbol{n}) = f'(\boldsymbol{n})$, where functions f and f' are expressed in some logical clauses derived from CUV.

Note that usually $\boldsymbol{n}$ is a vector of bounded integers, such as a 32-bit integer, thus, the number of check cases is finite.

Form Theorem 3, if we suppose that functions f, and f' always terminate then the expression can be efficiently checked using SAT/SMT solvers[6]–[12].

SAW and SeaHorn[4] are recently appeared tools to efficiently check all inputs cases.

2.2 SAW

SAW (Software Analysis Workbench)[3] is developed by Galois inc. It is an open source software. It verifies the code written in C or Java using a compiler that generates LLVM, or JVM (Java Virtual Machine) . Some recent formal verification techniques [13]–[15] use JVM and LLVM as their targets. An LLVM file is compiled from a C, C++, or Objective-C source file. LLVM is a virtual machine instruction set (Intermediate Representation) and usually used for code optimization in compilers. It, therefore, supports three-address code scheme and Static Single Assignment form, which facilitate static analysis for optimizing compiled code LLVM has pointer types as well, which is mandatory for compliers of C-family languages.

SAW supports equivalence checking between two C functions given in the LLVM format. Both symbolic execution and equivalence checking functions are provided as commands of a script language used in SAW. SAW also supports property checking. SAW has been successfully applied to security domains such as Cryptographic Protocol Analysis.

Figure 2 shows the architecture of SAW.

The followings summarizes the feature of SAW.

- It uses its own verification script called SAWScript, which is a kind of functional program languages.
- It has several verification packages that support each of their specific fields.
  - llvm_extract
  - llvm_symexec
  - llvm_verify
  - crucible_llvm_verify
- It has a build-in solver, ABC[17]. It can also use three SAT/SMT solvers, Z3[6], Yices[7], and CVC4[8].
- It can generate proof constraints with a form of aig, and smtlib2[18]. Using the format file, other external solvers can be available.

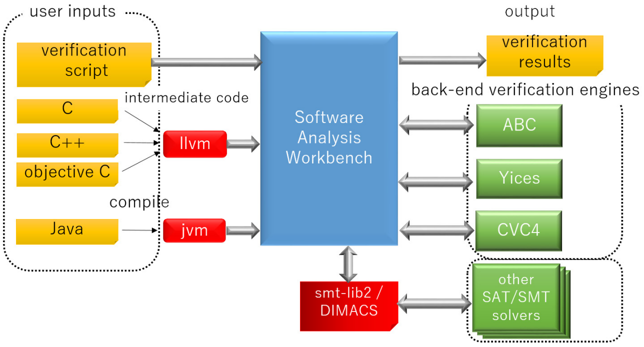


Figure 2: The architecture of SAW

Package llvm_symexec is the original package SAW used. Package crucible_llvm_verify is recently provided. It support pointers and data structures in C.

2.2.1 Verification Examples using SAW

The following example shows verification process for two functions that output twice of the input values. See Listing 1, 2, and 3

Listing 1: Twice Program

```
// reference function
unsigned int reference_function(unsigned int x){
  return x * 2;
}

// implementation
unsigned int implementation_function(unsigned int x){
  return x << 1;
}
```

List1 has two functions reference_function() and implementation_function(). Function reference_function() just outputs the value of input multiplied by two, while implementation_function() outputs arithmetic left shift of the input value by 1 bit. Though the code differs from each other, those functions output the same value for any value of the same input. Proof scripts Listing 2 and 3 prove by different approaches.

Listing 2: Verification script for twice program(llvm_symexec)

```
// llvm_symexec
// .bc is llvm format
load <- llvm_load_module ”add.bc”;

// reference_function
// variable x is defined as 32bit integer
x <- fresh_symbolic ”x” {| [32] |};
// alloc is used when pointer is used
let alloc_ref  = [];
//
let input_ref  = [(”x”, x, 1)];
//
let output_ref = [(”return”, 1)];
```

```
t1 <-
  llvm_symexec load ”reference_function”  alloc_ref
  input_ref output_ref true;

// implementation_function
let alloc_imp  = [];
let input_imp  = [(”x”, x, 1)];
let output_imp = [(”return”, 1)];

t2 <-
  llvm_symexec load ”implementation_function” alloc_imp
  input_imp output_imp true;

// verification
thm <- abstract_symbolic {{ t1 == t2 }};
result <- prove z3 thm;
//
print result;
```

Listing 3: Verification script for twice program(crucible_llvm_verify)

```
// crucible_llvm_verify
// add.bc
load <- llvm_load_module ”add.bc”;

// reference_function
let add_setup = do {
//
    x <- crucible_fresh_var ”x” (llvm_int 32);
//
    crucible_execute_func [crucible_term x];
//
    crucible_return (crucible_term {{ x << 2 : [32] }});
};

crucible_llvm_verify load ”reference_function” [] false
  add_setup abc;
```

Listings 4 and 5 show the results, respectively.

Listing 4: Result(SAW:llvm_symexec)

```
$saw llvm_symexec.saw
Loading file ”llvm_symexec.saw”
Running reference_function
Finished running reference_function
Running implementation_function
Finished running implementation_function
Valid
```

Listing 5: Result(SAW:crucible_llvm_verify)

```
$saw crucible_llvm_verify.saw
Loading file ”crucible_llvm_verify.saw”
Proof succeeded! @reference_function
Running reference_function
```

Messages “Valid” and “Proof succeeded! @ reference_function” show that the two functions have the same behaviors, for llvm_symexec and crucible_llvm_verify, respectively.

Listings 6, 7, and 8 show the case that FCV outputs counter-examples.

Listing 6: Wrong implimented code

```
// Reference Function
unsigned int reference_function(unsigned int x){
  return x * 2;
}

// Implementation
// (llvm_symexec)
unsigned int implementation_function(unsigned int x){
  if(x == 10){
    return x * 3;
  }
  return x << 1;
}
```

Listing 7: Result(SAW:llvm_symexec)

```
$saw llvm_symexec.saw
Loading file ”llvm_symexec.saw”
Running reference_function
Finished running reference_function
Running implementation_function
Finished running implementation_function
prove: 1 unsolved subgoal(s)
Invalid: [x = 10]
```

Listing 8: Result(SAW:crucible_llvm_verify)

```
$saw crucible_llvm_verify.saw
Loading file ”crucible_llvm_verify.saw”
Subgoal failed: @reference_function safety assertion:
 literal equality postcondition
SolverStats {solverStatsSolvers = fromList [”ABC”],
 solverStatsGoalSize = 60}
———————Counterexample———————
(”x”,10)
user error (”crucible_llvm_verify”
(crucible_llvm_verify.saw:8:1-8:21):
Proof failed.)
```

For both cases, when x equals to 10, the behavior differs. SAW correctly shows the counter-examples. This is the most useful advantage of SAW.

2.2.2 ABC

A user of SAW can analyze the LLVM using symbolic execution. The result of the execution is stored in AIG (And-Inverter Graphs) [16]. AIG data can be verified by a theorem prover called ABC[17]. ABC is especially good at equivalence checking [19] between two functions represented in AIG. ABC is the default solver for SAW.

2.3 SeaHorn

SeaHorn[4] also verifies C program code using LLVM. SeaHorn has the following features.

- It is easy to use because programmer can directly write assertions in the code. The notation based on a simple notion of Design by Contract[20].
- Learning times for the tool is shorter than other formal based tools.

2.3.1 Verification Example using SeaHorn

Listing 9 shows an example of verification using SeaHorn.

Listing 9: Verification script for twice program(seahorn)

```
#include ”seahorn/seahorn.h”
extern int nd(void);

// code under verification
unsigned int implementation_function(unsigned int x){
  return x << 1;
}

int main(){
  int x, val;

  x = nd();
  val = nd();
  val = implementation_function(x);

  // assrtion
  sassert(val == x * 2 );
}
```

Here, function nd means non-deterministically. It returns arbitral value. sassert(P) states that P is true.

Listing 10 shows the result of the verification.

Listing 10: Result(seahorn)

```
$sea pf double.c --show-invars
  -- omit --
unsat
Function: main
main@entry: true
main@entry.split: true
```

Note that usually SeaHorn checks unsatisfiablity. In other words, unsat is printed if and only if the assertion holds in SeaHorn.

2.4 SAT/SMT solvers

Recently efficient SAT (SATisfiability problem) solvers are emerging and these solvers prove a lot of constraint base problems. Satisfiability problem is usually given as a set of clauses, where each clause is logical disjunction of Boolean variables. the set of clause is evaluated as logical conjunction of clauses. Therefore, the set can be evaluated as satisfiable or unsatisfiable. Satisfiability problem is known as a NP-complete. However, SAT solvers efficiently proves most of instances.

SMT (Satisfiability Modulo Theories) is an extension of SAT. Each Boolean variable is substituted for inequality over integers or reals. Several classes are known for SMT. Some of these classes are decidable and there are tools which can efficiently proves.

2.4.1 Z3

Z3[6] is one of famous SMT solvers developed by Microsoft Research. In SMT-COMP, an SMT solver competition, it has excellent results in every year. It is one of the built-in SMT solvers by SAW.

2.4.2 CVC4

CVC4[8] is one of the CVC (Cooperating Validity Checker) series in the flow of the theorem proving system SVC developed by Stanford University. In SMT-COMP 2017, it won the victory in many divisions.

2.4.3 Yices

Yices[7] is an SMT solver developed at SRI and was upgraded as Yices 2 since 2009. It also has excellent results with SMT-COMP 2017. It is one of the built-in SMT solvers by SAW.

2.4.4 SMT-RAT

SMT-RAT[10] is an SMT solver that can perform parallel processing written in C ++. Since SAW is not built in as standard, it is necessary to output a file such as smtlib2.

2.4.5 minisat

Minisat[11] is one of the representative SAT solvers. It has the minimum set of functions as SAT solver, and its source code is about 2000 lines. Because SAW does not built in it as standard, it is necessary to apply minisat after outputting in Conjunctive Normal Form (CNF) or AIG[16] format file format.

3 OUR PROPOSED METHOD

Program code with recursive data structures has a loop structure which has a termination condition. The termination condition depends on the recursive data structures, thus we cannot bound the number of iteration a fixed finite value. For this reason, verification on such program code faces so-called termination problem.

In other words, verification on program code with recursive data structures is essentially undecidable.

In order to overcome this problem, in practice, we usually approximate the problem by introducing the idea of bounded verification.

Our proposed method also uses bounded verification by bounding the size of recursive data structures.

3.1 The idea

Bounded verification usually terminates iteration of loop body by a fixed value. We limit the size of recursive data structures. This is essentially the same idea of the usual manner of bounded model checking approaches.

For example, let us consider a linear list. We fixed the size of list namely n. We produce every pattern of verification script for the data structure with in size n. See Figure 3.

Then we perform each formal verification for the produced scripts.

For example, we choose 100 as n. Then we perform formal verification with size 1 to 100 of the liner list. If all of the verification passed, we strongly assume that the program is valid for any size of a list.

The scheme has advantage that we can choose any feasible value of n, but as we notice, the verification time becomes large as n becomes large.

In similar way, for fixed vaule n, we produce every pattern of binary trees with size 1 to n, where n is the number of nodes in the binary tree. Figure 4 shows the every pattern of binary trees with size of 3.

For every pattern we perform each formal verification for the produced script.

We use the above idea with package crucible_llvm_verify for SAW. We evaluated the effectiveness of our proposed method.

3.2 Verification Subjects

We perform experiment for the following three data structures.

- Two-level nests
- Linear list with recursive definition

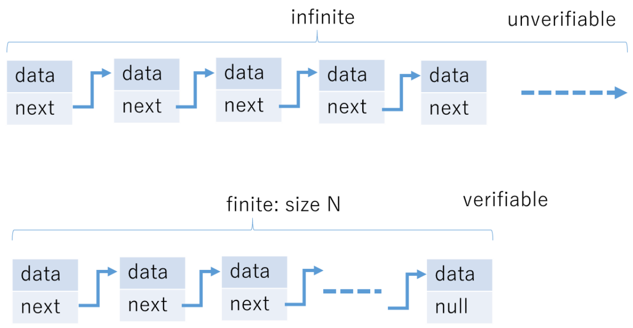


Figure 3: Bounded Model Checking

The upper of Fig 3 shows that verification does not end due to infinite size of the list, while lower of Fig 3 shows that verification terminates due to the specifying the limit of size.

- Binary tree with recursive definition

In this section, we show every programs under verification.

3.2.1 Two-level nests

The program calculates summation of 32-bit integers in the parent and children nodes (Figure 5).

Listing 11 shows the data structure.

Listing 11: Two-level nest

```
typedef struct s {
    int a;
} s_t;
//parent
typedef struct t {
    int x;
    s_t n;
} t_t;
//reference function
int f_ref(t_t *p) {
    return (p->n).a + p->x;
}
//implementation
int f_imp(t_t *p) {
    return  p->x + (p->n).a;
}
```

The difference between the reference function and the implementation is trivial. We just change the left term and the right term.

3.2.2 Linear List

The program calculates summation of 32-bit integer in every cell of the list (Figure 6).

Listing 12 shows the data structure of the program.

Listing 12: Lienear List

```
struct NODE{
  uint32_t val;
  struct NODE* next;
};
typedef struct NODE* node_t;
// reference function
int linear1(node_t node){
  if(node->next == NULL){
    return node->val;
  }else{
    return node->val + linear1(node->next);
  }
}
// implementation
int linear2(node_t node){
  if(node->next != NULL){
    return node->val + linear2(node->next);
  }else{
    return node->val;
  }
}
```

The difference between the reference function and the implementation is that the form of if statement.

3.2.3 Binary Tree

The program calculates summation of 32-bit integer in every node of the binary tree (Figure 7).

Listing 13 shows the data structure of the program.

Listing 13: Binary Tree

```
// node
struct BTREE {
    uint32_t val;
    struct BTREE* left;
    struct BTREE* right;
};
// reference function
uint32_t pre_order(struct BTREE* tree){
    if(tree == NULL){
        return 0;
    }
    return pre_order(tree->left) +
    pre_order(tree->right) + tree->val ;
}
// implementation
uint32_t in_order(struct BTREE* tree){
    if(tree == NULL){
        return 0;
    }
    return in_order(tree->left) +
    tree->val + in_order(tree->right);
}
```

The difference between the reference function and the implementation is the order of traversal. Thus the difference is not trivial.

4 EXPERIMENTAL EVALUATION

The environment is summarized as follows.

- OS: Windows 10 64 bit
- CPU: Intel core i7-4500U CPU @ 1.80GHz 2.39GHz
- Memory: 8.00 GB
- Docker
  - version: 18.01.0-ce
  - Memory: 4096 MB
  - The number of CPUs: 2
- SAW: 0.2 (2018-01-31)
  - LLVM: 3.8.0
  - Z3: 4.5.0
  - Yices: 2.5.2
  - minisat: 2.2.0
  - SMT-RAT: 2.1.0
- SeaHorn: 0.1.0-rc3
  - LLVM: 3.6.0

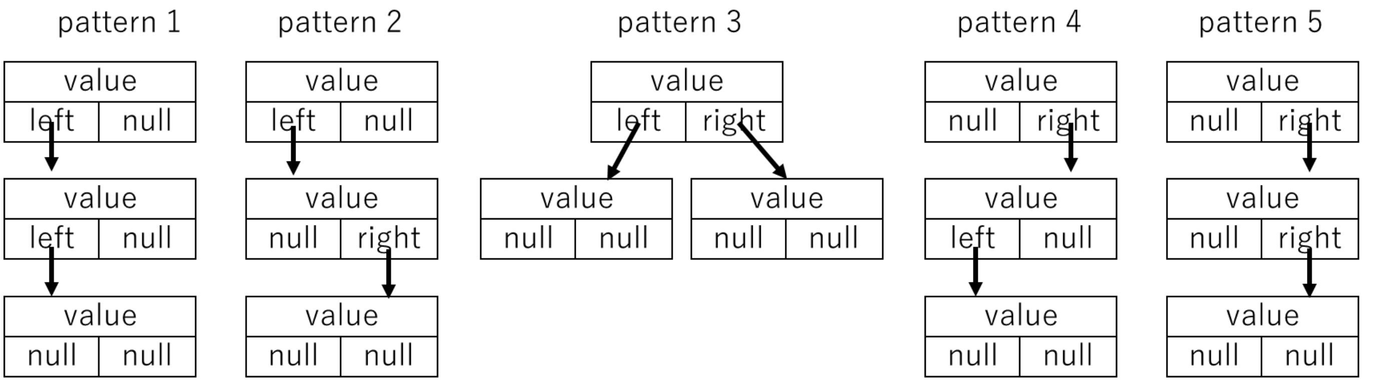


Figure 4: Patterns of Binary Trees

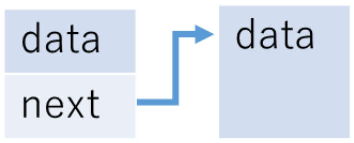


Figure 5: Two level nest

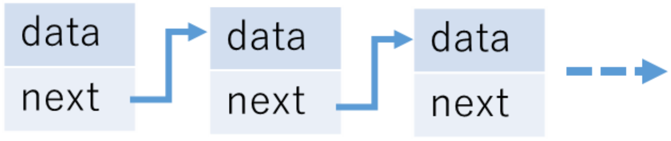


Figure 6: Linear List

4.1 Comparison of SMT Solvers: EXP 1

We evaluate verification results and CPU times using the built-in function "Output a file for SAT/SMT solver" of SAW. For ABC, we use a proof package named crucible_llvm_verify. For other SAT/SMT solvers, we use a proof package named llvm_symexec.

4.2 Comparison of SAW and SeaHorn: Exp 2

We evaluate verification results and CPU times using SAW and SeaHron. We use a proof package named crucible_llvm_verify.

4.3 THE RESULTS

For all results, T/O specifies that verification time is over than 3,600 sec. Symbol '–' shows failure of verification. Unit of time is second.

4.3.1 Comparison of SAT/SMT solvers

Table 1 summarizes verification of Two-level nest with varying SAT/SMT solvers.

Table 2 summarizes verification of Linear List of size 100 with varying SMT/SAT solvers.

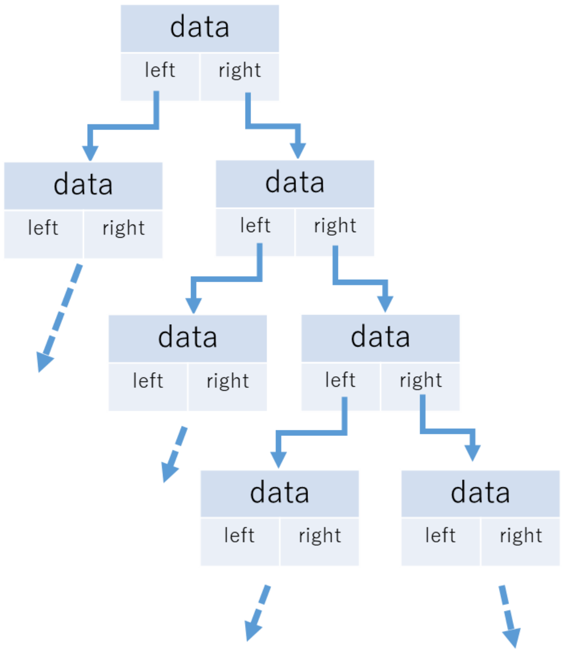


Figure 7: Binary Tree

Table 3 summarizes verification of Binary Tree of depth 5 with varying SAT/SMT solvers.

4.3.2 Comparison with SAW and SeaHorn

Table 4 summarizes verification of Two-level nests with varying SAT/SMT solvers.

Table 5 summarizes verification of Linear List of size 100 with varying SAT/SMT solvers.

Table 6 summarizes verification of Binary Tree of depth 5 with varying SAT/SMT solvers.

5 DISCUSSION

We verified by applying bounded verification to functions that deal with structures including recursion. We can verify two-level nest and linear list structure correctly using cru-

Table 1: Results for Two-level nests

SMT/SAT	ABC	Z3	Yices	minisat	SMTRAT
CPU time	1.06	1.23	1.47	1.24	T/O

Table 2: Results for Linear List

SMT/SAT	ABC	Z3	Yices	minisat	SMTRAT
CPU time	1.47	–	–	–	–

cible_lvm_verify. SeaHorn can only verify two-level nest. It was not possible to verify the binary tree structure by all verification methods.

In Exp1, in order to output smtlib2 format, we have to use "llvm_verify." However llvm_verify does not support data structure, we failed to verify with SAT/SMT solvers. On the other hand, verification succeeded when we used "crucible_lvm_verify."

Timeout has occurred in SMTRAT, this is presumed to be a file format not supported by SMTRAT.

In Exp2, it was possible to verify Linear List structure with 100 elements. When we investigated how many elements it could go up to, the number of elements was about 5000. In realistic verification, since sufficient verification can be performed even with the number of elements of the list structure up to 1000, the bounded verification method can be applied. The timeout occurred in the verification of the program of the binary tree structure, but it seems that the binary tree structure is more complicated in structure than the linear List structure and the verification becomes difficult by the depth of the binary tree (the number of elements).

In the verification using SeaHorn, the verification time for Two-level nest is about one tenth of the SAW and therefore it is superior to the SAW in view of the verification time. However, when trying to verify a program dealing with recursive structure, recursive functions. it automatically skips the analysis of the structure. Thus, it is impossible to handle programs handling recursive structures. As a result, it can be said that SAW that can handle recursive structures is superior to SeaHorn at the present.

6 CONCLUSION

This paper proposed a new method for Formal Conformance Verification based on bounded model checking for programs with recursive data structures. We also conducted experimental evaluation using SAW. It shows that the proposed method works well for a simple program which deals with calculation over a linear list.

As future work, we want to show better performance for binary trees and other complex data structures.

Acknowledgment

Funding from Mitsubishi Electric Corp. is gratefully acknowledged. The authors wish to thank Mr.Karashima for assistance of new experiments.

Table 3: Results for Binary Tree

SMT/SAT	ABC	Z3	Yices	minisat	SMTRAT
CPU time	T/O	–	–	–	–

Table 4: Results for Two-level nests (SAW and SeaHorn)

Verification Tool	SAW	SeaHorn
CPU Time	1.06	0.104

REFERENCES

[1] E. Clarke, A. Biere, R. Raimi, Y. Zhu: "Bounded Model Checking Using Satisfiability Solving," Formal methods in system design, Vol.19 Issue 1, pp.7-34 (2012)

[2] T. Liu, M. Nagel, and M. Taghdiri, "Bounded Program Verification using an SMT Solver: A Case Study," Proceedings of the 5th International Conference on Software Testing, Verification and Validation, pp.101-110 (2012)

[3] R. Dockins, A. Foltzer, J. Hendrix, B. Huffman, D. McNamee, and A.Tomb: "Constructing Semantic Models of Programs with the Software Analysis Workbench," Proceedings of VSTTE 2016 (2016)

[4] A. Gurfinkel, T. Kahsai, A. Komuravelli, and J. A. Navas: "The SeaHorn Verification Framework," International Conference on Computer Aided Verification, pp.343-361 (2015)

[5] J. E. Hopcroft, R. Motwani, and J. D. Ullman: "Introduction to Automata Theory, Languages, and Computation (2nd Edition)," Addison Wesley (2000)

[6] L. de Moura and N. Bjørner: "Z3: An efficient SMT solver," Proceedings of TACAS 2008, LNCS Vol. 4963, pp.337-340 (2008)

[7] B. Dutertre: "Yices 2.2," Proceedings of CAV2014, LNCS Vol.8559, pp.737-744 (2014)

[8] C. Barrett, C. L. Conway, M. Deters, L. Hadarean, D. Jovanović, T. King, A. Reynolds, and C. Tinelli: "CVC4," Proceedings of the 23rd international conference on Computer aided verification (CAV'11) pp. 171-177 (2011)

[9] A. Cimatti, A. Griggio, B. Schaafsma, and R. Sebastiani: "The MathSAT5 SMT Solver," Proceedings of TACAS 2013, LNCS Vol. 7795, pp.93-107 (2013)

[10] F. Corzilius, G. Kremer, S. Junges, S. Schupp, and E. Abraham: "SMT-RAT: An Open Source C++ Toolbox for Strategic and Parallel SMT Solving," Proceedings of the International Conference on Theory and Applications of Satisfiability Testing (SAT 2015) pp.360-368 (2015)

[11] N. Een, A. Mishchenko, and N. Sörensson: "Applying Logic Synthesis for Speeding Up SAT," Proceedings of the 10th International Conference on Theory and applications of satisfiability testing pp. 272-286 (2007)

[12] A. Biere, M. Heule, H. Van Maaren, and T. Walsh: "Handbook of Satisfiability," IOS press (2009)

Table 5: Results for Linear List (SAW and SeaHorn)

Verification Tool	SAW	SeaHorn
CPU time	1.47	–

Table 6: Results for Binary Tree (SAW and SeaHorn)

Verification Tool	SAW	SeaHorn
CPU time	T/O	–

[13] K. Okano, S. Harauchi, T. Sekizawa, S. Ogata, and S. Nakajima: "Equivalence Checking of Java Methods — Toward Ensuring IoT Dependability —," Proceedings of the 26th International Conference on Computer Communications and Networks, pp.1-6 (ICCCN 2017) (August 2017)

[14] C. Belo Lourenco, Si-Mohamed Lamraoui, S. Nakajima, and J. Sousa Pinto: "Studying Verification Conditions for Imperative Programs," Proceedings of 15th International Workshop on Automated Verification of Critical Systems, AVoCS'15 (2015)

[15] Si-Mohamed Lamraoui, S. Nakajima: "A Formula-based Approach for Automatic Fault Localization of Multi-fault Programs," Journal of Information Processing, Vol.24, No.1, pp.88-98 (2016)

[16] A. Darringer, W. H. Joyner, Jr., C. L. Berman, and L. Trevillyan: "Logic synthesis through local transformations," IBM Journal of Research and Development, Vol.25 (4), pp.272-280 (1981)

[17] R. Brayton and A. Mishchenko: "ABC: An Academic Industrial-Strength Verification Tool," LNCS Vol.6174, pp.24-40 (2010)

[18] C. Barrett, A. Stump and C. Tinelli: "The SMT-LIB Standard Version 2.0," (2010)

[19] A. Kuehlmann, V. Paruthi, F. Krohm, and M. K. Ganai: "Robust boolean reasoning for equivalence checking and functional property verification," IEEE Transaction on CAD, vol.21 (12), pp.1377-1394 (2002)

[20] B. Meyer: "Object-Oriented Software Construction, second edition," Prentice Hall (1997)

Occurring Frequency for Selecting Optimal Window Length in Motif Discovery

Makoto Imamura*, Mao Inoue*, Tadao Yagou*, Daniel Nikovski**

** School of Information and Telecommunication Engineering, Tokai University, Japan
** Mitsubishi Electric Research Laboratories, USA
imamura@tsc.u-tokai.ac.jp

Abstract - Repetitive patterns in time series obtained from sensors attached to humans or machines often show typical behavior, and finding such repetitive patterns is useful in various domains such as smart factory, health care, and seismology. Motif discovery is not only a fundamental method for finding repetitive subsequences in a longer time series, but is also used as a sub-routine in higher-level analytics including classification, clustering, visualization, and rule-discovery. However, existing motif discovery algorithms depend critically on knowledge of the correct subsequence length. Therefore, deciding on a suitable subsequence length is necessary before using those algorithms. In this work, we investigate criteria how suitable a subsequence is for a motif in order to select top K motifs among subsequences with different window lengths. We propose a new index 'occurring frequency' that counts the number of similar subsequence occurrences with removing trivial matching subsequences. We also show normalized motif occurring frequency, which we call 'motif coverage' in this paper, can decide optimal window lengths of motifs in simulated time series containing motifs with different window lengths.

Keywords: Time series data mining, Motif discovery

1 INTRODUCTION

Time series motifs [1][2] are approximately repeating subsequences embedded in a time series. Motifs are one of the most important primitives in time series data mining, and motif discovery has been used as a sub-routine in higher-level analytics, including classification, clustering, visualization and rule-discovery. Moreover, motif discovery has been applied to domains as diverse as factory operation [3], medicine [4], and seismology [5]. The notion of a motif is useful for a wide range of applications, because a repeated and frequently occurring pattern implies a latent system that occasionally produces a repeatable output. For example, this system may be an over-caffeinated heart, sporadically introducing a motif pattern containing an extra beat [6], or the system may be a factory worker, producing repetitive movement in a series of assembly operations [3].

Since the Matrix Profile [7], a fast and scalable algorithm subsequence all-pairs-similarity-search in time series was introduced, it has helped to develop new innovative ideas for time series data mining [8]. However, as a motif is defined as a pair of subsequences the distance between which is the smallest, it does not necessarily imply that similar subsequences to the motif occurred frequently in a time series. That is, there are not necessarily very many subsequences in the neighborhood of a motif. Furthermore, motif discovery algorithms expect that a subsequent length be chosen beforehand, which usually means in practice that users must try several possible lengths, and must confirm that the discovered motif indeed has frequent similar subsequences in a time series.

In this work, we introduce a new index 'occurring frequency' to measure the number of subsequences within a given range from the motif. We will further define a novel index '*motif coverage*' by normalizing occurring frequency with respect to a window length so as to decide which the best motif is among motifs with different lengths.

2 RECONSIDERRATION ON THE DIFINITION OF MOTIF

This section describes the commonly used definition of motif and summarizes the problems for deciding optimal length of motif.

2.1 Definition of a Motif

Motif is defined by using nearest neighbor distance in the space consisting of subsequences in a time series.

Definition: time series X

A *Time Series* $X=[x_1,\cdots,x_m]$ is a continuous sequence of real values. We denote the value of the i-th time point as $X[i] = x_i$.

Definition: subsequence $X[p:q]$

A *subsequence* $s = [x_p, x_{p+1},...,x_q] = X[p:q]$ is a list which consist of continuously occurring values in X, starting at position p and ending at position q. We also denote a subsequence $X[p:q]$ as $X_w(p)$, which means a subsequence staring at p with length w.

The *length* w of a subsequence S is $w = q - p + 1$ and we denote it as *length*(s).

Definition: support of a subsequence

The *support* of a subsequence s is a set of time at which a subsequence s have a value. The support of $X[p:q]$ is $[p:q] = [p, p+1,..., q-1, q]$ and we denote it as *support* (s).

Definition: subsequence space $S_w(X)$

Subsequence space is the set of all the subsequences with length w in a time series X. We denote it as $S_w(X)$. Subsequence space $S_w(X)$ is a w dimensional vector space. Therefore, for given subsequences s_i and s_j, the distance between s_i and s_j, which we denote as $\text{dist}\left(s_i, s_j\right)$, can be defined the same as vector space. We use L_1 distance in this paper.

$$dist\left(X_w(p), X_w(q)\right) \equiv \sum_{1}^{w} |X(p+i-1) - X(q+i-1)|$$

Definition: disjoint subsequence

Let s_i and s_j be subsequences. When *support* (s_i) and *support* (s_j) are disjoint, that is *support* (s_1) ∩ *support* (s_2) $= \emptyset$, we call that s_i and s_j are disjoint.

Definition: Motif subsequence (1-NN)

Let w be a window length and X be a time series. A subsequence s with length w in X which satisfies the below condition is called a *motif.*

There is a subsequence s' with length w, such that

$$\text{dist}(s, s') = \min_{i,j} \{ \text{dist}(s_i, s_j) \mid s_i, s_j \in S_w(X) \text{ and } \text{support}(s_i) \cap \text{support}(s_j) = \emptyset \}$$

The above definition is based on one nearest neighbor (1-NN) distance. We can extend this definition for k nearest neighbor distance by replacing minimum with k-th minimum in the above condition.

2.2 Problems in Defining Motif Criteria

The intuitive meaning of a motif is a subsequence which has many similar subsequences in a time series X. The first idea for defining motif criteria, which are indexes to measure how suitable a subsequence is for a motif, is to give indexes to measure 'similar' and 'many' in the above intuition. A similar sequence can be measured by a distance between subsequences. With regard to the definition of "many", it seems to be possible to count the number of subsequences which are similar to a given motif. We call this number "occurring frequency". There are several problems for defining occurring frequency and motif criteria. Those problems can be summarized in the following three points.

(1) Error dependency

When we call a sequence s_i is similar to a subsequence s means that *dist*(s_i, s) is small. Therefore, the threshold of an error parameter ϵ is necessary for counting similar subsequences to s. A simple definition of occurring frequency of s is $|\{s_i \mid s_i \in S_w(X) \; and \; dist(s_i, s) \le \epsilon\}|$ where |A| means the number of elements of a set A. This definition needs a window length w and an error ϵ as parameters.

(2) Window length dependency

As the previous simple definition shows, a window length w should have been given before we define the distance between subsequences. When we select the better motif from motif candidates with different window lengths, we need the criteria for comparing the appropriateness of subsequences with different lengths as motifs.

(3) Trivial match

Subsequences close to a subsequence s are similar to s, if time series is continuous and has a small vibration. We call this property "trivial match". Most time series which have motif patterns satisfy this assumption. Trivial match is described formally by that dist(X[p': $p' + w - 1$] , X[p: $p + w - 1$]) is small, if $|p - p'| \ll w$. Trivial matching suggests that the simple definition may count subsequences duplicately. When we count the similar subsequences, we should take trivial matching into consideration.

3 MOTIF OCCURING FREQUENCY AND COVERAGE

3.1 Our Approaches

This subsection describes our approaches for each of problems described in the precious section.

(1) Error dependency

We will define a neighborhood of a subsequence in $S_w(X)$ for a given time series X with two parameters, a window length w and a distance error ϵ.

(2) Window length dependency

If an error is equal, the longer subsequence seems to be more appropriate than the shorter subsequence as a motif. Therefore, we introduce an error per subsequence length for a longer length motif to be advantageous. Furthermore, we will define '*motif coverage*' as an index for deciding the best motif among subsequences with different window lengths.

(3) Trivial match

When we define a neighborhood of a subsequence s in subsequence space $S_w(X)$, we remove trivial matching subsequences of s.

3.2 Neighborhood of a Subsequence

First, we introduce a disjoint neighborhood for treating a trivial match problem. Then, we define a maximal disjoint neighborhood of a subsequence for defining occurring frequency.

Definition: *Disjoint neighborhood of a subsequence*

Let X, w , ϵ and s are a time series, a window length, a positive real number and a subsequence with length w respectively. Disjoint neighborhood of s, which we denote as $DB_w(s)$, is a subset of $S_w(X)$ where for every pair $s_i, s_j \in DB_{w,\epsilon}(s)$, support $(s_i) \cap$ support $(s_j) = \emptyset$.

The above definition allows little important subsets. In an extreme case, a null set satisfies the above definition. Our concern is the one which has maximal number of disjoint subsequences.

Definition: Maximal neighborhood *of a subsequence*

Maximal neighborhood of a subsequence $B_{w,\epsilon}(s)$ is a disjoint neighborhood $DB_{w,\epsilon}(s)$ which has the largest number of elements.

There are several choices in selecting the elements of a maximal disjoint neighborhood. However, there is no problem, because it is sufficient to know only the number of elements of a set without knowing the contents of a set for defining occurring frequency.

Theorem: Construction of maximal disjoint neighborhood
Let X, w , ϵ and s are a time series, window length, a positive real number and a subsequence with length w. When we repeatedly select the nearest disjoint subsequence from s towards right to the end of X (we call the selected subsequences right neighborhood) and conversely search towards left to the beginning of X (we call the selected subsequences left neighborhood), the union of the right and left neighborhoods is a maximal neighborhood of s . We call this construction method 'nearest order selection'.
Proof:
When a $B_{w,\epsilon}(s)$ is a maximal neighborhood of s, we will show that the subsequence set constructed by the above procedure has the same number of elements. We will prove the case of right neighborhood, because the proof of the left case is similar to it.

When we sort the element of $B_{w,\epsilon}(s)$ by time ordering, we get
$B_{w,\epsilon}(s) = \{\ldots, s = X_w(p), X_w(p_1), X_w(p_2), \ldots, X_w(p_n)\}$
where $p < p_1 < p_2 < \cdots < p_n < length(X)$.
By nearest order selection, the first selected subsequence $X_w(p_1{}')$ satisfies $p_1{}' \leq p$ (inequalty 1). Because $X_w(p_1{}')$ which has the smallest $p_1{}' > p$ where
$dist(X_w(p_1{}'), X_w(\mathrm{p})) \leq \epsilon$
and support $(X_w(p_1{}')) \cap$ support $(X_w(\mathrm{p})) = \emptyset$.
Next, secondly selected subsequence $X_w(p_2{}')$ satisfies
$p_2{}' \leq p_2$ (inequalty 2) similarly.
If we continue these procedures, i-th selected subsequence $X_w(p_i{}')$ satisfies $p_i{}' \leq p_i \;\; i = 1,2, \ldots, n$ (inequalty i)
The above inequalities show that the disjoint neighborhood constructed by nearest order selection
$B'_{w,\epsilon}(s) = \{\ldots, s = X_w(p), X_w(p_1{}'), X_w(p_2{}'), \ldots, X_w(p_n{}')\}$
has equal or larger number of elements to $B_{w,\epsilon}(s)$.
However, because $B_{w,\epsilon}(s)$ is a maximal disjoint neighborhood, the number of elements of $B'_{w,\epsilon}(s)$ must be equal to that of $B_{w,\epsilon}(s)$. This means that $B'_{w,\epsilon}(s)$ is also a maximal neighborhood of a subsequence s.

3.3 Occurring Frequency and Motif Coverage

This subsection defines occurring frequency and motif coverage.

Definition: Occurring frequency
Let X, w , ϵ and s are a time series, a window length, a positive real number and a subsequence with length w respectively. Occurring frequency of a subsequence s on an error ϵ is the number of the elements of a maximal neighborhood of a subsequence $B_{w,\epsilon}(s)$, *that is,* $|B_{w,\epsilon}(s)|$.

Definition: Motif Coverage
Let X, and ϵ are a time series and a unit error per window length respectively. Motif coverage of a subsequence s with length w is $w \times |B_{w,w\epsilon}(s)|$

We can select the best motif among subsequence with different window lengths by the following procedure.
1. Give a list of window lengths W.
2. Select the subsequence which has the maximal occurring frequency for each window length in W. The selected subsequences are each best motifs for each window lengths in W.
3. Select the subsequence that has the maximal motif coverage among the motifs obtained by procedure 2. The selected subsequence is the best motif among all the subsequences with all the window lengths in W.

4 ALGORITHM

We can get algorithms for calculating occurring frequency and motif coverage by operationally interpreting the definitions and the theorem in the previous section.

TABLE 1 shows an algorithm that counts the occurring frequency of a given subsequence. The inputs are a time series *X*, a window length *w* of the given subsequence *s*, a starting time *t* of *s* and an error per window length ϵ. The outputs are the occurring frequency and the motif coverage of the given subsequence *s*.

Line 01 calculates each distances between the given subsequence *s* and each subsequences in $S_w(X)$. Line 02 counts the number of elements of a maximal neighborhood subsequence set whose elements are in the right side of the given subsequence *s*. Line 03 counts the number of the elements of a maximal neighborhood subsequence set whose elements are in the left side of *s*. Line 04 counts the total occurrence frequency of *s* by adding the right side one obtained by line 02 and the left side one obtained by line 03. Line 05 calculates the motif coverage of *s* by multiplying the window length *w* and the occurring frequency obtained by line 04.

TABLE 1. CountOccurringFrequency Algorithm

```
Algorithm: CountOccurringFrequency (X, w, t, ϵ)
[Input]  X: Given time series
         w: Length of a given subsequence s
         t:  Stating time of a given subsequence s
         ϵ: Error per unit length
[Output]  OF: Occurring frequency of s in error wϵ
          MC: Motif coverage of s
01  DL = distanceListFromS(X, t, w);
02  OFR = countRightOccurence (DL, t, w, ϵ)
03  OFL = countLeftOccurence (DL, t, w, ϵ)
04  OF = OFR + OFL;
05  MC = OF * w;
06  return (OF, MC);
```

TABLE 2 shows an algorithm that counts the number of elements of a maximal neighborhood subsequence set whose elements are in the right of the given subsequence s. The inputs are the distance list DL obtained by line 01 in TABLE 1, the window length *w* of a given subsequence *s*, a starting time *t* of *s* and the error per window length ϵ. The output is the number of maximal neighborhood subsequences in the right side of *s*.

Line 01initializes a time cursor 'Cur' and a normalized error 'Err'. Line 02-13 is a while-loop that chooses maximal subsequences that are in the right side of the given subsequence *s* to the end of the time series *X*. Line 03-05 is a while-loop that searches the next disjoint subsequence whose distance from *s* is smaller than 'Err'. Line 06-09 increments 'Right' when the line 03-05 found a new disjoint subsequence. Line 10-12 exits while-loop 02-13 after checking all the subsequences in the right side of *s*.

TABLE 2. CountRightOccurence

Algorithm: *countRightOccurence* (DL, *t, w,* ϵ)

[Input] DL: Given time series
w: Window length of a given subsequence *s*
t: Stating time of *s*
ϵ: Error per unit length
[Output] Right: the number of maximal neighborhood subsequences in the right of *s*.

```
01  Cur = t+1;  Err = $\epsilon * W$;
02  while Cur <= length(DL)
03    while DL(Cur) >  Err or Cur <= length(DL)
04      Cur = Cur + 1;
05    end
06    if DL(Cur)  <=  Err
07      Right := Right + 1;
08      Cur := Cur + w – 1;
09    end
10    if Cur > length(X)
11      break;
12    end
13  end
14  return Right;
```

TABLE 3 shows an algorithm that counts the number of maximal neighborhood subsequences which are in the left side of the given subsequence s. The left case can be reduced the right case by reversing the time series values from right to left.

Line 01 reverses the distance list 'DL' from right to left. Line 02 reverses the starting time *t* of *s* from right to left. Line 03 gets the value of the left case by calling the algorithm 'CountRightOccurence' with the reversed arguments.

TABLE 3. CountLeftOccurence Algorithm

Algorithm: *countLeftOccurence* (*X, w, t,* ϵ)

[Input] DL: Given time series
w: Window length of a given subsequence *s*
t: Stating time of a given subsequence *s*
ϵ: Error per unit length
X: Given time series
[Output] Left: the number of maximal neighborhood subsequences in the left of *s*.

```
01  DL_rev = fliplr(DL);
02  t_rev  = length(X) – t + 1;
03  Left = countRightOccurence (DL_rev, t_rev, w, $\epsilon$)
```

5 EXPERIMENTAL EVALUATION

We show that our proposed index 'motif coverage' can select the best motif among subsequences with different window lengths.

5.1 Window Length Selection

This subsection shows that for two simulated time series in which motif patterns are intentionally inserted, motif coverage can select the best window length of the inserted motif.

(1) Experiment 1

First, we will show that 'motif coverage' can decide the best window length 15 by selecting window length that has the highest 'motif coverage' among each best motifs with window lengths 5,9,15 and 31 in time series shown by Figure 1.

Figure 1 is a simulated time series that combines sine curves with length 15 and random subsequences with various length.

Figure 2 shows motif coverage values on each times and each window lengths in case that an error per window length is 0.01. The top graph is a motif coverage trend graph in window length 5. The second, third and fourth trend graphs from the top to the bottom are each motif coverage ones with window length 5, 9, 15 and 31 respectively. The third trend graph for length 15 shows that times at which motif patterns start have high and sharp peaks. The trend graphs for length 5 and 9 shows that times at which sub-patterns of the best motifs start with length 15 have relatively high motif coverage values and longer peak duration than those of length 15. These observations match the fact that the best motif pattern includes motifs with smaller window lengths.

Figure 3 shows each best motifs in each window lengths. The best motif in length 15 is an intentionally inserted one. The best motifs for window lengths 5 and 9 are the sub-patterns of the best motif with length 15. The best motif with length 31 is a subsequence including the best motif with length 15.

Figure 4 shows motif coverage values for each best motifs with each window lengths. The window length which has the highest motif coverage is 15. It shows that motif coverage can decide the optimal window length. It also shows that the nearer window length to 15 is, the higher motif coverage is. This result is what we expect for motif coverage.

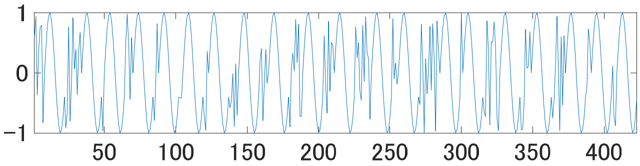


Figure 1. A time series with a motif of length 15 samples

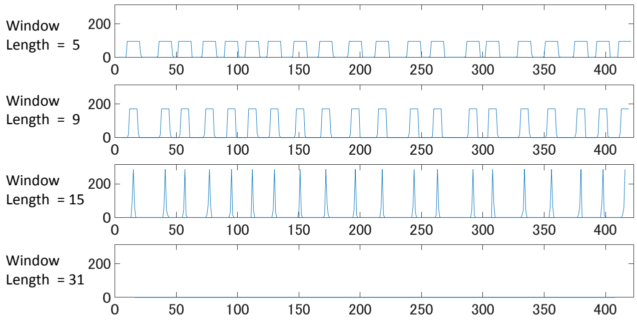


Figure 2. Motif coverage of each times for each window lengths

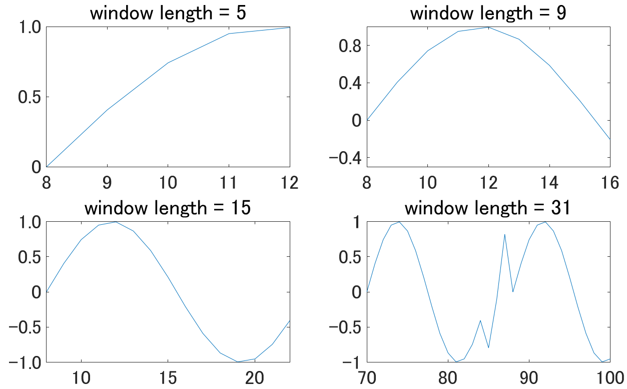


Figure 3. Best motifs for each window lengths

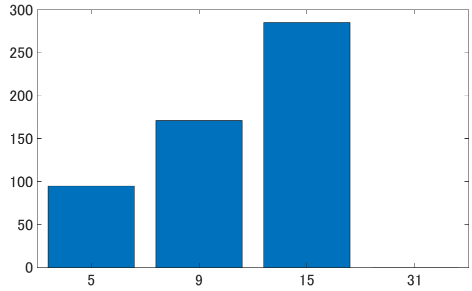


Figure 4. Motif coverage of best motifs with each window lengths

(2) Experiment 2

Next, we will show that ‘motif coverage’ can decide the best window length in a time series including two motif patterns with different lengths.

Figure 5 is a simulated time series that combines sine curves with length 15 and 31 and random subsequences with various length.

Figure 6 shows motif coverage trend graphs in each window lengths in case that an error per window length is 0.01. Each trend graphs are for window lengths 5, 9, 15, 31 and 47 from the top to the bottom. As is the first experiment, the trend graphs for 15 and 31 have high peaks of motif coverage values at times when motif patters start. The trend graphs for 5, 9 and 15 have relatively high motif coverage at times when sub-patterns of the motif patterns with length 15 or 31 start.

Figure 7 shows each best motifs in each window lengths. The best motifs in window lengths 15 and 31 are intentionally inserted ones. The best motifs in each window lengths 5 and 9 are sub-patterns of the best motifs with lengths 15 or 31. The best motif with window length 47 is a subsequence including the best motif in window length 31.

Figure 8 shows motif coverage for each best motifs with each window lengths. It shows that 15 and 31 are top 2 window lengths.

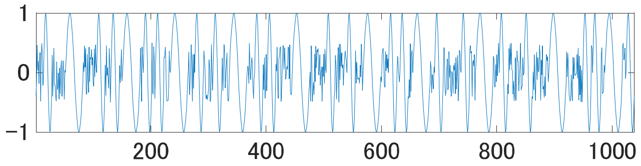


Figure 5. A time series with length 15 and 31 motifs.

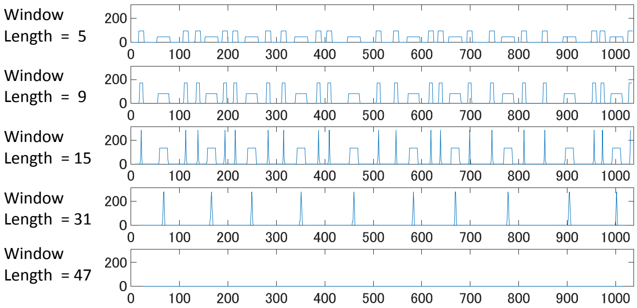


Figure 6. Motif coverage of each times for each window lengths

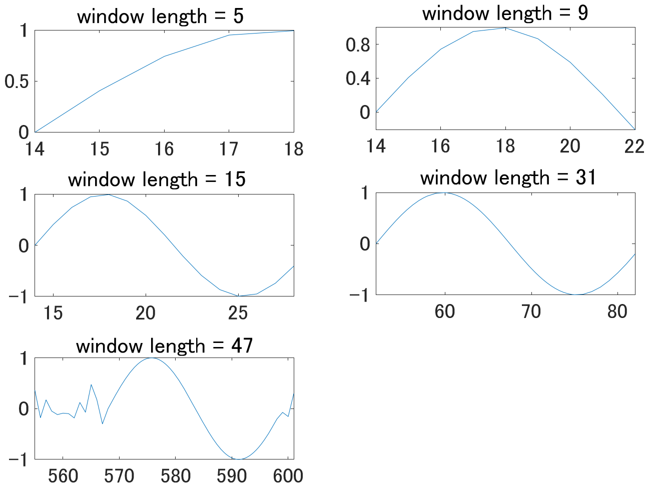


Figure 7. Best motifs for each window lengths

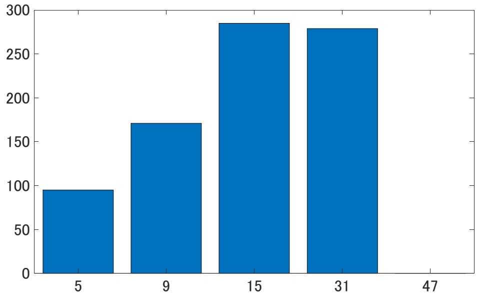


Figure 8. Motif coverage of best motifs with each window lengths

5.2 Error dependency of Motif Coverage

The previous subsection shows that 'motif coverage' with an appropriately selected error per window length can decide the best window length. However, the problem how to decide an error per window length still remains. This subsection shows error dependency of motif coverage and the consideration on this remaining problem.

Figure 9 shows error dependency graphs of each motif coverages of the best motifs with each window lengths for the time series shown in experiment 1. This graph shows that the best window length has overwhelmingly highest value in small error values. However, as the larger the error value is, the smaller the difference is. This observation suggests that an occurring frequency should be measured in a small error value, because motif is local characteristics in a subsequence space $S_w(X)$. This observation also suggests that a rising point in an error dependency graph seems to be a candidate for deciding the optimal error per window length.

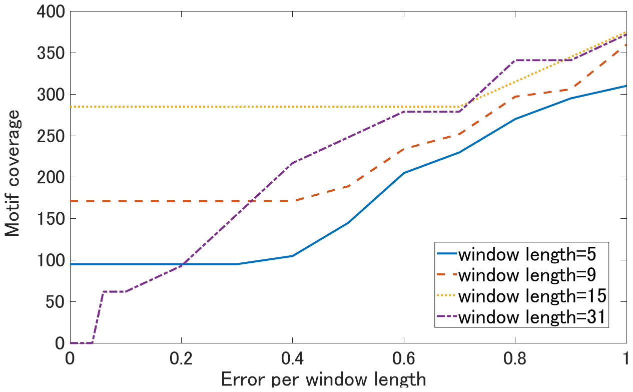


Figure 9. Error dependency of each motif coverage of each best motifs with each window lengths (Data 1)

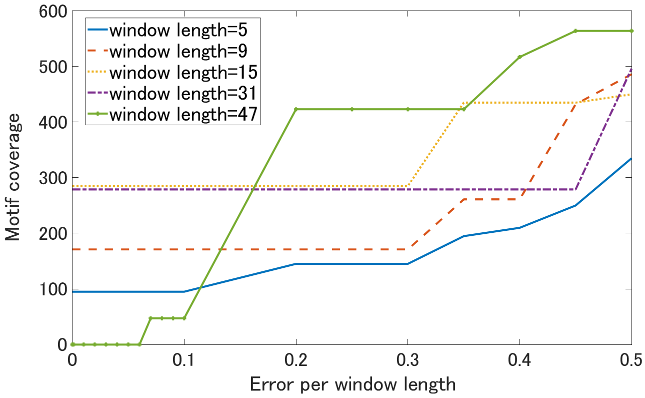


Figure 10. Error dependency of each motif coverage of each best motifs with different window lengths (Data 2)

Figure 10 shows error dependency for time series shown in experiment 2. This graph shows that the best window length 15 and 31 have overwhelmingly higher values in small error values. However, as larger the error value is, the smaller the difference is. This observation is the same as that in data shown in experiment 1.

6 CONCLUSIONS

We proposed novel indexes 'occurring frequency' and 'motif coverage' for deciding optimal window length in motif discovery. We also showed that 'motif coverage' can decide the best window length by selecting window length that has the highest 'motif coverage' among each best motifs with each window lengths in simulation data.

From a theoretical point of view, a future work is how to decide an error per window length, which is a remained parameter of motif coverage. From an experimental point of view, we plan to apply our algorithms to more complex simulated data as well as real data.

This work is supported by JSPS KAKENHI Grant Number 17K00161.

REFERENCES

[1] Pranav Patel, Eamonn J. Keogh, Jessica Lin, Stefano Lonardi: Mining Motifs in Massive Time Series Databases. IEEE ICDM pp. 370-377 (2002).

[2] Abdullah Mueen, Eamonn J. Keogh, Qiang Zhu, Sydney Cash, M. Brandon Westover: Exact Discovery of Time Series Motifs. SDM pp. 473-484 (2009).

[3] Takuya Maekawa, Daisuke Nakai, Kazuya Ohara, Yasuo Namioka: Toward practical factory activity recognition: unsupervised understanding of repetitive assembly work in a factory. UbiComp pp. 1088-1099 (2016).

[4] Zeeshan Syed, Collin M. Stultz, Manolis Kellis, Piotr Indyk, John V. Guttag: Motif discovery in physiological datasets: A methodology for inferring predictive elements. TKDD 4(1): 2:1-2:23 (2010).

[5] Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, Eamonn J. Keogh: Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins. IEEE ICDM, pp. 739-748 (2016).

[6] William R. Lovallo, Michael F. Wilson, Andrea S. Vincent, Bong Hee Sung, Barbara S. McKey, Thomas L. Whitsett: Blood Pressure Response to Caffeine Shows Incomplete Tolerance After Short-Term Regular Consumption, Hypertension.43.4 p.760-765 (2004).

[7] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, Eamonn J. Keogh: Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. IEEE ICDM, pp. 1317-1322 (2016).

[8] Yan Zhu, Makoto Imamura, Daniel Nikovski, Eamonn New Primitive for Time Series Data Mining (Best Student Paper Award). IEEE ICDM, pp. 695-704 (2017).

[9] Eamonn J. Keogh, Jessica Lin, Wagner Truppel: Clustering of Time Series Subsequences is

Meaningless: Implications for Previous and Future Research. IEEE ICDM, pp. 115-122 (2003).

[10] Tsuyoshi Idé: Why Does Subsequence Time-Series Clustering Produce Sine Waves? PKDD pp. 211-222 (2006).

Visualization of Skills and Techniques Using Big Data Analysis for Vocational Skill Development

Masaki Endo[*], Takuo Kikuchi[*], Shigeyoshi Ohno [*], Makoto Imamura[**], and Hiroshi Ishikawa[***]

[*] Division of Core Manufacturing, Polytechnic University, Japan
{endou, kikuchi, ohno}@uitec.ac.jp
[**] School of Information and Telecommunication Engineering, Tokai University, Japan
imamura@tsc.u-tokai.ac.jp
[***] Graduate School of System Design, Tokyo Metropolitan University, Japan
ishikawa-hiroshi@tmu.ac.jp

Abstract— At manufacturing sites throughout Japan, labor force reduction and industrial structure changes are progressing. To meet these and related challenges, many small and medium-sized enterprises need not only to advance skills and techniques, but also to hand down skills and techniques to new groups of workers. Therefore, the need for human resource development in small and medium-sized enterprises is increasing. Although many such enterprises acknowledge the importance of vocational training, a shortage exists of human resources responsible for guiding vocational ability development. Visualization of skills using ICT and advancement of skills are urgently needed at manufacturing sites in Japan. This study examines the visualization of skills using big data analysis at sites of vocational ability development. As described herein, we explain construction of the analytical environment using the operation history in programming practice as the first stage.

Keywords: big data; big data utilization technology; curriculum model; job training; polytechnic science

1 INTRODUCTION

With the dramatic development of information and communication technologies such as the spread of smartphones and Internet of Things (IoT), enormous amounts of information generated in society and many academic research fields are accumulated as big data. Such big data include vast knowledge and potential value. Therefore, their effective use for analysis and processing of data using artificial intelligence (AI) is crucially important for future industrial development. Actually, big data utilization is anticipated not only to strengthen the international competitiveness of companies, but also to solve social problems, to create new businesses and services, to improve the convenience of individual lives, etc. Already, intense international competition has begun. In the industrial world, Germany has promoted Industry 4.0; the US advocates an Industrial Internet. Moreover, a struggle for international standardization has developed. Comprehensive efforts including human resource development to cope with this trend have become urgent in Japan as AI, IoT, big data, and structural reforms for technological innovation of robots.

Social media data, website data, sensor data, log data, multimedia data, customer data, office data, operation data are all examples of big data recognized by the Ministry of Internal Affairs and Communications (METI) [1]. Big data is defined as "a large amount of data that is difficult to manage with existing technology" and which is represented by Volume, Variety, and Velocity: the 3 Vs (generation speed / update frequency). Large varieties of data having various forms and structures can be generated, collected, accumulated, and otherwise processed using information and communications technology (ICT). Furthermore, big data are fundamentally heterogeneous large-scale datasets with non-stationarity that captures data with different frequencies and accuracies from moment to moment.

Fundamentally, conventional data analysis processes input data collected for some purpose with an algorithm, yielding necessary information as an output. However, analysis of big data makes it possible to realize scientific discovery, prediction, and knowledge acquisition that could not be achieved through conventional data analysis. However, representing natural phenomena and social life requires the use of analyzed results of data of various kinds gathered using sensor arrays, monitors, and recorded information transmitted on a daily basis. Furthermore, for information processing, society is shifting from interpretation of data and from appropriate modeling of analysis methods to information and knowledge acquisition.

For the reasons enumerated above, the E-Science Data Center Science Subcommittee of the Informatics Committee of the Science Council of Japan [2] and the Ministry of Education, Culture, Sports, Science and Technology (MEXT) [3] recommend the necessity of developing big data utilization technicians who have acquired new information processing technologies such as data acquisition technology dealing with big data, data utilization technology, machine learning, and statistical modeling. Furthermore, they are tackling human resource development. In this way, top-down efforts centering on MEXT, METI, and IT industries are being developed, such as academic research responding to technological innovation and further improvement of technical skills of information processing engineers.

Nevertheless, bottom-up training for practical engineers to use big data mainly in manufacturing industries in Japan has not been implemented as public vocational training. Figure 1 portrays the situation in Japan and an outline of this research, which was conducted to construct a curriculum model for learning Big Data utilization technologies with the current training contents of public vocational training. At universities and large enterprises designated by MEXT, research and efforts are being conducted on advanced data science. Efforts related to big data utilization are limited at production sites of small and medium enterprises. Among them, the introduction of data science to production sites of small and medium enterprises represents an urgent issue for Japan's manufacturing industry, where the labor force is decreasing and the aging of skilled technicians is progressing. Therefore, this study was undertaken to construct a data collection infrastructure to promote the utilization of big data for engineers at the production site. Based on data gathered in this research, we assess a curriculum through which on-site technicians acquire basic knowledge of utilization technology using data science and by which they can collaborate with advanced data scientists.

Therefore, we consider data necessary for analyses from a person, machine, material (3M) perspective of the model used for problem identification and analysis at the production site and aimed at modeling. As described in this paper, we explain programming practice as the first approach. As described herein, we present an approach to standardize the data model for analysis by accumulating data of a trainee's programming work, targeting programming practice. Chapter 2 describes earlier research related to this topic. Chapter 3 explains our proposed method for estimating proficiency in education and for training using data from a trainee's programming work. Chapter 4 describes experimentally obtained results for our proposed method and a discussion of the results. Chapter 5 presents a summary of contributions and future work.

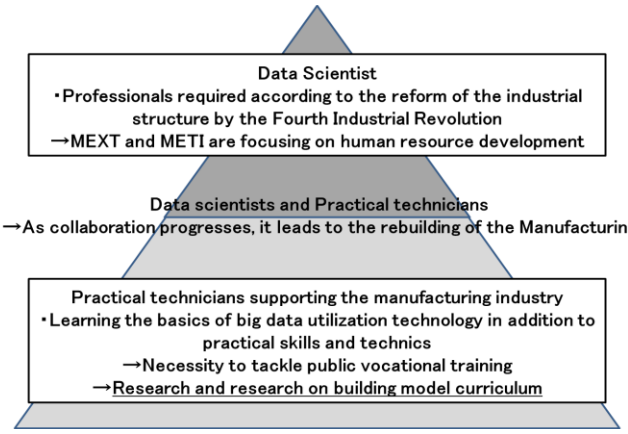

Data Scientist
・Professionals required according to the reform of the industrial structure by the Fourth Industrial Revolution
→MEXT and METI are focusing on human resource development
Data scientists and Practical technicians
→As collaboration progresses, it leads to the rebuilding of the Manufacturing
Practical technicians supporting the manufacturing industry
・Learning the basics of big data utilization technology in addition to practical skills and technics
→Necessity to tackle public vocational training
→Research and research on building model curriculum

Figure 1: Present situation in Japan and outline of this research.

2 RELATED WORK

Research investigating proficiency related to skills and techniques is conducted in various fields. Furukawa et al. [4] specifically examine such sophisticated physical skills as playing sports and playing instruments and introduces research activities aimed at elucidating and verbalizing those skills. They introduce approaches from physical modeling, measurements and data analysis, cognitive science, and human interfaces. Yamagiwa et al. [5] simplify the recording of human activities as body movement big data and present approaches to visualize differences between expert skills and novice skills using a technique called skill grouping. They identify skill differences in sports activities by comparison with skilled athletes. Miyadera et al. [6] present a system in which teachers monitor real-time learning of beginner programmers in the classroom and provide support to groups of students with common problems. Their system has a program animation function that passes through the program with understanding of the student's historical record of operations. By analyzing this record, the system accurately notifies the teacher of common problem areas and an inferred cause of the difficulty. Tanigawa et al. [7] reports that a teacher can not teach efficiently when a student's source code is read at the time of questioning in the programming exercise. Furthermore, the trial and error process from the record of the function calling order at the time of source code creation by the student is analyzed. A calculation method is proposed for student acquisition items. In their method, by deriving a pattern common to many students as a calling pattern, a user can ascertain learning items efficiently and can then pursue effective teaching.

Techniques related to skills and techniques have been proposed in various fields, but examination of data collection and analytical methods for the manufacturing field have not been systematized. In many cases at many large enterprises, a PDCA cycle such as kaizen is incorporated into a business, leading to productivity improvement. However, in small and medium-sized enterprises supported by public vocational training, implementing such a cycle effectively and efficiently is difficult because of labor shortages and individualization. Therefore, we adopt a scientific approach for visualization and sophistication against skills required of manufacturing experts possessed by skilled technicians such as craft skills and Olympic champion skills. In addition to engineering, we are considering the application of methodologies to fields such as social system science, pedagogy, and human information science [8].

3 OUR PROPOSED METHOD

This section presents a description of the analytical method used for target data collection. As the first step of systematizing data collection that is applicable to analysis of the skill, which is the ultimate objective, we present a data collection method modeling the operation history at the time of programming practice. Our proposal is presented in Fig. 2. This paper introduces data collection in programming practice, but the ultimate goal of this research is to collect large-scale data targeting the Polytechnic Centers of 64 public training centers and Polytechnic Colleges of 25 schools nationwide. Furthermore, the data collection target area is expected to include mechanical systems, electrical systems, construction systems, and others.

We describe an acquisition level estimation method using analysis with the operation history at programming practice. Section 3.1 describes gathering of the operation history at programming practice. Section 3.2 describes preprocessing for conducting analyses. Section 3.3 describes the method of estimating the learning degree. With an analytical method using the operation history during programming practice, one can estimate the learning degree to programming practice. Section 3.4 presents output of the estimation result.

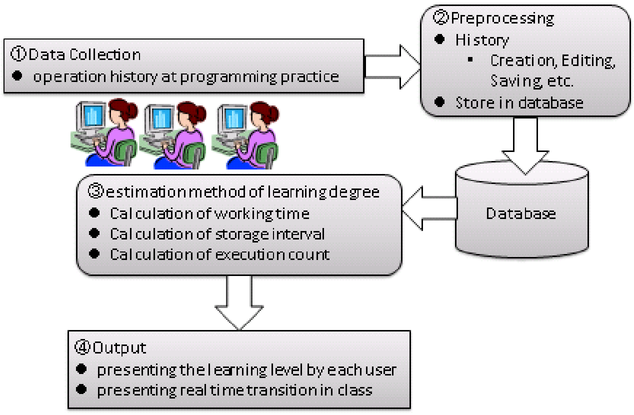

①Data Collection
● operation history at programming practice
②Preprocessing
● History
▪ Creation, Editing, Saving, etc.
● Store in database
Database
③estimation method of learning degree
● Calculation of working time
● Calculation of storage interval
● Calculation of execution count
④Output
● presenting the learning level by each user
● presenting real time transition in class

Figure 2: Summary of our proposal.

3.1 Data Collection

This section presents a description of the Method of (1) data collection presented in Fig. 2. As described in this paper, the file operation history at the time of programming practice is the collection target. Items to be monitored are file creation, editing, and saving. This data collection was done for Arduino programming practical training using folder monitoring free software [9] in a Windows environment and the inotifywait [10] command in Linux.

Next, we describe the number of collected data. As the initial stage of standardization this time, we obtained data of several cooperators. To classify the programming experience, data were classified as those of programmers with no experience, with less than five years experience, or with over five years of experience. We analyzed the acquisition level by applying processes using these data.

3.2 Preprocessing

Preprocessing stores the creation, editing, and saving of each user's code file and the creation, editing, and saving history of executable files generated by compilation in the database. The file name is targeted for programming practice. Therefore, it is specified beforehand. The code file and the executable file are associated. In the database, the operation history of the user's coding file and executable file are stored together with time information.

3.3 Estimating acquisition level

Analysis of the degree of acquisition uses data accumulated through preprocessing. The time from creation of the coding file by each user to the final time at which the execution file was created by compilation is set as the working time. Analysis of the acquisition level was done during that time according to the relation between the storage time of the coding file and the update time of the execution file. Figure 3 presents an overview of acquisition level estimation.

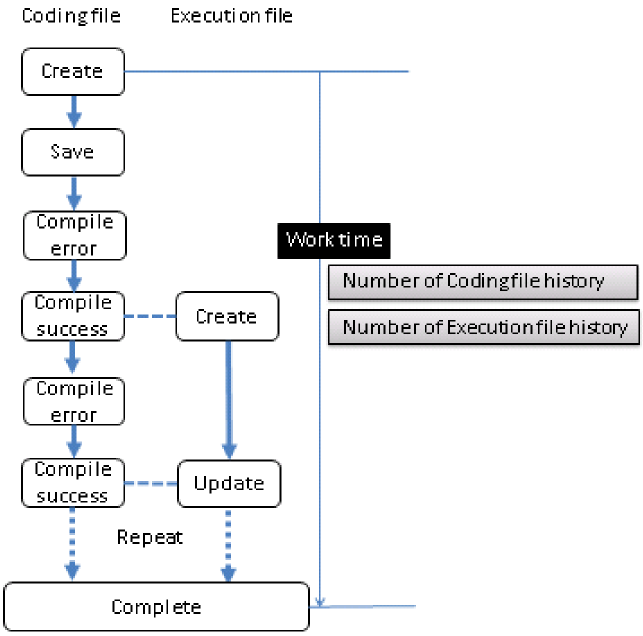


Figure 3: Estimating acquisition level.

Here, the number of times the coding file was saved might be saved for backup, but it is assumed that many saves are done during compilation. Writing to the microcomputer is performed when the executable file is created or updated. Because Create is the first write to the microcomputer, a file update occurs. Therefore, judging the possibility of overlapping compilation errors means that one file is saved many times. When the corresponding execution file is not updated at the time of saving of the coding file, it is judged that a compilation error has occurred. Furthermore, if the coding file is saved after the execution file is updated, it is judged that a runtime error has occurred and that an operation to specify has not been completed.

Therefore, using the proposed method, the shorter the work time, the higher the degree of acquisition. Additionally, we conducted an analysis assuming that a higher skill level is associated with fewer saves of the coding file and fewer updates of the executable file within the working time. The proposed method is a very general analytical method. As programming analysis methods, advanced and highly accurate analysis methods such as empirical software engineering have already been examined in earlier reports. Nevertheless, no report of the relevant literature in the field of public vocational training and analyzing beginner learning technology describes a study conducted to establish an effective public vocational training system by constructing a large-scale data collection infrastructure. Therefore, this study was conducted to construct a data collection base conforming to the actual circumstances of public vocational training sites.

3.4 Output

This section presents a description of (4) the output method portrayed in Fig. 2. Output considers visualization using the acquisition degree estimated by processing, as explained in the preceding section. We assess the use as an auxiliary system for programming learning by showing each user's learning level and the real time transition in class.

4 EXPERIMENTS

This chapter explains experiments conducted to estimate the acquisition level using the method explained in Chapter 3. Section 4.1 describes the dataset. Section 4.2 presents examples of the problem of programming practice for collecting data. Section 4.3 presents the trend of the degree of acquisition obtained using the collected data.

4.1 Dataset

Datasets used for this experiment were collected using data presented in Table 1. As described in this paper, the results for two people are shown by experimentation. Therefore, the data are few; the experiment results are unreliable. Currently, we are selecting procedures for a similar curriculum and obtaining permission for data collection; we are also collecting data individually by soliciting cooperation with the experiment. This time we gathered data from experimental collaborators after explaining the gist of data collection in advance and after obtaining consent. Programming experience, embedded programming experience, and embedded programming experience using Arduino are presented in the table. We use two people's data: both A and B have programming experience, but A has embedded programming experience; B does not. Neither A nor B has Arduino programming experience. The number of hours A is 72 hr for C language and 72 hr for embedded programming. B is 36 hr in C language. The programming time using research and other languages is not included. Furthermore, in general public vocational training, the programming practice time is often implemented in units of 36 hr. Therefore, we believe it will be a reference for actual training measurements.

Table 1: Datasets

Experimental collaborators	Programming experience	Embedded programming experience	Embedded programming experience using Arduino
A	3 years	3 years	No
B	4 years	No	No

We gathered data by having each experimental collaborator perform the same task. Because an Arduino is used for the first time in some cases, we use an explanation for beginners to instruct everyone in the prior explanation and programming method. Therefore, the data collection flow occurs according to the following procedure.

Procedure
1. Arduino description
2. Explanation of the programming method using a sample program
3. Explanation of compilation and operation confirmation method
4. Explanation of elements necessary for the task
5. Execute task
6. Repeat 4 and 5 for each task

4.2 Examples of the problem of programming practice

This section presents practice tasks prepared for data collection. Figure 4 and Fig. 5 respectively present examples of tasks involving tact switches and LEDs. Simple examples illustrate the foundation of embedded programming, but experienced persons and inexperienced people exhibit differences in their times to perform tasks. For the experiment, we prepared five tasks and gathered the operation history of the experiment collaborators. Tasks 1–5 confirmed the following points.

Task 1: Using timer and external interrupt
Task 2: Port operation
Task 3: Basics of Interrupt Processing
Task 4: Application of interrupt Processing
Task 5: Application of Task 1-4

Moreover, in this experiment, hardware parts such as LEDs and switches were prepared in advance. Data collection and analysis including tasks other than programming, such as hardware assembly, are the next step of this research.

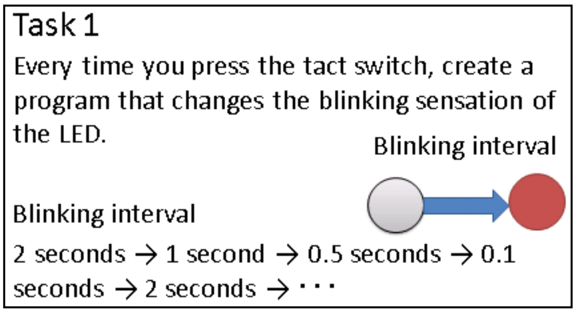

Task 1
Every time you press the tact switch, create a program that changes the blinking sensation of the LED.
Blinking interval
2 seconds → 1 second → 0.5 seconds → 0.1 seconds → 2 seconds → · · ·

Figure 4: Example of task 1.

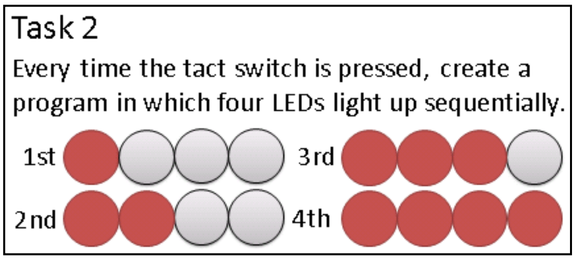

Task 2
Every time the tact switch is pressed, create a program in which four LEDs light up sequentially.

Figure 5: Example of task 2.

4.3 Experimental result

Table 2 and Fig. 6 present results from Task 1 of the student serving as the instructor of this experiment. The student had learned about Arduino programming beforehand and had instructed the experiment examinees. The working time was 4 min. Additionally, the coding file was generated at 14:20 and was updated twice. Regarding the 14:22 update, it was impossible to generate an executable file because of a compilation error, but compilation was successful at the 14:23 update. Therefore, one can confirm that the executable file was created at the same time. In addition, the execution file has two events: 14:23 compiling success and 14:24 updating. The second event at 14:24 is an event by writing to the microcomputer. Therefore, A finished task 1 in 4 min by two compiling works and one writing work.

Table 3 presents results of tasks from each of the two collaborators shown in Table 1. A has experience with embedded programming. Therefore, the work time of Task 1 is shorter than that of B. Adaptation to Arduino programming is apparent. In addition, Task 3, an intrinsic programming specific task that uses interrupts and delay functions, is applicable in a shorter time than B. Therefore, embedded programming experience is apparently superior for dealing with this Arduino programming. By contrast, B tends to have fewer compile errors than A, and tends to compile less frequently until completion of a Task. Although B had no experience with embedded programming, B’s programming proficiency is thought to have been higher than that of A. In Task 4 and Task 5, which are application tasks, the problem is a summary of the fundamental programming of Arduino, which was undertaken this time. Both A and B show almost no difference between these two problems. Therefore, I think that basic skills can be learned through the experiment Task. Achievement obtained using the curriculum was obtained.

Although this is still in a hypothetical stage, we believe that grammatical errors and algorithmic errors can be reduced by improvement of programming skills. Therefore, at the time of programming, the skill is analyzed based on the number of times used for compiling.

Table 2: Data of task 1 of instructor

Times of Day	History	Filename
14:20	Create	kadai_01.ino
14:22	Update	kadai_01.ino
14:23	Update	kadai_01.ino
14:23	Create	kadai_01.ino.hex
14:24	Update	kadai_01.ino.hex

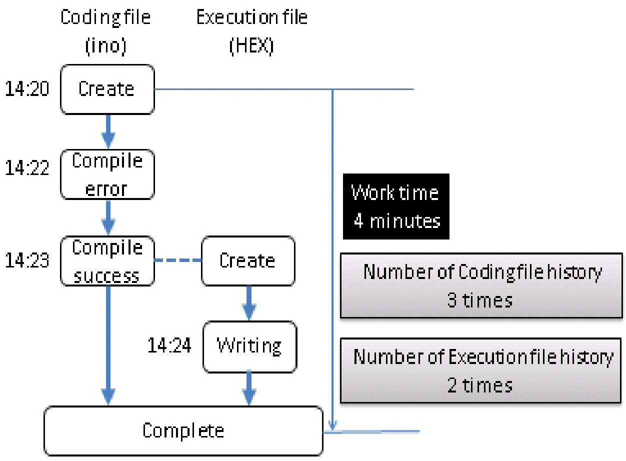


Figure 6: Data of task of instructor.

Table 3: Data of task 1 of experiment collaborator A

Experimental collaborators	Task	Work time [min]	Number of Coding file history [time]	Number of Execution file history [time]	Work time average [min]	Number of Coding file history average [time]	Number of Execution file history average [time]
A	Task 1	11	7	7	4.6	3.6	3
	Task 2	4	2	2			
	Task 3	5	5	4			
	Task 4	1	2	1			
	Task 5	2	2	1			
B	Task 1	14	4	4	6.4	2.4	2.4
	Task 2	4	2	2			
	Task 3	11	4	4			
	Task 4	2	1	1			
	Task 5	1	1	1			

Trends of embedded programming experience and programming proficiency can be confirmed using this method. Therefore, if one were able to acquire a large amount of data using this system, then analytical objectives such as the proficiency level of detailed programming and difficulty judgment of tasks could be considered using data analysis. Furthermore, collaborators implemented the task only once for this experiment, but it is expected that repeating the programming practice will shorten the work time and reduce the compiling errors for the task. Future studies must include experiments designed to collect and analyze numerous histories of programming practices of single subjects over a long time.

5 CONCLUSION

As described in this paper, for vocational ability development, we have constructed a system to collect operation histories during programming practice as the first step to visualizing skills and techniques with big data analysis. The research, which started this fiscal year, is at the stage of consultation with the ethics committee on data collection related to personal information. Therefore, we have not reached large-scale data collection. However, experimentally obtained results of collaborators during these experiments confirmed the differences between experienced programming and mastery level programming. Results demonstrate the possibility that accumulating large amounts of data using a data collection system in the future can support analysis of practical programming skills and techniques. Additionally, although targeting data in programming practice now, the author would like to conduct research to clarify differences between inexperienced and skilled workers by sequentially expanding and collecting data including human movement and biometric information.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 18K13254.

REFERENCES

[1] Information and Communications Council, Regarding how to utilize big data, Ministry of Internal Affairs and Communications (2012).

[2] Japan Science Council Information Informatics Committee, Recommendation Human Resources

Supporting the Big Data Era, National Research and Development Corporation Japan Science and Technology Agency, (2014).

[3] Information and Systems Research Organization, About developing specialized human resources to utilize big data, Inter-University Research Institute Corporation Research Organization of Information and Systems (2015).

[4] K. Furukawa, K. Ueno, T. Ozaki, S. Kamisato, R. Kawamoto, K. Shibuya, N. Shiratori, M. Suwa, M. Soga, H. Taki, T. Fujinami, S. Hori, Y. Motomura, and S. Morita, Research Trend of Physical Skill Science – Towards Elucidation of Physical Skill –, Transactions of the Japanese Society for Artificial Intelligence, Vol.20, Issue2, pp.84-93, Japanese Society for Artificial Intelligence (2005) (in Japanese).

[5] S. Yamagiwa, Y. Kawahara, N. Tabuchi, Y. Watanabe, and T. Naruo, Skill Grouping Method: Mining and Clustering Skill Differences from Body Movement BigData, Proceeding of International Conference on BigData 2015, IEEE (2015).

[6] Y. Miyadera, K. Kurasawa, S. Nakamura, N. Yonezawa, and S. Yokoyama, A Real-time Monitoring System for Programming Education using a Generator of Program Animation System, Journal of Computers, Vol.2, No.3, pp.12-20, Academy Publisher (2007).

[7] K. Tanigawa, D. D. Phuong, F. Harada, and H. Shimakawa, Grasping Learned Items in C Programming Exercise Using Function Call Logs, IEICE Transactions on Information and Systems (Japanese edition), Vol.95, No.12, pp.2079-2089, IEICE (2012).

[8] Study Group on Polytechnic Science, Introduction to Polytechnic Science, JUSE Press Ltd. (2018) (in Japanese).

[9] Folder monitoring, tukaeru, URL <http://www.saberlion.com/tukaeru/soft/folders.html> (2016).

[10] Ubuntu manuals, Ubuntu Manpage Repository, URL <http://manpages.ubuntu.com/manpages/bionic/man1/inotifywait.1.html > (2018).

Session 6:
Systems and Applications

(Chair: Ohki Tetsushi)

Interactive topic confirmation on decision support system for information dissemination

Kei Utsugi[*], Hajime Mori[*], Kiyohiro Ohara[*], Masahide Okamoto[*] and Yuuki Shimizu[*]
[*]Hitachi Ltd., Japan
{kei.utsugi.nz, hajime.mori.vh, kiyohiro.obara.pc, masahide.okamoto.ex, yuuki.shimizu.kf}@hitachi.com

Abstract - A decision-support system for quality-assurance activities using existing business documents was implemented as a prototype, and its novel features were experimentally evaluated. The prototype system helps the user decide to which division within an organization information of the current problem should be disseminated by using the similarity between the problem document and existing documents held by each department. The novel point concerning the prototype system is its user-interaction methodology used in recommendation process. The system presents feature terms for explaining classifications given by local interpretable model-agnostic explanations method and finds other related documents from feature terms in a smooth interaction. The user-interaction ensures the user's perspective and attention from behavior of checking system output, the perspective and behavior of the user are included in the output of the prototype system and a user's action logs can be obtained and used to analyze how expert users read and check the target problem.

Keywords: Computer-supported collaborative work, Human interaction.

1 INTRODUCTION

Quality management of products and service is a high priority for every responsible company. Such companies are always devoted to learning from information about cases of failure through the whole product life cycle—from design and development to production, operation. They believe that good decisions always rely on facts and experience, and analysis of failure cases is the best teacher for those who build new products and processes. However, as the amount of information on case studies increases, it becomes increasingly difficult for a single person to grasp a whole picture of that information. To make full use of such information, the help of a technology called "natural language processing" (NLP) is required. In the present study, to meet that requirement, a prototype of a decision-support system for *quality assurance* (QA) activities utilizing existing business documents was proposed and evaluated.

Current QA activities have been basically based on empirical knowledge by experts who are proficient in handling information within an organization. However, given that changes in organization and replacement of talented people are inevitable, the methods that rely upon such experience and empirical knowledge involve the risk of overlooking the target department in which the alert information should be deployed.

The prototype system assists in finding divisions that should disseminate the target issue on the basis of the similarity of the document describing the target issue and other existing documents describing similar issues in each division (Figure 1). The basic methodology for finding appropriate departments is to search for similar documents, namely, selecting similar technical documents to the one with the current context. Many basic technologies for finding similar documents have been proposed, but many practical problems must be overcome before they can be practically applied directly.

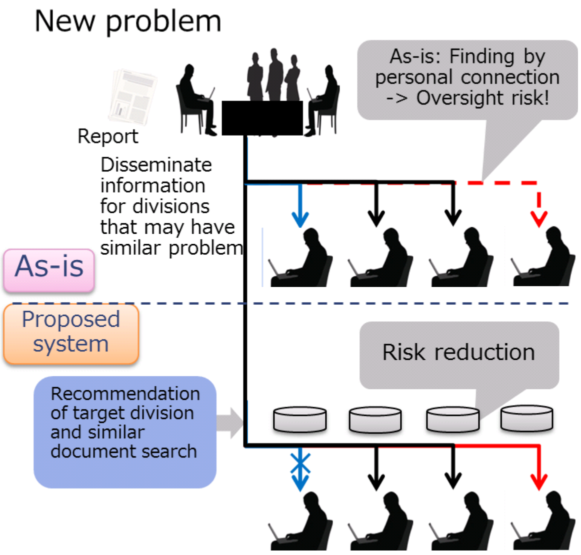

New problem
Report
As-is: Finding by personal connection -> Oversight risk!
Disseminate information for divisions that may have similar problem
As-is
Proposed system
Risk reduction
Recommendation of target division and similar document search

Figure 1: Use case: dissemination of information concerning quality-assurance activities

Hereafter, this problem is mainly dealt with from a practical perspective, and examples of our prototype construction are described. Especially, solving the problem involves addressing two challenges. First, it is necessary to establish a methodology for a mechanical decision-support system based on document information accumulated in every organization so far. The other challenge is to establish a method for accumulating know-how in preparation for the staff exchange accompanying generational change. To address these two challenges, the following two approaches are proposed:

1) Document similarity measurement and search system based on grouping of technical terms.
2) Getting feedback from users' behavior to confirm the reason for a judgment resulting from machine learning and improve the next judgement.

Although these two challenges are different problems on the surface, in terms of implementation, either process can be represented as a technique that deals with elements that define a topic of a document. The first one is a generating a topic from a document corpus, and the second one is evaluating the topic.

In the following chapters, the techniques used in our prototype decision-support system for quality assurance activities utilizing existing business documents and reports is introduced. Section 2 describes related research. Section 3 introduces some ideas established for this development. Section 4 describes the structure and procedure of the prototype system and show actual system images. Section 5 presents the results of a basic experimental evaluation of the prototype system. Note that the novelest point concerning the system is its user-interaction methodology used in the recommendation process (see Sections 3-4) that

1. shows feature terms for explaining classifications by using *LIME* [8];
2. finds other related documents from feature terms used in intuitive interactions.

Such interactions ensure the user's explicit attention for checking recommendation results and provide user's-action logs that are useful for analyzing the thinking process of expert users.

2 RELATED STUDIES

2.1 Document-similarity metrics

In the research field of NLP, the most-classical ways to represent a document are to use a bag-of-words (BOW) and term-frequency inverse document frequency (TF-IDF) in which each document is mapped into a vector space. These representations, however, cannot capture relations between different but similar words. There are several methods that attempt to circumvent this problem by mapping them into lower-dimensional representations. For example, latent semantic indexing (LSI) [1] and latent dirichlet allocation (LDA) [2] make vector spaces in which similar words are mapped into close points probabilistically, however, the effects are limited.

In 2013, a novel word-embedding procedure, called *word2vec* (W2V), was introduced by Mikolov et al. [3] This procedure makes a model that learns a vector representation for each word by using a shallow-neural-network language model and generates a representation of word embedding (or *word vector space*) of unprecedented quality and scales naturally to very large data sets. While W2V is widely adopted in most NLP projects, methods for providing document-similarity metrics are still controversial. For example, Doc2vec (paragraph vector) [4] is an extension of W2V and associates arbitrary documents with labels. As for this method, each document is assigned to a specific point in a *document vector space*. Although it seems a natural implementation, its results depend on the feature of the original label in the learning process. Also, it is impossible to know which element in the document defined the distance. Word movers' distance (WMD) [5] is another extension of W2V. WMD represents a document as a set of embedding words in W2V's word vector space, and it makes a document metric by formulating the distance between two documents as an optimal transport problem between the embedded words. Theoretically, in WMD, it is easy to analyze the influence of each word on the distance between documents, but its processing cost is very high. Furthermore, WMD differs from Doc2Vec in that document distances are entirely unsupervised and a mechanism for incorporating supervision is unavailable [6].

As a requirement of this application subject, the prototype system requires a means to explain document classification a posteriori, and a means to give feedback from human behavior as supervision.

2.2 Explaining classification

The classification decisions made by machine-learning models are quite difficult to explain to non-data scientists. Ribeiro's *Local Interpretable Model-Agnostic Explanations* (LIME) [7] is a method for making these models at least partly understandable. In the case of LIME, a model function is approximated by locally fitting linear models to permutations of the original training set. At least in the local approximation, the top-*n*-most-significant features for the classification can be shown. The explaining processes were mainly used by data scientists to validate the model. Our attempt in this study is to incorporate the explanation method into a user interface for daily use and to utilize it to obtain feedback from users.

3 METHOD

The goal of this study was to build a prototype system that classifies a given document that describes a target problem to particular divisions in a company in accordance with information from existing technical documents. Taking existing documents as data sources, the constructed prototype system uses a document-similarity metric for searching documents with various technical terms from a given corpus and automatically recommends not only documents but also divisions in which they should be deployed.

Before the prototype system is described in detail, current business activities and the target problem are described in Section 3.1, and the proposed method is explained in Section 3.2.

3.1 Target activity

A traditional and typical workflow of the target business activity, namely, a problem-checking process, is illustrated in Figure 2. When a new problem concerning existing products is encountered, members of staff of the QA department ask all other departments if they have a similar risk potential and explain the problem in detail and recommend countermeasures. In this department-searching process, QA staff elicit empirical knowledge from experts who are proficient in handling meta-level information within the organization (Figure 2).

However, when the organization inevitably changes, and talented people are replaced, the methods that rely upon such experience-based knowledge risk overlooking the target divisions in which the alert information should be deployed. Especially when the structures of the organization and human assets have been drastically changed owing to, for example, the merger with another company, there is a big risk of information development.

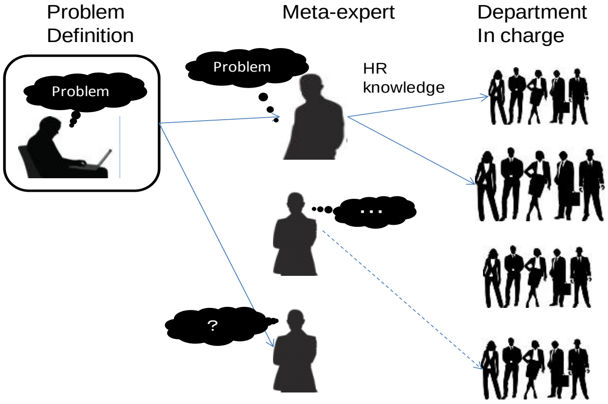


Figure 2: Current process for disseminating QA

3.2 Concept and outline of proposed system

The challenge in this study was to develop a system that identifies divisions that deal with similar cases and possibly have similar problems from document information on the basis of NLP. The fundamental framework of the proposed system is shown in Figure 3. It has two functions: to search for documents with similar context to the current context from a database of technical business documents ("1: Similar case finding" in the figure) and present the identified documents to the user. A new and additional function for searching for departments requesting confirmation of something on the basis of the similar documents ("2: Dissemination support") is introduced.

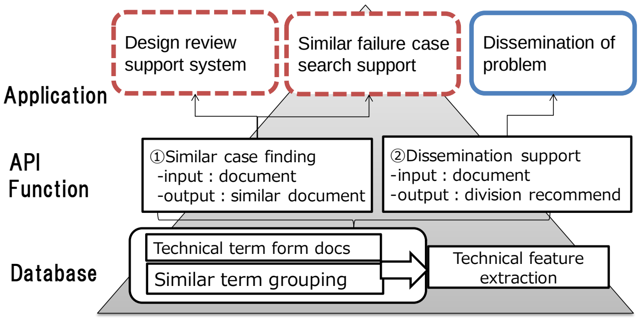


Figure 3: Total concept of the proposed system.

We describe the issues of the method along the following three points. The three key challenges to overcome in developing the proposed system are summarized below.

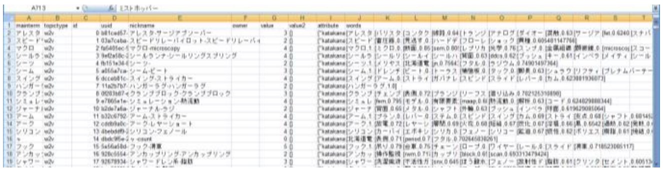
1. Document topics and similarity: How to define feature topics and similarity of documents (Section 3.3).
2. Modeling of classifier: How to extract the relationship between feature topics and specific divisions (Section 3.4).
3. Interaction and feedback: How to design the user interface of the system for enabling effective feedback of expert's knowledge (Section 3.5).

3.3 Topics and document similarity

Measuring the similarity of documents is a "situation-dependent" issue. Rather than using all words for judging the similarity, it is better to extract the specific technical terms related to the concerned context and to judge the document similarity on the basis of the similarity measure. Such a measurement process would make a more reasonable judgment of similarity. For example, under the assumption that groups of different corpus types are gathered together, such as design documents and operation manuals, accident-case documents, the frequency of occurrence of terms varies from corpus to corpus. To extract similar documents from these different corpora, it is effective to select the characteristic terms with more emphasis on the vocabulary elements commonly used across these corpora than to select terms locally used in each corpus. The method for extracting such a vocabulary set was studied and is described as follows.

<u>Topic grouping</u>

Term groups, named *topics*, are generated as follows.[1] Each topic consists of a main technical term (seeds) and several related technical terms tied by a particular feature that can be understood by humans. An example of such topics is shown in Figure 4.



	A	B	C	D	E	F	G	H	I	J
1	maintem	topictype	id	uuid	nickname	owner	value	value2	attribute	words
2	アレスタ	w2v	0	b81ced57-	アレスタ-サージアブソーバー		3	[]	["katakana	[アレスタ
3	スピード'	w2v	1	03a7caba-	スピードリレー-パイロット-スピードリレーパイ		2	[]	["katakana	[スピード
4	マクロ	w2v	2	fa6405ec-f	マクロ-microscopy		4	[]	["katakana	[マクロ.1
5	シールラ	w2v	3	9ef2a58c-2	シールランナ-シールリングスプリング		4	[]	["katakana	[シールラ
6	シーツ	w2v	4	fb151e34-f	シーツ-		2	[]	["katakana	[シーツ.1
7	シーム	w2v	5	a855a7ca-	シーム-ビート		3	[]	["katakana	[シーム.1
8	スイング	w2v	6	dcceb81c-	スイング-ストライカー		4	[]	["katakana	[スイング
9	ハンガー'	w2v	7	11a2b7b7-	ハンガーラグ-ハンガーラグ		2	[]	["katakana	[ハンガーラグ.1.0]
10	クランプ'	w2v	8	0f283b87-(	クランプブロック-クランプブロック		3	[]	["katakana	[クランプ
11	シミュレ'	w2v	9	e7865e1e-	シミュレーション-熱流動		4	[]	["katakana	[シミュレ
12	ジャーナ'	w2v	10	b2de7a6a-	ジャーナル-ラジ		2	[]	["katakana	[ジャーナ
13	アーム	w2v	11	b32c6792-	アーム-ストライカー		4	[]	["katakana	[アーム.1
14	アーク	w2v	12	cddb9abc-	アーク-レヤーショート		4	[]	["katakana	[アーク.1
15	シリコン	w2v	13	4bebddf9-2	シリコン-フェノール		4	[]	["katakana	[シリコン
16	ヶ	w2v	14	dbdc9f5e-2	ヶ-count		0	[]	["katakana	[北海道電
17	フック	w2v	15	5a56a58d-	フック-清掃		5	[]	["katakana	[フック.1
18	アンカッ'	w2v	16	928c5554-	アンカップリング-アンカップリング		2	[]	["katakana	[アンカッ
19	シャワー	w2v	17	92678934-	シャワー-ドレン系-脱脂		3	[]	["katakana	[シャワー

Figure 4: Example of topics (in Japanese)

In each document, the similarity of sentences is judged by the presence or absence of these topic terms in concerned sentence and the number of occurrences of that topic terms in each document. Each topic group is managed in the database and have a importance value that is updated with feedback from users' actions.

<u>Generating a topic</u>

A topic definition is composed of one representative technical term, a *seed* selected by a rule-based method, and groups of words selected as related terms in word2vec space.

<u>Selecting a seed of a topic</u>

In regard to the handling of documents, management of technical terms is very important. Technical documents used in manufacturing industries include a lot of specialized compound words, abbreviations, code numbers, etc. that are rarely used in general documents. These special terms are very important factors in defining similarity of such kinds of documents.

[1] ※The dimensionally compressed language model such as LSA[1] or LDA[2] is also called *TopicModel*, but *topic* we define here is a grouping of semantic information of finer granularity, and used in the meaning of topic tagging group proposed in [9].

The seed of each topic is a special technical term selected by various rule-based methods as listed below.

- Technical terms extracted from existing glossaries used in the business domain.
- English words expressed in Japanese *katakana* script.
- English abbreviations found in a Japanese corpus
- Various code numbers (e.g. product ID codes, project ID codes, and so on)
- Composite technical terms by continuation of *kanji (*extracted from documents by a statistical method [7])

Japanese technical documents used in business contain many specialized compound words, abbreviations, code numbers, etc. that are rarely used in general documents. They also include not only compound words with a series of *kanji* characters but also technical terms based on expressions in which English is written in *katakana* or a phonetic notation of foreign words. The characteristics of these Japanese documents can provide hints for retrieving technical terms.

Selecting related terms

The neighboring words of the seeds in each of the W2V spaces with different learning circumstances in common are selected as topic group words. Terms related to the selected species are then generated as follows.

1) Generate multiple Word2Vec model spaces from different kinds of corpus. By changing the combination of corpus used for learning and the order of learning, a W2V space of different learning models is made.
2) In each W2V model, the top-*N* vocabulary of the distance close to the species word is selected. (In this experiment *N* (number of words) is about 30.)
3) Pick up terms that are commonly selected in all W2V spaces as relatives of seeds.
4) Calculate the weight to each relative term. (The weight is a function of the distance between the seed and the term and the frequency of occurrence of the term.)

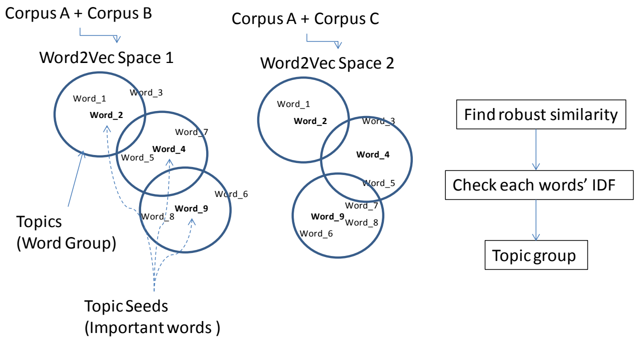


Figure 5: Extracting topic groups from multiple-distributed-word vector space.

Grouping attributes of topics

The prototype system uses about 30,000 kinds of topics. To support human understanding, 72 kinds of attribute tags for grouping topics were made. Each topic is assigned some of the attribute tags. The assignment is derived from the rule-based estimation. Each attribute tag represents a group of technical topics, e.g., "heat," "fluid," and "monitoring." The attribute tag is used in the characterization of the classifier (described in Sections 3.4 and 3.5).

Measuring similarity of document on the basis of topics

Measuring similarity of documents on the basis of topics is defined in Figure 6 as the following procedure.

1) Transform the bag-of-words representation into a topic matrix by linear transformation.
2) When a user selects a specific topic,
   ① add a weighting multiplier
   ② select similar topics
3) Take the dot product of the vectorized topics and perform threshold processing for each component.

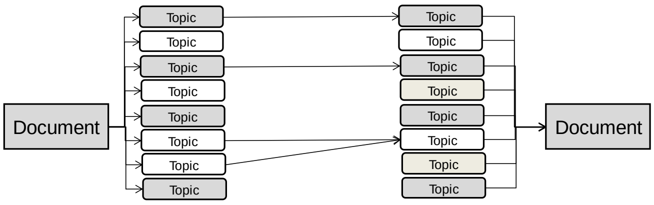


Figure 6: Document similar based on topic method. Measuring similarity of documents on the basis of topics

Evaluating topics on the basis of feedback

Each topic group is continuously maintained, and its evaluation value is updated with human feedback information.

3.4 Classification

For judging which department should be announced to the concerned case, a classifier is constructed so that its input is an existing document vectorized by topic expression and its output is a label (i.e.,, the name of the division that handled the case in question).

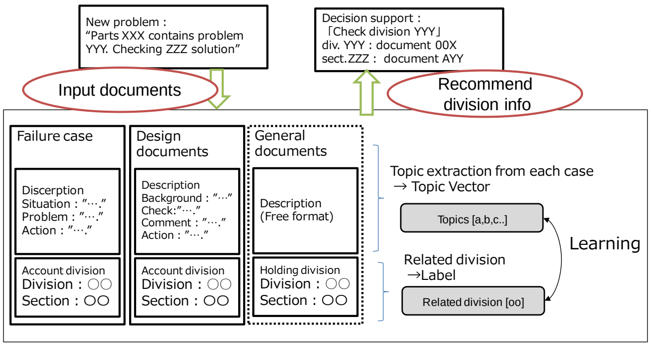


Figure 7: Classification of related departments by using existing document data.

As for the proposed system, multiple classifiers, each called an *agent*, are prepared, and each agent judges a division independently. To support human understanding, each agent specializes in a particular area, after which it is named. For example, *"agent of thermal attribute"* gathers only vocabulary related to elements such as heat, temperature,

gas, control, friction, melting, and output, and judges their classification independently.
The system consolidates the judgments of the agents to make an integrated judgment. The purpose of dividing the judgment task into multiple agents is to make an opportunity for human interaction. It enables to record information on what kind of agent's opinion is confirmed by the user and to obtain information of the perspective in the judgment process. As a result, information on what kinds of agent's opinion are confirmed by people can be recorded, and the agent's judgement process can be put into perspective.

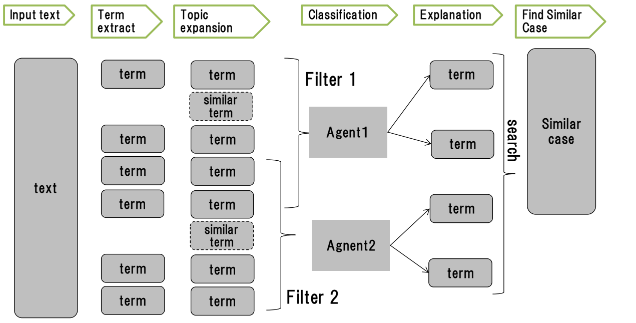


Figure 8: Process for searching for divisions and documents by using classification agents.

3.5 User interaction and feedback

A distinctive point concerning the proposed system is the interactive process that explains the reason for the recommendation of target departments The entire user interaction process is designed to answer the following three questions and store the answers as operation logs:

(a) Which perspectives are used in classification?
(b) Which topics (terms) are focused on by users to find similar problems?
(c) Which documents identified the target department?

The flow of the process for the user to check the current problem is illustrated in Figure 9.

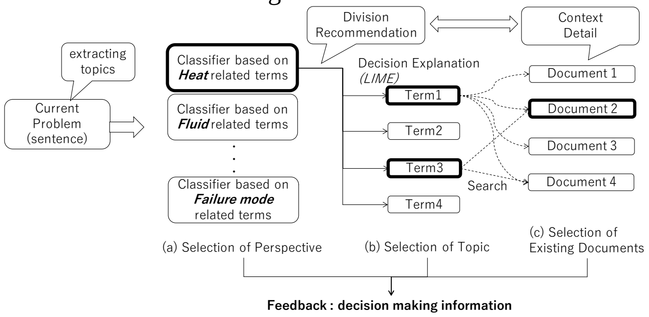

Division Recommendation
Context Detail
extracting topics
Current Problem (sentence)
Classifier based on *Heat* related terms
Classifier based on *Fluid* related terms
Classifier based on *Failure mode* related terms
Decision Explanation (LIME)
Term1
Term2
Term3
Term4
Search
Document 1
Document 2
Document 3
Document 4
(a) Selection of Perspective
(b) Selection of Topic
(c) Selection of Existing Documents
Feedback : decision making information

Figure 9: Process for user interaction. Among agents' explanations, a topic selected by the user is used for checking related documents.

(a) Selection of perspective
Several topics are extracted from the subject problem, and the topic vector obtained is applied to the classifier of agents. There are several agents, and each one is assigned an attribute that it is responsible for. The classifier filters the given data and selects only those words that are relevant to the agent's own attribute and topics, and it performs the judgment processing of the related department on the basis of the screened data. The overall result is judged from the result of the processing of each classifier. On the other hand, from the classifiers, the user selects a particular classifier that is similar to their own perspective of interest, and checks the topic that became the reason for judgment.

(b) Selection of topics
The system displays the topic terms that are important reason elements in the judgment of the selected classifier. The method for selecting the terms explaining this decision is used method. LIME [7] was used to select terms that explain the reason for this decision. From the explanatory topics, the user selects some important topics that need to be confirmed and searches for related documents.

(c) Selection of existing documents
The system searches for similar documents that contain the selected words and displays the documents in a group of related departments. The user checks the contents of the documents held by the recommended department and selects the documents that may need to be specifically confirmed. To the selected document, the user adds a comment to the department's staff related to the topic, and then the system adds the document and comment to the checklist to ask for confirmation.

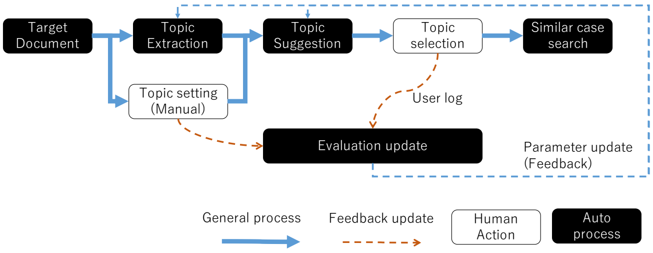


Figure 10: Procedure for feedback from the user

By logging the user's UI operation by the above three processes, it is possible to estimate what kind of the perspectives (classification agents), topics, and documents are used to determine the expert users' checking points.

4 APPLICATION OF PROPOSED SYSTEM

The prototype system operates according to the following procedure in three stages.

Preparation stage: Each document is labeled by each according to its corresponding division and stored in the database. Topic elements are then extracted from the documents. The classification agents learn the feature terms of each department on the basis of the extracted information in advance.

Sending stage: A concerned document to be disseminated and additional topics given by a QA stuff (sender) are represented as a topic vector. The documents related to this topic vector are used as a search key, and the relevant-department candidates to contact and the related documents are confirmed.

Receiving stage: In each department, the recipient checks the incoming message. At that time, the system explicitly presents the documents and topics referenced by the sender. It supports the consideration of the problem on the recipient.

4.1 Procedure and interaction

The process of the sending stage is shown in detail in Figure 11. The proposed system works along five steps to help the sender find the target. The sending stage consists of the following five blocks.

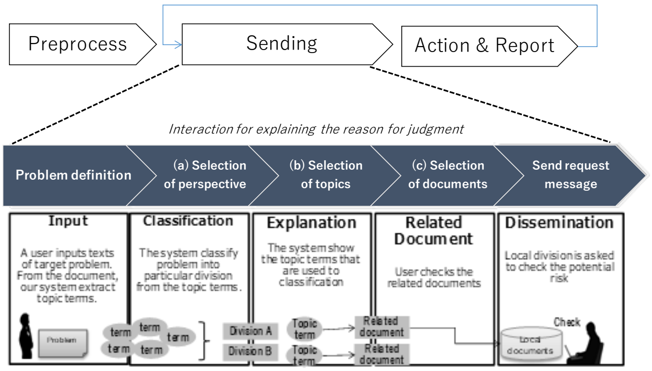
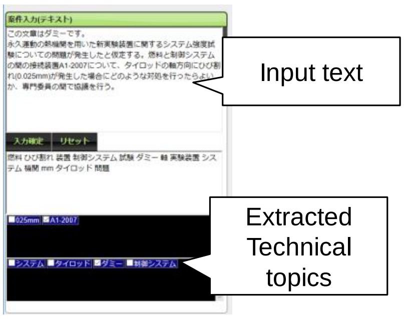

Preprocess
Sending
Action & Report
Interaction for explaining the reason for judgment
Problem definition
(a) Selection of perspective
(b) Selection of topics
(c) Selection of documents
Send request message
Input
A user inputs texts of target problem. From the document, our system extract topic terms.
Problem
term
Classification
The system classify problem into particular division from the topic terms.
Division A
Division B
Explanation
The system show the topic terms that are used to classification
Topic term
Related Document
User checks the related documents
Related document
Dissemination
Local division is asked to check the potential risk
Check
Local documents

Figure 11: Procedure of sending process

Step 1: Input problem
An example of the input screen is shown in Figure 12. The user inputs a sentence to define the context of the problem. Topic keywords are extracted from the input sentences and displayed on the screen. From them, the user selects topics to be focused on.


案件入力(テキスト)

この文章はダミーです。
永久運動の熱機関を用いた新実験装置に関するシステム強度試験についての問題が発生したと仮定する。燃料と制御システムの間の接続装置A1-2007について、タイロッドの軸方向にひび割れ(0.025mm)が発生した場合にどのような対処を行ったらよいか、専門委員の間で協議を行う。

入力確定 リセット

燃料 ひび割れ 装置 制御システム 試験 ダミー 軸 実験装置 システム 機関 mm タイロッド 問題

☐025mm ☑A1-2007

☐システム ☐タイロッド ☑ダミー ☐制御システム

Figure 12: Display image of input step. Technical topics are extracted from the input sentence.

Step 2. Check classification result
By comparing the extracted keywords and the document of the accumulated past cases, the system recommends possible relevant departments presumed to cope with the problem. Each agent recommends departments independently, and an overall recommendation is made according to the total judgments of all agents.

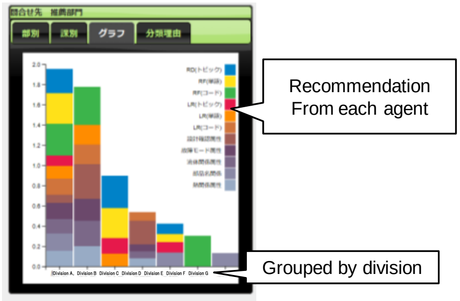
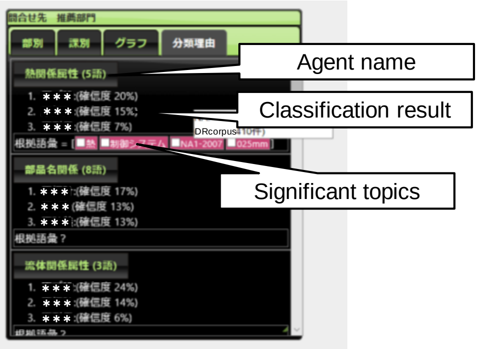


Figure 13: Display image of classification step.

Step 3. Check explanation terms
Using LIME, each agent presents topics that have a great influence as “explanation.” These topics are limited to the vocabulary that may be related to the agents' attributes. As shown in Figure 14, the user can check each agent's judgement result and the reason for that judgement.


問合せ先 推薦部門

部別 課別 グラフ 分類理由

熱関係属性 (5語)
1. ＊＊＊:(確信度 20%)
2. ＊＊＊:(確信度 15%)
3. ＊＊＊:(確信度 7%)
根拠語彙 = [☐熱 ☐制御システム ☐NA1-2007 ☐025mm]

DRcorpus(410件)

部品名関係 (8語)
1. ＊＊＊:(確信度 17%)
2. ＊＊＊(確信度 13%)
3. ＊＊＊:(確信度 13%)
根拠語彙？

流体関係属性 (3語)
1. ＊＊＊:(確信度 24%)
2. ＊＊＊:(確信度 14%)
3. ＊＊＊:(確信度 6%)
根拠語彙？

Figure 14: Display image of explanation step. The user checks topic terms to be focused on.

Step 4. Checking for related documents
By selecting the topic, After the topic is selected, the documents that include the topics are retrieved from the document database. Examples of a UI for checking these documents are shown in Figure 15. The chart shows the documents in a timetable grouped according to department. The system logs the referenced documents and gets feedback for verification data on the importance of the vocabulary.

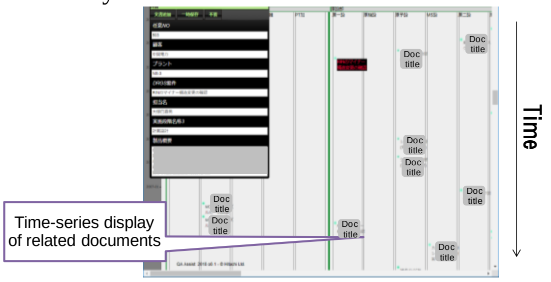


Figure 15: Display image of drill-down step. The user checks related documents in this phase.

Step 5. Dissemination of documents to target division
The system converts the list of checked related documents during the operation into a message document for the confirmation request to the target division.

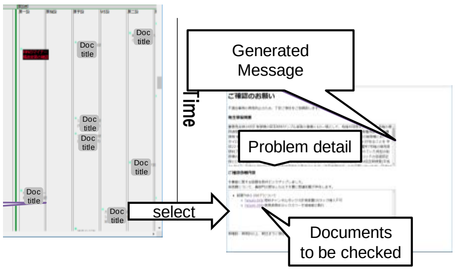


Figure 16: Display image of dissemination. The system makes a request template message for local divisions.

Finally, the user sends the confirmation documents created by the above process to the contact person in each department.

5 EXPERIMENT

As a basic experiment to evaluate the basic concept of the proposed system, the following contents were evaluated quantitatively: (i) the effect of using the proposed method (topic grouping) and (ii) whether the subject and the target department can be identified from the frequency of occurrence of the topic of the business document.

Experiment 1

A case study of a similar-case search in the design-review business is explained as follows. In this evaluation, the evaluation score obtained with the proposed method and that obtained with an existing similar-document-search method were compared. The existing search engine used as the benchmark is based on associative-search technology [8], which selects feature words from a given document group and selects the similar documents from that group. It is highly regarded as a general-purpose similar-document retrieval system. This approach is similar to that of the proposed content, and the difference is that the words are selected directly by using W2V without using a topic group. The effect of the topic group that was adjusted as a benchmark for this existing system was verified.

In this evaluation, four kinds of design-review tasks were targeted. First, 50 corpora considered to have a high probability related to the target document were extracted. For the primary-extraction data, the user decided whether each corpus was related to the context of the design review and labeled the success and failure.

Table 1: Comparison of results of similar-document search by proposed (topic-based) and existing method.

	Proposed precision	Existing precision	Improvement
Case#1	0.36	0.36	1.0
Case#2	0.5	0.44	1.13
Case#3	0.95	0.8	1.18
Case#4	0.52	0.26	2.0
average	0.583	0.465	1.328

Table I lists conformity rates of the label that was assigned by a user using the evaluation value given by this system. The rate of conformity is less than 58% on average. Compared with the existing method, the proposed method shows an average improvement in precision of 30%.

Experiment 2

It was experimentally confirmed whether the subject and target department can be identified from the occurrence frequency of the topic of the business document. The total number of documents used in the test experiment was 1954, separated into 1600 data for study and 354 test data. As for the number of divisions that were classified, divisions were classified into 23 categories, and the sections (sub group of divisions) were divided into 136 categories.

The results of experiments on identifying related divisions (23 departmental units) from each sentence are listed in Table 2. "Top choice" is the rate at which the highest rating was the actual responsible department. "Three choices" is a case in which three possible related departments are recommended, and the actual responsible division was among the three options. Other results when each agent classified cases into 136 sections are listed in Table 3.

As a goal set at the start of this study, a target of 70% was set as the provisional recommendation accuracy. However, the actual value is only about 60% when the recommendation value of top choice is considered, and it is inadequate for the process for estimating automatically (without human assist) especially in units of departments. However, from perspective actual operation, since human interaction is carried out from people interact in selecting three types of recommendation candidates to determine the deployment destination, it is important to select the top-three problem areas.

At the present stage, in terms of the accuracy required for QA activities, we could not obtain sufficient accuracy by automatic recommendation alone, and human interaction (focusing on appropriate topics) is necessary for practical use. In future works, we must conduct an experiment to accumulate information by actually introducing the proposed method in the field and verifying its effect on improvement of classification

Table 2: Classification accuracy (division: 23 labels)

Classifier unit name	Top choice	Three choices
Code	0.48	0.681
Term (general)	0.668	0.844
Topic (general)	0.603	0.857
Heat	0.467	0.772
Parts	0.584	0.792
Fluid	0.525	0.688
Failure and malfunctions	0.454	0.714
System design	0.467	0.746

Table 3: Classification accuracy (division: 136 labels)

Classifier unit name	Top correct	Three choices
Code	0.415	0.623
Term (general)	0.617	0.72
Topic (general)	0.558	0.766
Heat	0.467	0.668
Parts	0.526	0.727
Fluid	0.454	0.636
Failure and malfunctions	0.481	0.642
System design	0.396	0.59

6 CONCLUSION

A prototype decision-support system for QA activity was proposed and experimentally evaluated. The proposed system classifies documents that describe target problems and assigns them to particular divisions in accordance with the similarity of existing technical documents. It also utilizes a description of the machine’s decision as a clue to user interaction.

The goal of this method is to improve performance by human feedback. However, long-term accumulation of user data in the field operation is indispensable in regard to this goal. At present, we are working on an experimental phase in the field with various additional developments and brushing up the method for data accumulation and performance improvement.

There is concern that AI supporting human judgment may interfere with an expert user’s sense of responsibility. If the judgment of AI reaches a reliable level, confirmation of judgment by the user might become a ritual. As a means to avoid this situation, we think that it is important to create a procedure to obtain feedback that uses “checking reason for a machine’s judgment” for “another searching process” and involve the expert users in order to accumulate information of know-how. This prototype will be a typical case of such UI procedure.

REFERENCES

[1] T. K. Landauer and T. S. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, Vol. 104 No. 2, pp. 211-240. (1997)

[2] D. Blei et al., Latent Dirichlet Allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022. (2003)

[3] Y. Goldberg and O. Levy, word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method, Cornell University Library (2014)

[4] Q. Le and T. Mikolov, Distributed Representations of Sentences and Documents, PMLR, Vol. 32, pp. 1188-1196, (2014)

[5] M. J. Kusner et al., From word embeddings to document distances. ICML 2015, Vol. 37, pp. 957-966, (2015)

[6] G. Huang et al., Supervised word mover's distance, NIPS'16, pp. 4869-4877, (2016)

[7] M. Tulio et al, Why Should I Trust You?: Explaining the Predictions of Any Classifier, ACM SIGKDD 2016, pp. 1135--1144, (2016)

[8] H. Nakagawa and T. Mori, A simple but powerful automatic term extraction method, Computerm2, COLING 2002 Workshop, pp. 29–35, (2002)

Initial Attempt of Bluetooth Low Energy Water-Level Estimator

Ryo Orihara[†], Shigemi Ishida[†], Masahiko Miyazaki[†], Shigeaki Tagashira[‡], and Akira Fukuda[†]

[†]ISEE, Kyushu University, Fukuoka 819-0395, Japan
{orihara, ishida, miyazaki, fukuda}@f.ait.kyushu-u.ac.jp
[‡]Faculty of Informatics, Kansai University, Osaka, 569-1095, Japan
shige@res.kutc.kansai-u.ac.jp

Abstract -

In recent years, torrential downpours in a limited area occur frequently. To avoid flooding damages, it is important to continuously monitor the changes in river conditions, especially changes in the water-level. This paper proposes a low-cost water-level estimation method using off-the-shelf BLE (Bluetooth Low Energy) devices. BLE uses a 2.4-GHz band, where the radio signals decay at a rapid rate in the water as we increase the communication distance. We measure RSS (received signal strength) of BLE signals and estimate the river water-level. The influence of RSS fluctuations are reduced by employing the RSS difference between two BLE beacons installed in and by a river. We conducted an initial experiment to confirm the feasibility of our water-level estimation method. The experimental evaluation demonstrated that our water-level estimator successfully estimated water-level within a six classes of water-level with an absolute mean error of 2.75 centimeters.

Keywords: BLE (Bluetooth Low Energy), RSS (received signal strength), water-level

1 INTRODUCTION

In recent years, torrential downpours in a limited area occur frequently. The heavy rains sometimes cause river-bank collapse resulting in flooding disasters. Flooding damages not only houses but also people who fail to evacuate from the damaged area. To avoid such damage, it is important to continuously monitor the changes in river conditions, especially changes in the water-level.

Water-level is manually monitored at many rivers in Japan. To monitor river water-level in realtime, water-level monitoring system using sensors such as floating and bubble water-level gauges have been deployed at some rivers. The automatic water-level monitoring systems, however, come with high installation and maintenance costs, resulting in limited number of deployment. Monitoring of the river water-level in Japan is mainly carried out on class A rivers designated by the MLIT (Minister of Land, Infrastructure, Transport and Tourism) and class B rivers designated prefectural governor. For a total of 21,004 rivers managed by prefectures, 4,986 sensors have been installed, which indicates that less than 25 % of rivers are covered by automatic water-level monitoring systems.

In the heavy rain in northern part of Kyushu in July 2017, 32 rivers managed by each prefecture were flooding and caused serious damages. 30 rivers among the 32 lacked water-level gauges. Although many local governments are aware of the necessity of water-level gauge, high installation and maintenance costs of water-level gauges prevent the deployment.

This paper proposes a water-level estimation method using off-the-shelf BLE (Bluetooth Low Energy) devices. BLE uses a 2.4-GHz band, where the strength of radio signals decays at a rapid rate in the water as we increase the communication distance. We measure RSS (received signal strength) of BLE signals sent from a BLE beacon in a river water using a riverside BLE module and estimate the river water-level based on the RSS. Temperature or environmental changes may affect transmission and reception circuit operations, which results in changes of RSS. We therefore employ the RSS difference between two BLE beacons installed in and by a river. We also employ packet loss rate for water-level estimation because the estimation performance might degrade due to the decrease of the number of received packets as the water-level increases.

To verify the feasibility of the water-level estimation system using BLE, we experimentally evaluated the water-level estimation system. We changed water-level in six steps in a bathtub and collected RSS samples. The 10-fold leave-one-out cross validation demonstrated that our water-level estimation system successfully estimated water-levels in six steps with an accuracy of 85.3 %, which resulted in an absolute mean error of 2.75 centimeters.

Specifically, our main contributions are three-fold:

- We propose a new water-level estimation system that uses low-cost off-the-shelf BLE beacons. The water-level estimation system estimates water-level by machine learning using RSS (received signal strength) of advertising packets sent from BLE beacons.
- We present an initial design of the water-level estimation system that relies on two BLE beacons. Using two BLE beacons, we reduce the influence of RSS fluctuations.
- We show the feasibility of our BLE water-level estimator by initial evaluation conducted in a bathroom.

The remainder of this paper is structured as follows. Section 2 shows an overview of water-level estimation system using BLE and issues in a practical environment. Section 3 describes a design of the water-level estimation system in a practical environment and Section 4 conducts an initial experimental evaluation. Related work is presented in Section 5. Finally, Section 6 concludes the paper.

2 BLE WATER-LEVEL ESTIMATION SYSTEM

Figure 1 depicts an overview of water-level estimation system using BLE. In the water-level estimation system, BLE beacon and receiver are installed in a river and by the riverside, respectively. The BLE beacon periodically sends advertising packets. We receive the periodic advertising pack-

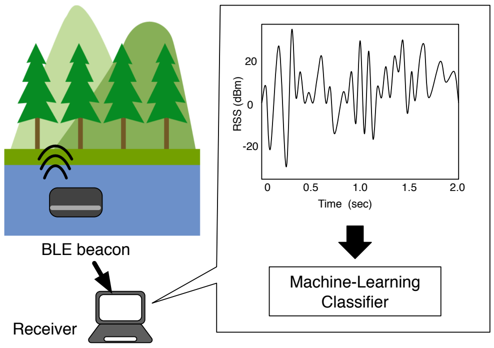


Figure 1: Overview of water-level estimation system

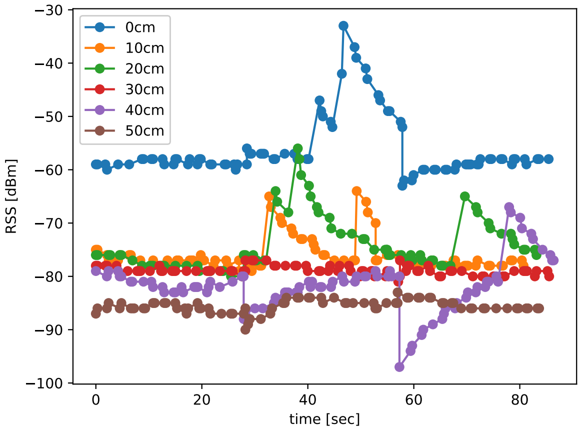


Figure 2: Variation of RSS

ets using the BLE receiver and measure RSS (received signal strength) of the packets. BLE uses a 2.4-GHz band and the strength of radio signals decays at a rapid rate in the water. As the water-level increases, RSS of the advertising packet decreases. We estimate the water-level from the RSS at the receiver by a machine-learning classifier.

In a practical environment, we need to solve the following two issues.

1) *How do we reduce the effects of RSS fluctuations?:* Even when the water-level is unchanging, RSS is too unstable for water-level estimation. Figure 2 shows an example of RSS of advertising packets sent from a BLE beacon installed in a bathtub as a function of time. We changed water-level from 0 to 50 centimeters with 10-centimeter step and collected RSS samples. Intuitively, RSS decreases as the water-level increases. Figure 2, however, shows that the intuition is sometimes not true because of the RSS fluctuations. The RSS fluctuations are mainly caused by changes in radio propagation environment and in operations of transmitter/receiver circuits.
2) *How do we reduce the influence of RSS error due to decrease in the number of received packets?:* RSS of packets sent from a BLE beacon in a river decreases as the water-level increases, which results in increase in packet error rate. In such a case, water-level estimation error increases because of the RSS measurement error due to the limited number of received packets.

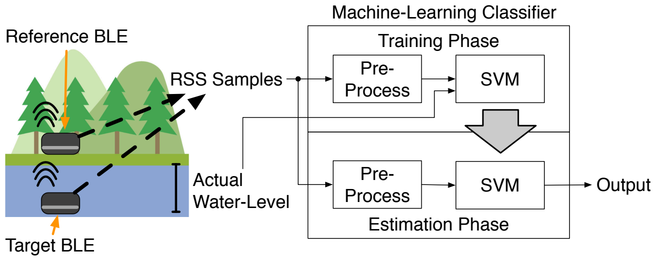


Figure 3: Overview of water-level estimation system

3 EXTEND TO PRACTICAL ENVIRONMENT

3.1 Basic Idea

We address the two issues presented in the previous section based on the following two basic ideas.

1) *Utilize RSS difference:*

   We utilize RSS difference between two BLE beacons installed in and by a river to estimate water-level. Using the RSS difference, we can remove the influence of operational changes in a receiver circuit.

2) *Utilize packet reception interval:*

   We use packet error rate as one of the features for water-level estimation to improve accuracy. Although packet error rate is not always independent of RSS of advertising packets, packet error rate is also affected by factors other than the decrease of RSS. There is no API defined in BLE standards to retrieve packet error information [1]. In addition, BLE beacons add a random delay on every transmission of advertising packets, which makes difficult to calculate packet reception rate from the number of received packets within a fixed duration. We therefore use packet reception interval instead of packet error rate.

3.2 System Overview

Figure 3 shows an overview of the water-level estimation system. As shown in Fig. 3, the water-level estimation system uses target and reference BLE beacons, and a BLE receiver. A target BLE beacon is installed at the bottom or the target depth in a river, and a reference BLE beacon is installed by the riverside. A BLE receiver installed by the riverside receives advertising packets from the BLE beacons and measure the packet RSS. The RSS data is then sent to a machine-learning classifier that consists of pre-processing and SVM blocks.

The following subsections describe the each process in details.

3.3 Pre-Processing Block

Figure 4 shows an overview of a pre-processing module. The pre-processing block reformats RSS samples to keep a specific sampling rate and calculates the RSS difference between target and reference BLE beacons. BLE beacons are periodically sending advertising packets with a random delay

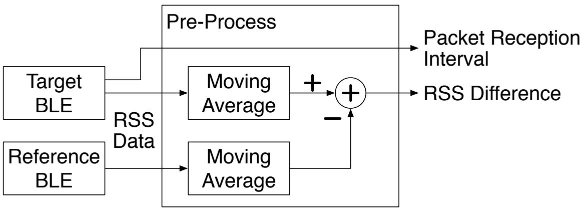


Figure 4: Overview of pre-processing block

on each transmission. The RSS samples of advertising packets are averaged over a specific period to derive the specific number of RSS data. The number of RSS data is dependent on packet reception errors and the random delay. We derive the inverse of the actual number of averaged RSS samples as a packet reception interval.

3.4 SVM Block

An SVM (support vector machine) block consists of training and estimation phases. The input of the SVM block is a vector with two components: the RSS data and packet reception interval which are passed from the pre-processing block.

In a training phase, test data that consists of the RSS difference, packet reception interval, and actual water-level is passed to an SVM classifier to construct an SVM machine-learning model. We used a multi-class SVM classifier with the RBF kernel as we changed water-level from 0 to 50 centimeters with 10-centimeter step in our experiment. The RBF kernel parameter and cost parameter are optimized by grid search.

In an estimation phase, the water-level is estimated by the pre-trained SVM classifier. The feature values, i.e., the RSS difference and packet reception interval, are passed from the pre-processing block in the same manner as the training phase and water-level is estimated within the six steps.

It is natural to use SVM regression for continuous water-level estimation. We, however, use SVM classifier as an initial evaluation in this paper because unstable RSS might fail to construct a valid regression model. In the near future, we will collect much more training data and will utilize SVM regression.

4 INITIAL EVALUATION

To verify the effectiveness of the water-level estimation system using BLE, we conducted experiments in a bathroom as an initial evaluation because the bathroom is easy to control the water-level.

4.1 Preliminary Experiment

A preliminary experiment was conducted to determine the time duration of the moving average in a pre-processing block. Frequency component analysis was performed on the RSS samples using FFT (Fast Fourier Transform). To apply FFT, sampling frequency must be constant. We first derived constant sampled RSS by linear interpolation because BLE beacons are sending advertising packets with a random delay on each transmission. We then analyzed frequency components of the interpolated RSS samples.

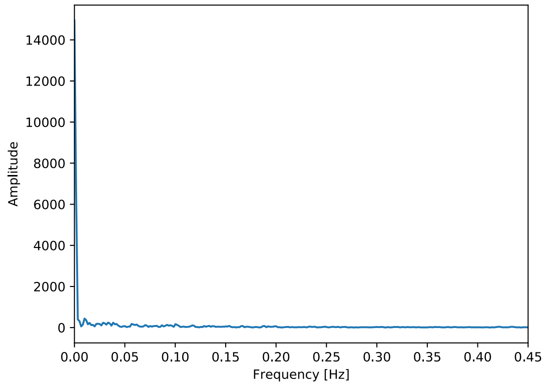


Figure 5: Frequency components of RSS samples

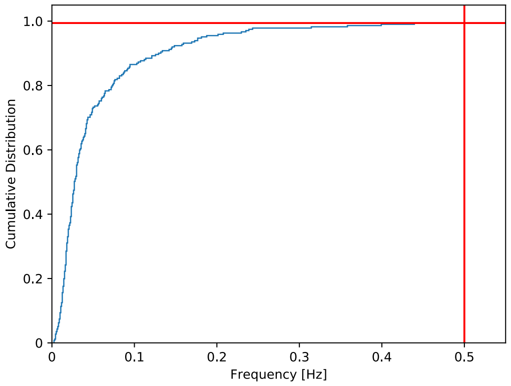


Figure 6: Empirical cumulative distribution of frequency components

Figures 5 and 6 shows the result of frequency component analysis and empirical cumulative distribution of the frequency components, respectively. The figures indicate that 99.4 % of frequency components were below 0.5 Hz. A BLE receiver receives a packet at least every two seconds. We determined to use 2-second window for moving average in a pre-processing block.

4.2 Implementation

Figure 7 shows the equipments used in our implementation. We used a MacBook Pro laptop running macOS Sierra 10.12.6 as a BLE receiver and two MyBeacon Pro (MB004 Ac-DR) BLE beacons, as shown in the figure. We implemented node.js program running on a BLE receiver to receive BLE advertising packets from the BLE beacons. Water-level estimation, i.e., pre-processing and SVM classifier are implemented using Python 2.7.13 with scikit-learn 0.19.1.

4.3 Evaluation Environment

An experimental evaluation was carried out in a condominium bathroom. The bathroom setup is shown in Fig. 8. Figure 9 shows experiment setup. Two BLE beacons were installed as shown in Fig. 9: a target BLE beacon at the bottom of the bathtub and a reference BLE beacon outside of the bathtub. A BLE receiver, namely, a laptop, was installed above the bathtub.

Figure 7: Equipments used in the implementation

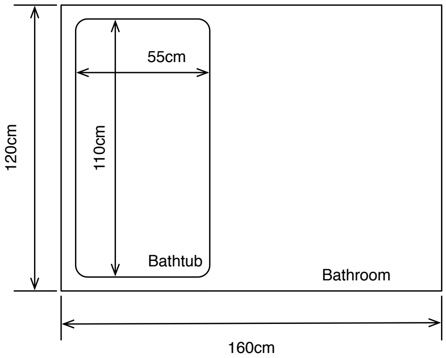


Figure 8: Size of bathroom

We changed the water-level from 0 to 50 centimeters with 10-centimeter step and measured RSS of received packets for 10 minutes for the each water-level. Using the RSS samples, we performed 10-fold leave-one-out cross validation to estimate water-level in six water-level classes. For each water-level, we randomly sorted the data. We then divided the data for the each water-level into ten chunks, nine of which were used for training and the remaining was used for testing.

Comparing the estimated results with the actual water-level, we calculated precision and recall for each water-level class using a confusion matrix. Macro precision and recall, which are the mean of precision and recall for each class, were also calculated. We also calculated macro F1 score from the macro precision and recall defined as

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{1}$$

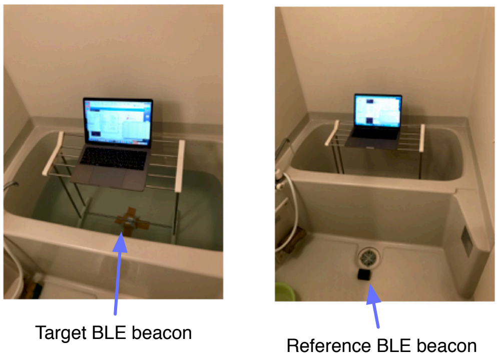


Figure 9: Experiment setup

Table 1: Confusion matrix of water-level estimation result

Estimated [cm]	Actual [cm]					
	0	10	20	30	40	50
0	297	0	0	1	0	0
10	0	280	19	0	0	0
20	0	26	159	4	93	0
30	0	8	2	284	4	0
40	0	18	57	0	211	13
50	0	0	0	0	1	299

Table 2: Precision and recall for each water-level class

Class	0	10	20	30	40	50
Precision	0.997	0.936	0.532	0.953	0.706	0.997
Recall	1.00	0.843	0.671	0.983	0.683	0.909

4.4 Classification Performance

Table 1 shows a confusion matrix of the water-level estimation result. The diagonal of the confusion matrix, i.e., gray-colored cells, represents the number of TPs (true positives), which are the cases that water-level is correctly estimated, for each water-level class. We can see that the water-level estimation system exhibited good estimation performance in water-level classes of 0 and 50 centimeters. This was mainly caused because the RSS of BLE signals above or under specific levels were easily estimated as 0- or 50-centimeter class, which resulted in the good performance.

Table 2 shows precision and recall calculated from the confusion matrix for each water-level class. As the confusion matrix implied, precision and recall for 10- and 50-centimeter classes were quite high. Precision and recall were the highest for a 0-centimeter class. 20- and 40-centimeter classes were difficult to be distinguished because RSS data of the both classes had small difference. It is considered that the estimation accuracy of 20- and 40-centimeter class happened to be low by chance, because the number of data was insufficient in this time. In the future, we plan to utilize RSS distribution to estimate water-level.

Macro precision, recall, and F1 score were 0.853, 0.848, and 0.851, respectively. Although the results were insufficient for practical use, the results demonstrated the feasibility of water-level estimation using BLE beacons as we derived the estimation performance above random selections.

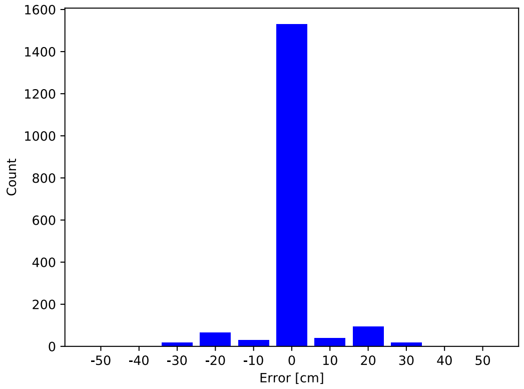


Figure 10: Distribution of water-level estimation errors

4.5 Water-Level Estimation Error

We evaluated water-level error because our system not for classification but for water-level estimation. The water-level error ε_i for trial i is defined as

$$\varepsilon_i = \widetilde{d_i} - d_i, \quad (2)$$

where $\widetilde{d_i}$ and d_i is estimated and actual water-levels for trial i. Note that we estimated water-level within six classes of water-level. The error is therefore in $\{0, \pm 10, \pm 20, \pm 30, \pm 40, \pm 50\}$.

Figure 10 shows distribution of water-level errors. From the figure, we can see that our water-level estimation system correctly estimated water-level for 85.3 % of trials. An absolute mean error $\overline{|\varepsilon|} = \sum_i |\varepsilon_i|$ was 2.75 centimeters. This result greatly exceeded the result that an absolute mean error was 4.79 centimeters when using a single BLE beacon, which demonstrated the effectiveness of the two-beacon approach. We believe that the results successfully demonstrated the feasibility of our water-level estimation method using BLE beacons.

5 RELATED WORK

To the best of knowledge, there is no study working on estimation of river water-level using off-the-shelf radio modules. This section reviews studies on water sensing technologies and sensing technologies using radio signals.

5.1 Water-Level Sensing

Water-level sensing technology is not much studied in literature because there are many off-the-shelf water-level sensors such as floating and bubble water-level gauges available. There is another form of water-level sensors relying on ultrasound, radar [2], and laser [3] available today. These approaches require high cost equipments or high installation and maintenance costs.

Image-based water-level sensing technologies have been reported [4,5]. River Eye [5] installs cameras by the riverside and other system components are deployed as a cloud system, which minimizes installation and maintenance costs. Combined with the cloud-based approach with our BLE-based sensing technology, we believe that we can much more reduce the installation and maintenance costs.

5.2 Wireless Sensing

There are some studies reporting wireless sensing technologies in terms of water sensing. Wi-Wheat presents a nondestructive and economic wheat moisture sensing system with commodity WiFi [6]. Wi-Wheat utilizes amplitude and phase shift of WiFi wireless links in CSI (channel state information) data derived from a commodity WiFi module. Machine-learning using SVM with features extracted by PCA (principle component analysis) on CSI data estimates wheat moisture level. The experimental results demonstrated the Wi-Wheat system can achieve higher classification accuracy for both LOS (line-of-sight) and NLOS (non line-of-sight) scenarios. The use of commodity WiFi drastically reduces the cost of moisture sensing compared to the existing grain moisture sensing technologies relying on magnetic or electric properties [7–12].

Wireless sensing technologies using commodity WiFi devices are extended to gesture recognition nowadays. WiGest leverages changes in WiFi signal strength to sense in-air hand gestures around the user's mobile device [13]. Smokey utilize CSI changes to detection smoking motion [14]. Both technologies utilize change of radio signal propagation environment. Utilizing the signal processing methods presented in these works to extract features, we may improve water-level estimation accuracy, which is one of our future works.

6 CONCLUSION

This paper proposes BLE-based water-level estimator as a new low-cost water-level measurement method. We conducted an initial experiment and confirmed that the BLE water-level estimator estimated water-level with small errors in a controlled environment. In our future work, we conduct experiments in a practical environment.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers JP15H05708 and JP17H01741.

REFERENCES

[1] Bluetooth Special Interest Group, "Bluetooth specification version 4.2," Dec. 2014, http://www.bluetooth.com/.

[2] W. Sun, H. Ishidaira, and S. Bastola, "Calibration of hydrological models in ungauged basins based on satellite radar altimetry observations of river water level," *Hydrological Processes*, vol. 26, no. 23, pp. 3524–3537, Nov. 2011.

[3] Y. Ji, M. Zhang, Y. Wang *et al.*, "Microwave-photonic sensor for remote water-level monitoring based on chaotic laser," *Int. J. Bifurcation and Chaos*, vol. 24, no. 3, pp. 1–7, Mar. 2014.

[4] J. Kim, Y. Han, and H. Hahn, "Embedded implementation of image-based water-level measurement system," *IET Computer Vision*, vol. 5, no. 2, pp. 125–133, Mar. 2011.

[5] Y. Kim, H. Park, C. Lee *et al.*, "Development of a cloud-based image water level gauge," *IT Converg. Pract.(INPRA)*, vol. 2, no. 1, pp. 22–29, 2014.

[6] W. Yang, X. Wang, A. Song *et al.*, "Wi-wheat: Contact-free wheat moisture detection with commodity wifi."

[7] W. Wang and Y. Dai, "A grain moisture detecting system based on capacitive sensor," *International Journal of Digital Content Technology and its Applications*, vol. 5, no. 3, pp. 203–209, 2011.

[8] Z. Liu, Z. Wu, Z. Zhang *et al.*, "Research on online moisture detector in grain drying process based on v/f conversion," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[9] S. O. Nelson, A. W. Kraszewski, S. Trabelsi *et al.*, "Using cereal grain permittivity for sensing moisture content," *IEEE transactions on instrumentation and measurement*, vol. 49, no. 3, pp. 470–475, 2000.

[10] K. Kim, J. Kim, C. Lee *et al.*, "Simple instrument for moisture measurement in grain by free-space microwave transmission," *Transactions of the ASABE*, vol. 49, no. 4, pp. 1089–1093, 2006.

[11] Y. Yang, J. Wang, C. Wang *et al.*, “Study on online measurement of grain moisture content by neutron gauge.” *Transactions of the Chinese Society of Agricultural Engineering*, vol. 16, no. 5, pp. 99–101, 2000.

[12] D. Nath K and P. Ramanathan, “Non-destructive methods for the measurement of moisture contents–a review,” *Sensor Review*, vol. 37, no. 1, pp. 71–77, 2017.

[13] H. Abdelnasser, M. Youssef, and K. A. Harras, “Wigest: A ubiquitous wifi-based gesture recognition system,” in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 1472–1480.

[14] X. Zheng, J. Wang, L. Shangguan *et al.*, “Smokey: Ubiquitous smoking detection with commercial wifi infrastructures,” in *Proc. IEEE Int. Conf. on Computer Communications (INFOCOM)*, Jul. 2016, pp. 1–9.

Development of a Self-guided Web-based Mental Healthcare Course for Day-To-Day Practice

Takeshi Kamita[*], Tatsuya Ito[*], Atsuko Matsumoto[**], Tsunetsugu Munakata[***], and Tomoo Inoue[****]

[*]Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
{s1730527, s1721654}@s.tsukuba.ac.jp
[**]Graduate School of Comprehensive Human Science, University of Tsukuba, Japan
s1130368@u.tsukuba.ac.jp
[***]SDS Corporation, Japan
munakata21@yahoo.co.jp
[****]Faculty of Library, Information and Media Science, University of Tsukuba, Japan
inoue@slis.tsukuba.ac.jp

Abstract - The need of mental health management in enterprises has become widely recognized. It is a big issue because of the huge number of employees, which is beyond the capacity of existing number of psychological experts such as medical doctors and counselors. Thus, self-guided method of mental healthcare is needed. A digital content of a self-guided mental healthcare course based on the counseling method SAT has been developed with the virtual reality technology so far. However, the virtual reality course was not so easy to use every day, although repetitive use of the course is desirable for keeping good mental health. Thus, we have developed the web-based self-guided mental healthcare course that can be practiced by a smartphone, by applying the same structure of the virtual reality course. The 14 days experimental use of the course in comparison with the conventional breath relaxation method is reported as well as the detailed design of the course.

Keywords: Self-guided mental healthcare, Healthcare course, SAT counseling method.

1 INTRODUCTION

Research of online courses on mental healthcare has become active, as the importance of keeping good mental health has been widely recognized. The stress check system to keep employees' good mental health in companies has even been legislated recently in Japan. This resulted in sudden increase of the number of potential clients for medical counseling most of whom are not in sick, whereas the number of industrial physicians specialized in psychology to carry out the counseling and to deal with employees' mental problems does not. Thus, at the same time as expanding the support system by experts, means to increase employees' own ability to conduct mental healthcare are more in demand.

A self-guided mental healthcare course[1] [2] based on a counseling and therapy method, the SAT (Structured Association Technique) method [3], adopting virtual reality (VR) were suggested as one of the means. The course which enables a user to carry out the therapy process for oneself with wearing a VR Head Mounted Display (HMD) obtained good stress relief evaluation.

In this research, we have developed a new web-based self-guided mental healthcare course (WEB course) that can be practiced by a popular device, a smartphone, using the same structure with the VR course, to improve the usability. It enables a user to repeat the course easier and acquire a self-care skill to cope with daily stresses so that the user can gradually stabilize the emotion for oneself and improve stress tolerance eventually. In order to examine the stress reduction effect of the WEB course, we conducted an evaluation experiment to compare the effect by one-time use of the WEB course and continuous use during 14 days with a breathing relaxation method.

2 RELATED WORK

In recent years, researches have been conducted to apply psychotherapy to digital contents and use them with mobile devices as a complementary tool for treatment and counseling, or a training tool. Researches on one of major psychotherapy, Cognitive Behavioral Therapy (CBT) have especially progressed [8] [9] [10], and there are many commercially available mobile applications [11]. In the CBT sessions, homework such as keeping a diary or changing a habit of behavior which is defined as an issue to be fixed through a session is often imposed on the client. In the next session, a counceller finds distortion of the clients cognition from the record of the diary and makes correction. Such the homework part of CBT is applied to digital contents and offered as mobile applications, however which does not cover whole CBT processes.

"Mood Mint" [12] was developed as a mobile training tool to reduce anxiety and depression with reference to the Cognitive Bias Modification [13], which attracted attention as a counseling technique.

Mindfulness-based Stress Reduction using meditation and Mindfulness Cognitive Therapy are also active in research and psychology clinic in Europe and the United States [15] [16]. As smartphones become more prevalent, mindfulness-based mobile applications [17][18] are attracting attention. Among them, a meditation voice guide service "Headspace" [19] that can be used in smartphone applications is commercially available. This application provides several courses for each purpose, such as coping with anxiety and depression, and assists the progress of meditation.

The above psychotherapy-based mobile applications are basically aimed at providing a training program and not practicing a therapy in one session. The purpose of our research is to realize a self-guided mental healthcare course through which a user can practice a therapy to reduce daily stresses and a training to acquire a skill to cope with such stresses.

3 SAT METHOD

3.1 Overview of SAT method

SAT counseling method is a structured and interview form counseling method proposed by Munakata. The SAT method has a wide effective range such as a mental disorder (such as Depression, bipolar disorder, obsessive compulsive disorder, personality disorder, schizophrenia, etc.) and various stress diseases. Unlike other conventional counseling methods focusing on the psychological aspects, the SAT method puts an importance on physicality, and approaches mental problems from the bodily symptoms. Therefore, instead of working thought by linguistic stimulus, use visual stimulus from the presented image. It is possible to grasp unconscious true feelings and an essential desire in a short time because it can functionalize an association and a flash and intuition well.

3.2 SAT Imagery Therapy using Light Image and Surrogate Face Representation Image

When a person who wants to have counselling recalls a stress scene, it is perceived as physical discomfort (such as Stomach shrinks, nervous, sweating hands, chest tightening). The SAT Imagery Therapy using Light image is a technique to change the discomfort to a good feeling and reduce the stress by watching the light image selected and perceived as a pleasant stimulus[5].

The SAT Imagery Therapy using Surrogate Face Representation image is a technique for transforming the image for self to a good one by replacing the primitive land-scape (for example scenery that many yell at around childhood) in the interpersonal relationship of the consultant with the image of surrogate face representation symbolizing pleasure. In psychology research, it is generally known that influence on self-esteem of a person is influenced by how the child care attitude of a child career is positive or negative. By allowing the person to select an image of surrogate face representation with a sense of good feeling and recalling the image of a scene that makes a sense of security and providing a feeling of security, person perceives a sense that is safely protected, enhances self-esteem and encourages stress reduction.

3.3 Self-Guided Course based on SAT Method

SAT counseling method does not need a client to tell his/her traumatic episodes or secrets and uses visual stimulation by images of light and positive face representation instead of nuanced linguistic expressions. Also, it is well structured which can be practiced in relatively short time in 5 to 10 minutes. But, in the conventional SAT Imagery Therapy, the expert evokes client's association through hearing, counseling or presenting thumbnails on paper media without images In some cases, the image is not sufficiently evoked by merely looking at the image on the paper medium, and the counselor has a supplementary voice call or encourages eyes to close to arouse the image while seeing the reactions such as the words and expression. Therefore, counselors play an important role in their progress. In this research, we created a course as a technique that can make self-progression even without counselor guidance support by converting SAT method to digital content and using a smartphone.

3.3.1 Course Composition

Based on the SAT method, the composition and procedures of WEB course are classified into three categories (Table 1): (1) knowing their own mental condition (Assessment Part), (2) stress reduction (Solution Part), (3) knowledge and training to improve mental resistance (Learning Part). In the learning part, based on the analysis of the data obtained in the assessment part and the solution part, learning contents suitable for individual stress characteristics are provided. In this research, the assessment part and the solution part are developed prior to the learning part and the stress relief effect of image therapy in the solution part are investegated.

Table 1. WEB course and SAT method content comparison

The course category	Content of WEB course	Corresponded contents of the SAT method
Assessment Part	To know mental conditions and characteristics.	The SAT psychological scale used for health coaching by a SAT therapist
Solution Part	To carry out a therapy process to reduce stress.	The SAT imagery work using Light Image and Surrogate Face Representation Image
Learning Part	To learn the methodology of the SAT method and understand the course.	Psycho education conducted by a SAT therapist.

3.3.2 Assessment Part

In the assessment part, we conducted a mental characteristic check test (Table 2) using the SAT four psychological scale with the aim of measuring the mental condition and characteristics of the user and clarifying the changes before and after the use of the system.

3.3.3 Solution Part

The solution department presents the set questions in order and is proceeded by the process that the user answers to reduce the stress. In the first half (Table 3), first, the user is asked to remember one of stress scenes in accordance with the question and make aware of how much stress it is. Next, by comparing the stress to color and form and making imagination as if the stress image compared by color and form are approaching the user oneself, the user is encouraged to perceive the physical discomfort. Furthermore, by specifying body part and type of the perceived physical discomfort, user is prompted to focus consciousness on the discomfort in the body. And then, by expressing the stress level caused by the discomfort as a numerical value (%), the user recognizes more clearly that the discomfort is occurred while feeling the stress.

In the second half, the course steps based on the process of SAT Imagery therapy using light image and surrogate face representation (Table 4) are proceeded to decrease the stress level%, in short, relieve the discomfort and reduce the stress.

Table 2. Mental characteristic check test

Scale	Content	Number of questions	Score range
State-Trait Anxiety Inventory(STAI)	It tends to fall into anxiety. The degree of vague anxiety that reflects the individual's past experience rather than state anxiety that changes over time [20]	20	20-80
Self-rating Depression Scale(SDS)	Evaluation of depression symptoms including mood, appetite, sleep [21]	20	20-80
Self-repression behavioral trait	Behavior characteristics that suppresses one's feelings and thoughts	10	0-20
Difficulty to feel emotion	Even if feel a painful thing, you tend not to be emotional, and tend to endure yourself	10	0-20

Table 3. First half of solution part

Order	Question
1	Please remind me again what you are concerned about now
2	What is it like? Please choose (Choose from 34 sources of stress such as your future, family health etc.)
3	How much is that degree? Please choose (Choose from 3 stages "not so" to "very much")
4	Does that stress comparable to color? (Choose from red, brown, black, gray, purple, navy blue, light blue)
5	If you compare the stress to the shape? (Square, rugosum, muddy, fluffy, pointed, flat, selected from spheres)
6	Close your eyes, thinking about where this thing comes and imagining this image, where do you feel strangeness in your body?
7	How is that strangeness? (Choose from throbbing, cold, heavy, dull, sore, tight, numb, stretch)
8	What is the stress level of current discomfort? (answer from 0%~100%)

Table 4. Second half of solution part

Order	Question
1	The part that feels that discomfort is healed by which color light is being protected?
2	Please choose a comfortable face that came into your eyes. Do you have anyone who smiled easily?
3	Looking at that face, what percentage of stress is the same as before? (Answer from 0%~100%)
4	What kind of character are you going to be when you see these people?
5	If such a personality, in the situation of stress, how can you handle it? It's okay with what you came up with intuition.
6	What do you think is the result if you do that?
7	Who is the most interested of those who have chosen?
8	What message will you give me?
9	How will you feel?
10	How did you feel about the stress that first came up when you were watching all the faces of these people?
11	How has the degree of stress changed? (Choose from 3 stages "not so" to "very much")

4 SELF-GUIDED MENTAL HEALTH CARE WITH SMARTPHONE

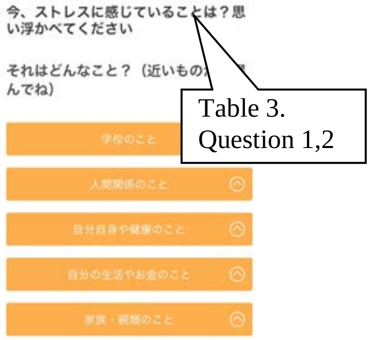
The WEB course using smartphones was developed in accordance with the composition of the digitized SAT method. It is constructed as a web site that can be realized with multiple platforms so as to flexibly respond to users' usage situations. Therefore, it can be accessed by using PC, smartphone, etc. In this research, we will describe contents assuming use on smartphones.

In the VR course, the immersive feeling image fits the SAT imagery therapy and can be as one of factors to bring effect on the stress reduction. However, such effect is not expected with the small flat display of the smartphone. On the other hand, from the viewpoint of operability, swiping and tapping operation on smartphone is easier and more familiar than moving the head to manipulate the cursor on the VR screen. In the SAT therapy process, it is desirable to intuitively perform selection operation in a short time rather than carefully selecting using long time. By using a smartphone, relatively complicated operations such as button selection, cancellation, page advancement and return can be performed more intuitively and quickly.

When the user logs in using the ID and password on the login page, a start page is displayed. The user select ether the button of the Assessment part or the Solution part.

The mental characteristic check test is displayed in the assessment part.

When the user select the button of the solution part, the page shown in Figure 4. In this screen, the user are



今、ストレスに感じていることは？思い浮かべてください

それはどんなこと？（近いもの

んでね）

Figure 4: The Stress source list screen

requested to recall the stress scene (Table 3, Question 1), and select from the list of prepared stress sources (such as things of own future, family health etc.) the one closer to the problem of the stress scene (Table 3, Question 3). Then choose from 3 options for the degree of stress. In the scene where stress is compared to color and shape (Table 3, Question 4), make a selection from the image list. Returning to the chat screen, while recognizing the color and shape of the selected color, perceiving physical discomfort and specifying the part and type (Table 3, Questions 6, 7) (Figure 5). Finally, answer by entering% of stress received by physical sense of discomfort.

Subsequently, questions are presented according to the latter half of the solution part (Table 4). The user is asked to select a light image (golden, green, peach, orange, blue, white, cream, yellow color, provided based on the light image of SAT method) (Table 4, Question 1) from image list. From the image list, after selecting the image that can make an image that will heal physical discomfort, we move to the screen and the selected image is displayed (Figure 6). Multiple images can be selected. After that, select a representative from the selected image and deepen the feeling of being protected by imagining speaking. Finally, it asked the user to answer how stress level against stress source confirmed in the first half has changed, and it ends.

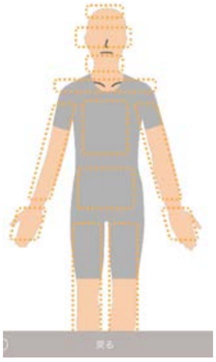


Figure 5: Identifying physical discomfort

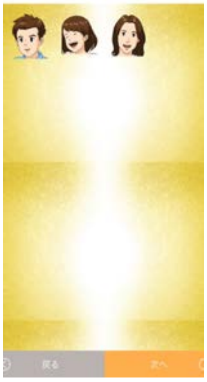


Figure 6: Selecting light image and face images

5 EXPERIMENT

In this research, the evaluation experiment was carried out with the approval of the ethics review committee in Faculty of Library Information and Media Science, University of Tsukuba (Notification No. 29-109)

In order to examine the stress reduction effect of one-time use and continuous use of WEB course and the difference of the effect, an evaluation experiment to have the subject continue to use this course for 14 days and to compare it with the breath relaxation method.

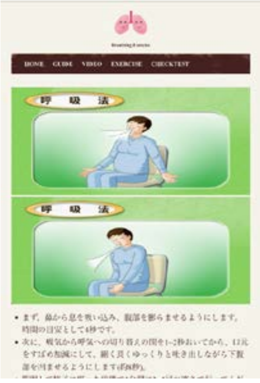
5.1 Breath Relaxation Method

We prepared a course to carry out the breath relaxation method as a control group.

The breath relaxation method is a training method aiming at improving the function of the mind and the body by breathing. It is introduced in the data of the Ministry of Education, Culture, Sports, Science and Technology in Japan [22] etc. as a relaxation method to consciously control breathing. It has also been reported that it is a technique that can be instructed safely and effectively in clinical practice [23]. Even a busy worker can do with a little time hanging on a chair and does not need special physical strength.

This method intends to increase the mobility of the diaphragm, and the respiratory movement that emphasizes the process of expiration is performed at a speed of 3-4 times per minute with the eyes closed and sitting on the chair [24][25]. First, take about 4 seconds, breath in through your nose and inflate your abdomen. Next, after 1-2 seconds between switching from inhalation to exhalation, pull out the lower abdomen while drawing slowly and slowly for about 8 seconds. In this experiment, according to this method, the subjects in the control group were asked to perform this method.

We created a website (BREATH course) to guide the experiment subject's breathing practice. This has a screen (Figure 7) for presenting the implementation method and a timer screen (Figure 8) for displaying the execution time.



- まず、鼻から息を吸い込み、腹部を膨らませるようにします。時間の目安として4秒です。
- 次に、吸気から呼気への切り替えの間を1-2秒おいてから、口元をすぼめ加減にして、細く長くゆっくりと吐き出しながら下腹部を凹ませるようにします(8秒)。

Figure 7: Breathing exercise practice guide screen

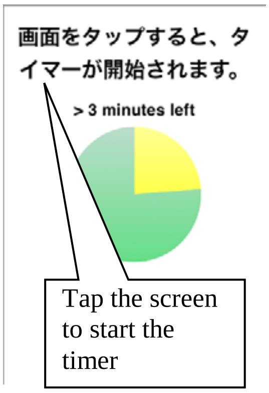

画面をタップすると、タイマーが開始されます。

Figure 8: 5 minutes timer screen

5.2 Procedure

33 college students and 7 office workers were selected as subjects, and randomly assigned to two groups, a WEB course conducted experiment group and a BREATH course running group (WEB course:N = 21, BREATH course:N = 19). For each group, after explaining the SAT method or breath relaxation method, the contents of the experiment and the course, we started to use the course and measured the state before the course by the mental characteristic check test of the assessment part. Next, we conducted the course at the solution part and asked the mental characteristics check test of the assessment department again after the course. This is the first day 's experiment.

After the second day, I instructed to carry out the course once a day. In addition, in order to confirm the implementation status of the course and the troubled points, we notified by e-mail every day until the 7th day after the start of the course and then notified two or three times.

The timing of the survey conducted mental characteristic check test using four psychological scales at 4 points before and after the first course, after the course on the seventh day and after the course on the 14th day. We used a

Table 5. Stress characteristic check scores (WEB course: N = 21、BREATH course: N = 19)

Scale	Course	One-time Before	One-time After	7th day	14th day
		Average ± SD	Average ± SD	Average ± SD	Average ± SD
STAI	WEB	42.95±9.233	41.85±11.03	38.00±9.94	36.88±9.52
	BREATH	37.68±6.57	33.84±6.91	37.94±7.01	33.67±8.51
SDS	WEB	32.48±13.04	33.05±11.26	31.53±7.08	29.50±7.67
	BREATH	31.16±7.37	30.37±4.97	28.53±6.27	27.83±6.44
Self-repression behavioral trait	WEB	10.00±2.67	9.85±3.30	8.88±1.50	8.75±2.08
	BREATH	9.79±3.33	10.21±3.88	10.35±3.84	9.22±3.74
Difficulty to feel emotion	WEB	9.24±2.74	9.40±3.32	9.60±3.07	8.00±2.63
	BREATH	8.63±4.19	9.26±5.30	9.12±4.09	8.94±4.28

questionnaire on the paper medium for the first experiment. On the seventh day and the 14th day, we used the search function on a smartphone.The stress situation difference was evaluated using the stress characteristic check test (Table 2). Four psychological measures (State-Trait Anxiety Inventory, Self-rating Depression Scale, Self-repression behavioral trait, Difficulty to feel emotion) used in the usual SAT method were used.

6 RESULT

The result of stress characteristic check test is shown on Table 5. The average score of STAI scale and Self-repression behavioral trait scale in the WEB course and SDS scale in the BREATH course are gradually decreasing step by step during the experiment period. The others are once increased, but decreased at 14th day eventually. The changes in these average scores suggest that the stress relieving effect was brought and accumulated in 14 days in each course. This result shall be tested and analyzed in detail in the future.

7 CONCLUSION

We developed a self-mental healthcare course based on the SAT counseling method aiming at long-term and continuously available self-mental healthcare tool, using a simplicity of smartphone and intuitive operation of web-based self-mental care course. In this research, we conducted an experiment to have the subjects continue to use the course over 14 days and examined the stress relieving effect of one-time use and continuous use. As a result, it is suggested that the stress relieving effect by using the course accumulated in 14 days. In the future, the result shall be tested and analyzed in detail to verify the effect of the course.

REFERENCES

[1]Asuki Nakanishi, et al.: A VR Self Mental Healthcare System by SAT-Based Meth-od,IPSJ SIG Technical Report,vol,2017-DCC-15, no.35, pp.1-8 (2017).

[2] Tatsuya Ito, et al.: A Self-guided Mental Healthcare Digital Content for Smartphone VR Based on the Counseling Technique SAT Method, IPSJ SIG Technical Report, vol.2018-DCC-18, no.37, 1-8 (2018).

[3] Tsunetsugu Munakata.: SAT therapy, KANEKOSHOBO, Japan, (2006).

[4]Tsunetsugu Munakata: The applicability of the simple edition of SAT method in promoting universal health, Journal of Health Counseling, 17, pp. 1-12 (2011).

[5]Tsunetsugu Munakata.: Does SAT Re-scripting Expression Imagery Enable Us to Overcome Un-endurable Hardships toward True Life Career ?, Journal of Health Counseling,15,pp.1-12 (2009).

[6] Yoshihumi Tsuji: Development of a self-image focusing coping strategy aimed at improving psychological competitive skills, PhD Thesis.Universitiy of Tsukuba,Japan, (2011).

[7]Kelley, Christina, Bongshin Lee, and Lauren Wilcox: Self-Tracking for Mental Well-ness: Understanding Expert Perspectives and Student Experiences, Proceedings of the SIGCHI conference on human factors in computing systems, CHI Conference 2017, pp. 629-641.

[8]P. J. Batterham, A.L. Calear, L. Farrer, S.M. McCallum, and V.W.S. Cheng: FitMindKit: Randomised controlled trial of an automatically tailored online program for mood, anxiety, substance use and suicidality, Internet Interventions, Vol12, pp 91-99, (2018).

[9]K.H.Ly, E. Jannib, R.Wrede, M.Sedem, T. Donker, P.Carlbring, and G. Andersson: Experiences of a guided smartphone-based behavioral activation therapy for depression: A qualitative study, Internet Interventions, Vol.2, Issue1, pp.60-68, (2015).

[10]D. Bakker, N. Rickard: Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism, Journal of Affective Disorders, Vol.227, pp.432-442, (2018).

[11]J. Torous, M.E. Levin, D.K. Ahern, and M.L. Oser: Cognitive Behavioral Mobile Applications: Clinical Studies, Marketplace Overview, and Research Agenda, Cognitive and Behavioral Practice, Vol.24, Issue 2, pp.215-225, May (2017).

[12]“Mood Mint”, http://www.biasmodification.com/, (view 2017-11-16).

[13]Wang Rui, et al,StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones, Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp3-14, (2014).

[14]F.B. Dickerson, et al. : The token economy for schizophrenia: review of the literature and recommendations for future research, Schizophrenia Research, vol.75, pp.405-416, (2005).

[15] Jon Kabat-Z: An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results, General Hospital Psychiatry, Vol. 4, Issue 1, pp.33-47, (1982).

[16] Rinske A. G. et al. : Standardised Mindfulness-Based Interventions in Healthcare: An Overview of Systematic Reviews and Meta-Analyses of RCTs, PLOS ONE, (2015).

[17]Inmaculada Plaza, Marcelo Marcos Piva Demarzo, Paola Herrera-Mercadal,and Javier García-Campayo: Mindfulness-based mobile applications: Literature review and analysis of current features. JMIR mHealth and uHealth 1, 2, e24 (2013). DOI: http://dx. doi. org/10. 2196/mhealth. 2733

[18]“Diana Wells: The Best Meditation Apps of 2016.”, http://www. healthline. com/health/mental- health/top- meditation- iphone-android- apps, (view 2016-07-18).

[19] “Headspace”, https://www.headspace.com/, (view 2018-06-01).

[20]Spielberger, C.D. : STAI manual, Palo Alto, Calif, Consulting Psychologist Press, (1970).

[21]Zung, W.K. K. : A self-rating depression scale, Archives of general psychiatry, 12, pp.63-70, (1965).

[22]"Safety measure materials for overseas educational institutions", Ministry of Education, Culture, Sports, Science and Technology Home page, http://www.mext.go.jp/a_menu/shotou/clarinet/002/003/010/004.htm (view 2017-10-14).

[23]Kikuyo Kosakabashi: Aiming to incorporate the relaxation method into clinical nursing - Research, education and practice of relaxation method as nursing intervention - , KMJ The Kitakanto Medical Journal, 65, pp.1-10, (2015).

[24]Setsuo Arai: Examination of effective posture for diaphragmatic respiration. Physical education science, 41, pp.813-817, (1991).

[25]“Useful for nursing care【3 volumes】”, https://www.igakueizou.co.jp/product/product_detail.php?product_code=NK (view 2018-01-08)

Healthcare Promotion System by Stimulating In-Group Contribution Mind and Inter-Group Competitiveness

Yuki Aso[†], Nobuyuki Ito[‡], Katsuhiro Naito*, Naoya Chujo*, Tadanori Mizuno*, Katsuhiko Kaji*

[†]Graduate School of Business Administration and Computer Science, Aichi Institute of Technology, Japan
[‡]Mitsubishi Electric Engineering Company Limited, Japan
* Department of Information Science, Aichi Institute of Technology, Japan
b18701bb@aitech.ac.jp

Abstract - Deskworker is a lack of exercise because there are few opportunities for exercise. Lack of exercise is a primary cause of most chronic diseases. On the other hand there have been many studies of health services combined with smart devices, gathering a user's health lifelog and managing his or her health for the improve. The purpose of this study is to add group elements to health services and promote exercise. Because the group elements can stimulate competitiveness and contribution mind, these will help to promote exercise. Other studies have concluded that we changed the behavior of the team based on life log technology and information sharing and promoted exercise[1]. The model with "competition" technique resulted the most effective performance for competitive teams such as sport teams. However, in this study it is inconvenient to require manual input by the user, and the presentation method is inadequate. In this paper, in order to automate acquisition of exercise amount and maintain continuous motivation, create a healthcare promotion system (Fig. 1) incorporating competitiveness and contribution mind. As a method of automating gathering of user's activity amount, use smartphone application and API provided by application. To improve competitiveness, compare with other groups using group evaluation values. Group members are ordered in descending order of exercise amount, ranking advantage is used to present group contribution. We did create and operate a prototype to find the best presentation method. Measure the number of steps of the person belonging to the four laboratory as the activity amount. For grouping, it is between laboratories, grade, between transportation, between favorite sweets etc. Number of people belonging to the group, the group evaluation value changes. Therefore, it is necessary to consider other calculation methods. It is insufficient in terms of privacy protection to present names, step counts and action rankings within the group. Therefore, the user should indicate that your position is in the top of the group. Information presentation using activity amount is inappropriate for maintaining motivation. Therefore, regular motivation is necessary. We implemented an event function aimed at regular motivation. After the event period and period, there was an increase or decrease of the activity amount of the user. From these results, we discussed the necessity of regular motivation.

Keywords: healthcare, group activity, contribution mind, competitiveness

1 INTRODUCTION

Deskworker is a lack of exercise because there are few opportunities for exercise [1]. Lack of exercise is a primary cause of most chronic diseases [2]. It is common for companies to grasp the health situation of employees and give exercise instruction to improve [3]. However, because that method is insufficient as a precautionary measure, companies need effective health management [4].

I believe that improving motivation has the effect of encouraging behavior leading to exercise. Ingress[1] and Pokémon GO[2], game portals are set for real world buildings and monuments. Therefore, in order to advance the game advantageously, there is a characteristic that it has to actually move to that place. For example, Pokémon GO got about 5 billion kilometers[3], about 10 billion km in 10 monthst[4]. Even for those who are not interested in healthcare, the number of steps in everyday life is increasing naturally in order to favorably perform these position information games. These location information games can be said to be examples that showed the possibility of improving motivation and transforming human behavior if the location information service is properly designed. In this case, the user is acting to lead to exercise based on the motivation to enjoy the game.

On the other hand there have been many studies of healtah services combined with smart devices, gathering a user health lifelog and managing his or her health for the improvement of the quality of his or her life, using various sensors. The purpose of this study was to add group elements to health services and promote exercise. Because the group elements can stimulate competitiveness and contribution mind, these will help to promote exercise.

2 RELATED RESEARCH

According to "Draft Motion Standards for Health Promotion" announced by the Ministry of Health, Labor and Welfare in January 2006, walking of 8,000 steps to 10,000 steps per day is considered good for prevention of lifestyle diseases [5]. In addition, excessive exercise, on the contrary, becomes stress, there is also a risk of lowering immune function [6].

[1] Ingress:https://www.ingress.com/

[2] Pokémon GO:http://www.pokemongo. For location information games for smartphones such as jp/

[3] https://pokemongo.nianticlabs.com/en/post/headsup

[4] Pokémon GO - Walk and go for adventure! http://pokemongolive.com/en/post/adventureweek2017

Therefore, it can be said that about 8,000 steps per day is a suitable exercise amount. In this research, the objective is to promote health by exercise focusing on walking. Therefore, it is necessary to introduce appropriate momentum based on medical grounds as a motion index in the system.

It is difficult to realize behavioral change by compulsion from outside, and sustainability is not seen. Meanwhile, behavioral changes realized based on endogenous motivation are sustained as long as their motivation is maintained. Therefore, promotion of healthcare of office workers is not necessarily merely realizable momentarily, but sustainability is a problem.

Human Based Computation's research will be helpful in considering stimulatable motivation. Human Based Computation is an approach in which people are asked to perform tasks that are difficult to automate by computers, and there are cases in which various motivations are stimulated. Here are the main motivations. I want to get results, monetary compensation, stimulation different from everyday life, aesthetic appetite, curiosity, volunteer spirit, contribution heart, competitive spirit, communication desire, desire to share knowledge, want to achieve what others can not do, game, fun is there. As an example, re-CAPCHA creates an account of a web system anew, or a mechanism for letting a person who is trying to download a certain file to input two words and judge whether it is a person or a BOT [7]. At this time, one of the two is a word that failed OCR in the digitization project of the old book. Behavior based on motivation to obtain results such as accounts and files, and behind the scenes contribute to digitalization of old books by human power.

To improve the motivation of exercise, a method using reward can be considered. An example using a game is BitHunters [8]. BitHunters is a real world game for smartphones who can enjoy adventure walking in the real world. Through walking, the user can acquire tokens issued in the game. This token can be exchanged for a virtual currency like a bit coin in the future. Therefore, the user can expect improvement in exercise motivation for reward.

At the Osaka Minami Power Station, Kansai Electric Power Co., Ltd. conducted group activities to improve lifestyle habits, and achieved [9]. As for the health improvement measures before group formation, the target person felt ”feeling being carried away”, and the expected effect was not obtained. As a breakthrough policy, we organized the group and practiced health improvement small group activities to formulate and practice health improvement measures by the members of the group themselves. Through this policy, we shared objective consciousness within the group and expected improvement of health with mutual sense of solidarity. In the first year of the group organization, it divided into three groups of ”Good Group”, ”Group that needs a little hard work”, ”Group that got worse” from the trend of past examination figures. Based on the advice of industrial physicians, the group discussed targets and all members worked on health examination to be carried out three months later. Of the 27 subjects, 7 people succeeded in health improvement, 11 people deteriorated more than before. For this grouping, because the cause varies for each subject, it is difficult to share the purpose consciousness. In order to facilitate the sharing of target consciousness among subjects within the group, we organized the ”eradication high-fat team”, ”resurrected liver function team”, ”overthrowing hypertension team” for each cause item. We set goals and take countermeasures in the same way as last time. Thirteen out of 25 participants succeeded in improving health. Five out of 13 people achieved all goals. From this result, grouping that is easy to give a common purpose consciousness is important.

Fitbit [10] which carries an acceleration sensor and an altitude sensor and measures the amount of activity is linked with wireless application of smartphone, and data management and the amount of activity can be visualized. Within the application, there is a function to share the momentum of friends and family members and make them compete. By utilizing this function, it can be expected to stimulate user's motivation.

The paper by Nishiyama [11] have concluded that They changed the behavior of the team based on life log technology and information sharing and promoted exercise. The model with ”competition” technique resulted the most effective performance for competitive teams such as sport teams. However, in this study it is inconvenient to require manual input by the user, and the presentation method is inadequate In this paper, in order to automate acquisition of exercise amount and maintain continuous motivation, create a presentation system incorporating competitiveness and contribution mind.

3 HEALTHCARE PROMOTION SYSTEM BY STIMULATING IN-GROUP CONTRIBUTION MIND AND INTER-GROUP COMPETITIVENESS

In this research, we focus on contribution and competitiveness. These focuses are very good for the combination of concepts of groups. It stimulates the feelings that contribute to the group as a member and the feelings that you do not want to lose to other groups. Improve motivation of exercise by doing these appropriately. Therefore, we focused on the group because we thought that maintaining motivation for exercise was easier when exercising as a group as members of arbitrary group than when doing voluntary exercise by individuals. When grouping office workers, there are groups such as departments and companies. It is also conceivable to form groups for each user profile such as gender, age, commutation method, favorite confectionery, and the like. To measure the amount of activity, use a smartphone or a wearable device that is considered to be carried around by an office worker. We use applications suitable for each device and collect and manage data. Visually display the amount of activity, based on the collected data. Users can switch, display among favorite groups.

3.1 Activity Amount Acquisition Method

In order to examine the presentation method effective for improving motivation for exercise, we have constructed an activity quantity presentation system among multiple groups.

Fig. 1 is a diagram of the entire system. We utilized existing field service and automated acquisition of momentum.

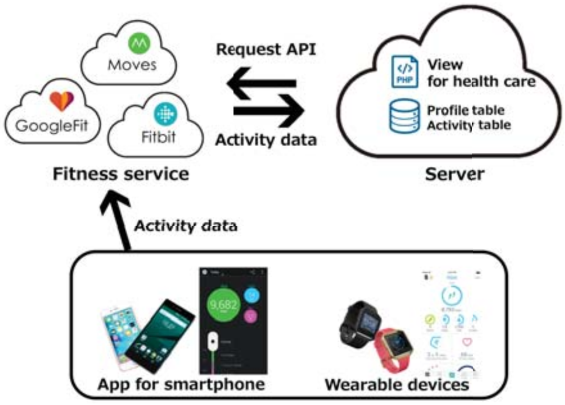
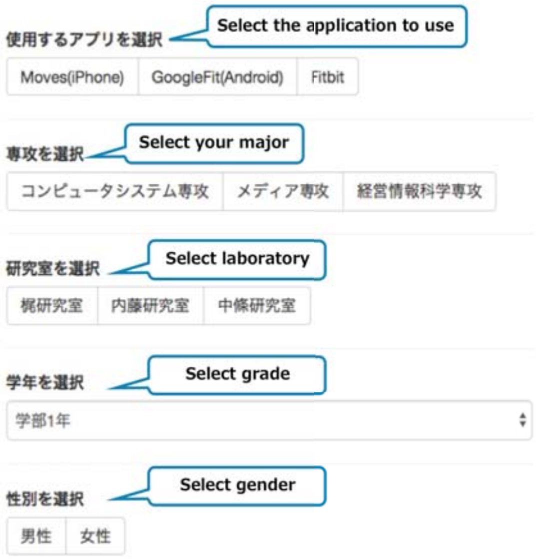


Figure 1: Automatic activity record system

As a location information service to acquire activity amount, use ”Moves” [12] on iPhone or ”Google Fit” [13] on Android device. Fitbit [10] which is one of wearable device installs and uses a dedicated application. People using the system shall always carry one of these.When the user uses this system for the first time, profile creation and registration of access tokens for exercise amount acquisition are performed. We created a questionnaire page to have users use their device attributes and their own attributes. The answer items are the device used, the major, the laboratory, the grade, the sex, the height, the weight, the way of commuting, the presence or absence of a circle of athletic meetings and cultures, Kinoko no Yama or Takenoko no Sato. When the input is completed and proceeding to the next step, the application authentication method suitable for the used Device is performed and the registration is completed. At this time, the web server side obtains the access token of the application and starts collecting the data of the exercise amount. In order to present information tailored to each individual, an authentication page for each person was created. For registration, ID and password authentication is common. To improve convenience, we implemented a mechanism called open ID to log in using an existing Google account. The user can save time of new registration and the system management side has advantages of simplifying user management and improving both sides.



学年を選択: 学部1年

Figure 2: User registration

3.2 Implementation of Activity Quantity Visualization System

Although the main object of this research is an office worker, information university students are also common in that there are many work sitting at the desk, We designed a system for university students. We grouped them by laboratory and constructed a system to graph the momentum in each group (Fig. 3).

Furthermore, grouping by user profile was done as a device to attract users interest. As an example of more subdivided grouping, there is a graphicization of activity amount of grade (Fig. 4). As an example of grouping according to hobby taste, there is a graphicization of activity amount of each favorite sweets (Fig. 5).

Fig. 3, 4, 5 graph the average number of steps per group for each day. When comparing the amount of activity, using the cumulative value of the activity amount, the number of people belonging to the group is different, and the small group is disadvantageous because it is disadvantageous. For this reason, we are currently comparing using the group’s average activity level.

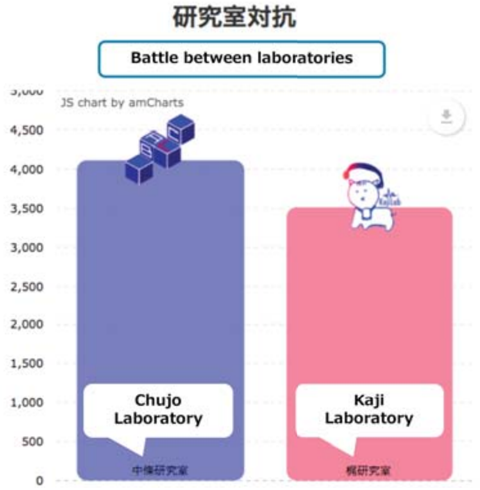


Figure 3: Amount of activity between laboratories

4 DISUCUSSION FOR OPERATION

We tested the proposed system in a short time. Also, we discussed the display method and handling of data. In this chapter, discussion items for practical application will be discussed.

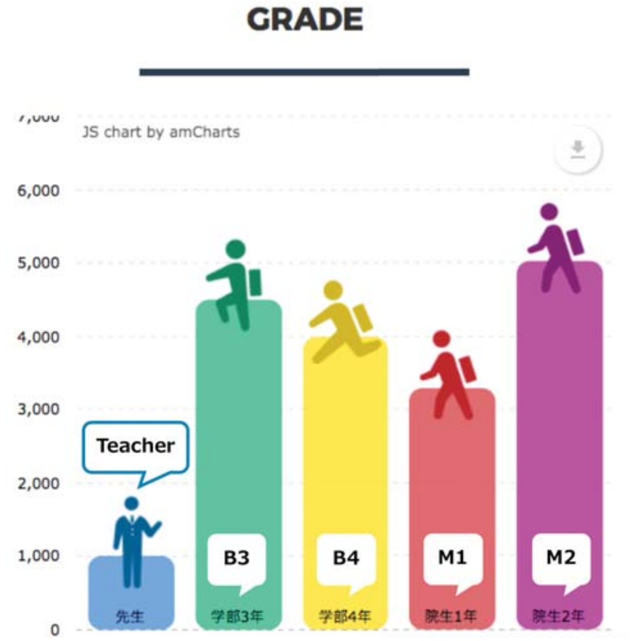


Figure 4: Activity amount between grade

4.1 Numericalization of Activity Amount

Although the average value of steps is used as the current activity amount, the influence per capita changes according to the number of belonging people (Fig. 6). We focus on the transition of the user's cumulative number of steps per day. As a method, evaluate when the user continuously achieves the goal or starts to work hard. The user who achieved the goal after increasing more than the usual step number. Evaluate these users (Fig. 7). Person A has achieved the target index for four days.Since this user is accomplished by daily effort, it is necessary to evaluate the contribution to the group. influence is weak in group B. For example, even if they all have the same momentum, if there is a difference in the number of people, there will be a difference in influence per person. For that purpose we are planning to use the combination of the variance of the activity amount of the group and the average value.

We focus on the transition of the user's cumulative number of steps per day. As a method, evaluate when the user continuously achieves the goal or starts to work hard. The user who achieved the goal after increasing more than the usual step number. Evaluate these user's (Fig. 7). Person A has achieved the target index for four days. Since this user is accomplished by daily effort, it is necessary to evaluate the contribution to the group. For Person C, the cumulative number of steps until 1st to 3rd day is low on average. However, on the fourth day, the figures showed an increase over the cumulative number of daily steps, achieving the target index. In this case, although it is impossible to read the intention of the individual himself, it can be inferred that we have started to achieve the target index. In order to help maintain motivation, it is necessary to evaluate the beginning of efforts to exercise.

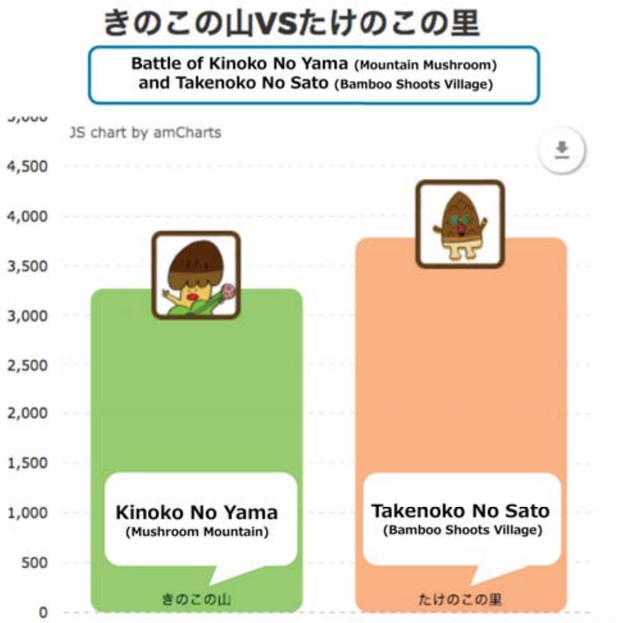


Figure 5: Activity amount by favorite confectionery

4.2 Reviewing grouping

Prior to visualization of the amount of activity, it is necessary to first classify groups without unfairness. As shown in the (Fig. 8), because there is a difference in the amount of activity due to differences in commuting methods, we are planning to use the amount of activity during the working hours for evaluation. However, since the administrator can acquire the activity amount of a specific user within an arbitrary time, there is a possibility that problems related to privacy may emerge from the viewpoint of behavior guessing and the like. Also, after using the system for a certain period of time, I plan to reorganize the group into a group more suitable for healthcare. For example, it is conceivable that people with less activity from various groups are gathered, and the user can be given a common conscience that "they must exercise".

4.3 Relationship between in-group contribution and privacy protection

To display the degree of contribution within the group, it is displayed using names and attributes, but it is considered to be insufficient from the viewpoint of privacy protection. For example, when names and steps are used (Fig. 9) to indicate the degree of contribution within a group, the recognition of the degree of hard work between members increases, and improvement of motivation within the group can be expected. However, for members with low group contribution, there is a possibility that exercise is forced from members with high competitive spirit or high contribution.

From the viewpoint of privacy, display that can specify the name and attribute of a member should be avoided. For example, comparing figures showing the amount of activity and degree of obesity (Fig. 10) and the figure of contribution within the group, there is a danger that the degree of obesity and name will be inferred. In general, obesity degree is informa-

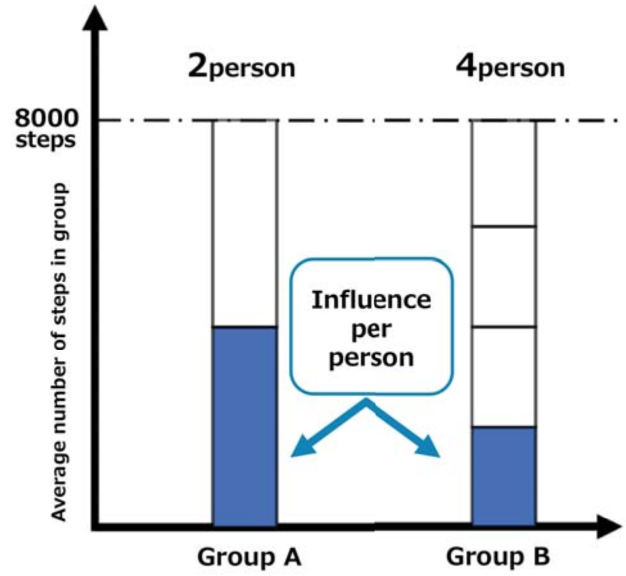


Figure 6: Evaluation changes depending on the number of members

tion that you do not want to know, so privacy protection is necessary.

Reducing the sensing acquisition interval increases real-time performance. However, privacy protection is insufficient from the viewpoint of user's behavioral inference. It is necessary to set the minimum necessary interval for presentation and consider privacy protection.

4.4 Examination of Information Presentation Method

It is difficult to obtain sense of accomplishment, superiority, and sense of crisis simply by presenting numerical values such as the number of steps. As a method to realize it, we are studying the following information presentation. For example, if you use Mt. Fuji as a motif, you can grasp the position between the groups and visually obtain a sense of accomplishment and superiority (Fig. 11). Also, display the activity level of each group in the same way, and measure the contribution degree. Next, if you use a balloon diving game as a motif, the amount of activity is proportional to the size of the balloon, and if the average value in the group exceeds 8000 steps, the balloon will break (Fig. 12). The upper group shows the number of steps required to achieve the goal and gives another group a feeling of crisis. Also, to small balloon group, present the number of steps to overcome and lead to behavioral change. To maximize motivation, it is easier to tackle if the way to overcome other groups is clear and simpler. We can build a strategy of winning with a little effort and think that if the effect gets better in a short period of time, it will lead to continuation of motivation. In order to have sustainability, it is necessary to devise not to get tired with a display method rich in variation. For example, giving each group a character, metabolizing or slimming the style makes it possible to obtain an evaluation of the image of each group. Also, comparing rice cultivation to games, encourage continued hard work. Daily efforts are required to harvest rice, and the achievement of rice growth and harvest can be obtained. Because rice plants dying when you skip along the way, you can expect participants to return to their original goals.

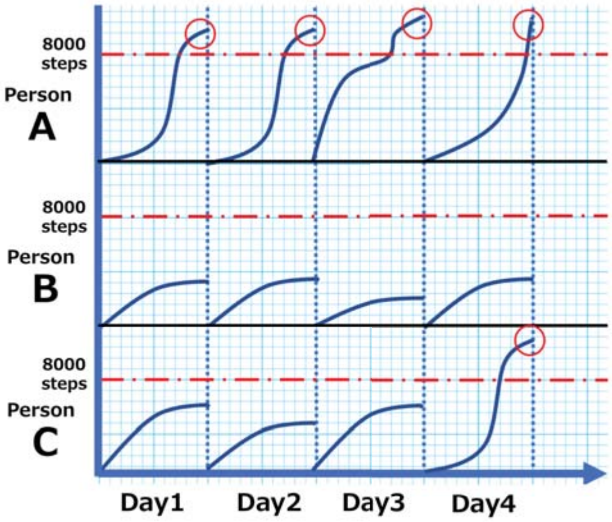


Figure 7: Activity comparison for each user

4.5 Ddeviceevice Difference of Acquisition Activity Amount

Although this system can acquire the activity amount on the smartphone or wearable device, there is a difference in the amount of activity acquired by the device. A smartphone is always carried by a user as compared with a wearable device worn on the body. Therefore, there is a difference in the amount of activity obtained by the device. Also, since the sensor accuracy built in iPhone, Android, Fitbit is different, there is difference in activity amount You can think of possibility. In the future, when the difference in activity amount due to accuracy becomes an adverse effect of comparison between groups, Or a method of comparing between groups within the same device may be considered. As an acquisition method other than smart phones, there is an employee ID card with embedded acceleration sensor [14]. It is possible to measure the number of steps from built-in sensors. Since the employee ID card is always carried by the office worker inside the company, it is expected that the capture rate of the momentum measurement can be improved more than the smartphone or the wearable device.

5 IMPLEMENTATION FOR ACTUAL OPERATION

Based on the discussion in Chapter 4, we propose individual and group evaluation method and presentation method. Evaluate the ongoing efforts and measures when evaluating. For this reason, composite evaluation values including step count are converted to points and evaluated (Fig. 16). In addition, as a presentation method, an event period is set for the purpose of evaluating ongoing efforts. In this chapter, we

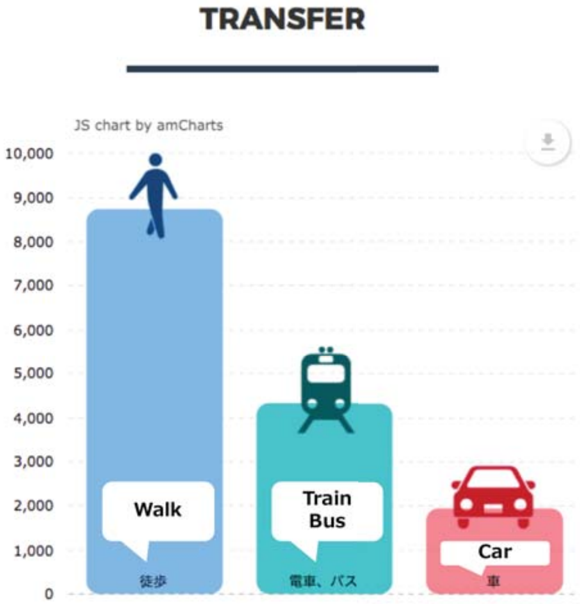
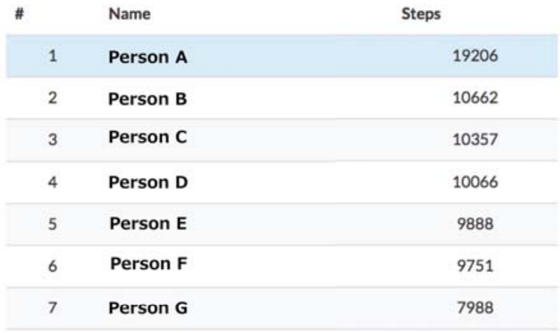


Figure 8: Activity amount by transportation method



#	Name	Steps
1	Person A	19206
2	Person B	10662
3	Person C	10357
4	Person D	10066
5	Person E	9888
6	Person F	9751
7	Person G	7988

Figure 9: Activity amount ranking

describe these proposals and implementations for evaluation experiments.

5.1 Implementation of Person Points

We propose an person points calculation method which is necessary for group points calculation. For person points, include bonus points to evaluate ongoing efforts. Person points (P_p) are defined as follows.

$$P_p = P_{basic} + P_{bonus}$$

Define the basic points (P_{basic}) for converting the number of steps to a point as follows (Fig. 14). Define the minimum target that is easy to achieve to 2000 steps and the maximum target to 8000 steps. Acquired points vary from 2 points to 10 points between 2000 and 8000 steps. 0 point if the minimum target has not been reached. It is fixed at 10 points with the number of steps exceeding the maximum target. The upper limit of the acquisition points is to prevent the immune function deterioration due to excessive exercise [6] and to have a certain degree of convergence for comparison between groups.

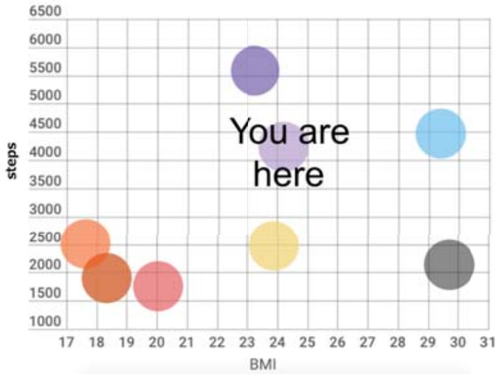


Figure 10: Correlation diagram of activity and body mass index

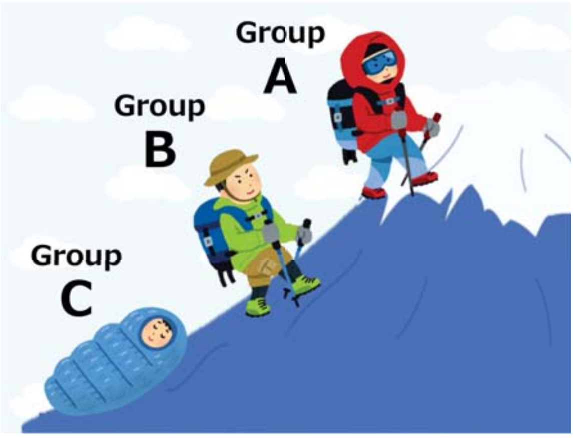


Figure 11: Visualization of activity amount with motif of Mount Fuji

Based on the discussion in Chapter 4, we propose individual and group evaluation method and presentation method. As an evaluation method, in order to evaluate the daily effort and the beginning of the initiative, a point which is a composite evaluation value including the step count is defined and used.

In order to evaluate the user 's ongoing effort, group points (P_{bonus}) are defined below.

$$P_{bonus} = P_{basic}(C_{it} - 1)scale$$

We set an intermediate target of 5000 steps between the maximum target (8000 steps) and the minimum target (2000 steps). The achieved number of intermediate targets is C_{it}.Bonus points will increase in proportion to the number attained the intermediate target.

$$scale = \frac{E_{bonus}}{\frac{1}{2} \times (E_{days} - 1) \times E_{days} \times max(P_{basic})}$$

Here, E_{days} is the number of days of the event. The scale is for setting the cumulative upper limit (E_{bonus}) of the bonus points. By achieving the continuous mid-term goal, points can be greatly acquired at the end of the event. By doing this,

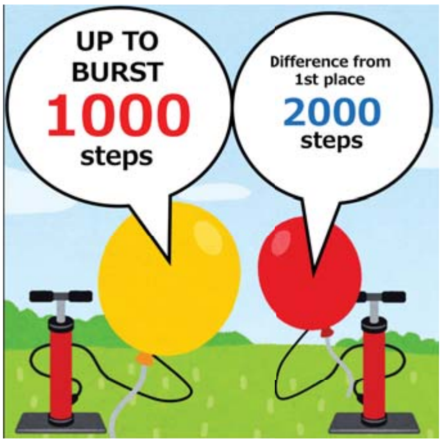


Figure 12: Presentdeviceing information for winning other groups

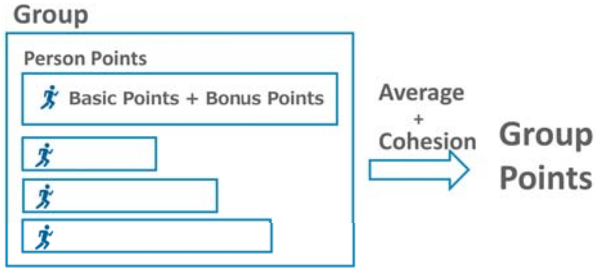


Figure 13: Point Calculation Schematic

it is possible to give appropriate points as compared to those who have not achieved target.

Fig. 15 is a graph of points obtained when walking 8000 steps or more each day. Earn daily basic points 10 and earn 70 points. Bonus points earned 30 points total.

5.2 Implementation of Group Points

Based on the discussion in Section 4.1, we implemented calculation of group points, including variations in individual points calculated in the previous section. For group points, we want to make the highest evaluation when the variation of group members is small like (Fig. 16) and the average value of individual points is high. This will better appreciate the group that everyone is doing their best. Therefore, group points (G_p) are defined as follows.

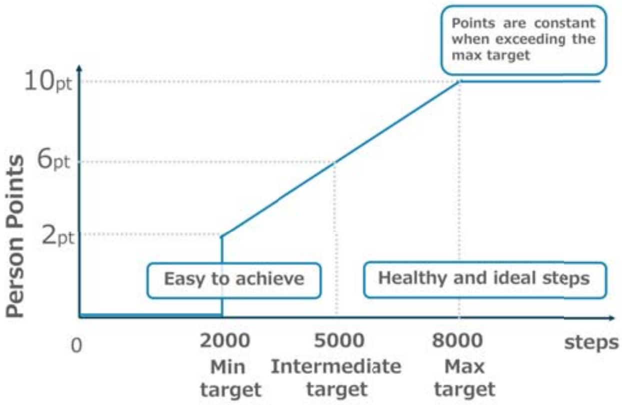


Figure 14: Basic Points

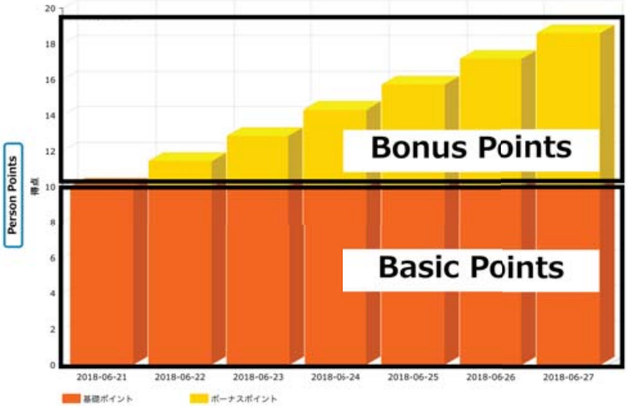


Figure 15: Person points when you continue to walk 8000 steps or more every day

$$G_p = \begin{cases} \mu + w\mu^{\frac{(\sigma_{max}-\sigma)}{\sigma}} & (\sigma \geq 1) \\ \mu + w\mu\sigma_{max} & (\sigma < 1) \end{cases} \quad (1)$$

Here, μ is the average of individual acquisition points of group members. σ is the standard deviation of individual acquisition points of group members.σ_{max} is the maximum possible value of standard deviation. w is a weight for how much variation is to be evaluated with respect to the average value.

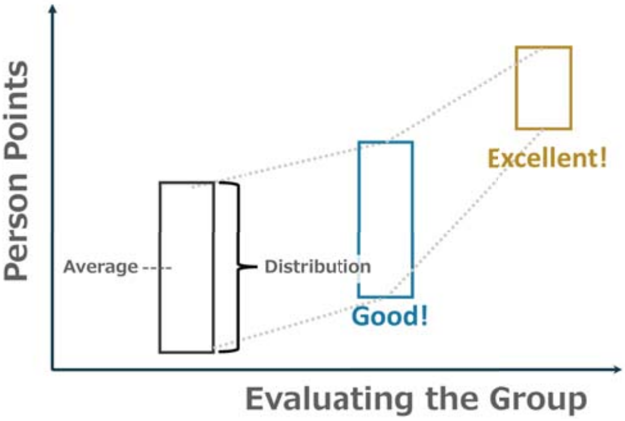


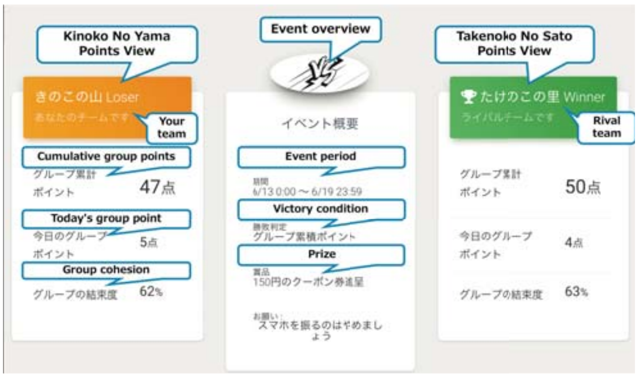
Figure 16: Approximate drawing about points

5.3 Information Presentation Method

In the upper part of the page, based on the discussion in chapter 4.4, in order to maximize motivation, we made an indication that improves competition among the groups (Fig. 17, 18).

Display group information on the left and right and implement a comparable screen. For the group cohesion degree, the standard deviation 0 is displayed as 100%, and the standard deviation maximum value is calculated as 0%, and the result is displayed. Also, group points are displayed in a time series in a line graph.

At the bottom of the page, we present information on individual acquisition points(Fig. 19, 20). Individual information indicates the degree of contribution to an individual group and individual acquisition points. To indicate contribution degree, what is the top. For privacy protection mentioned in the discussion in section 4.3, do not display ranking.



Kinoko No Yama Points View

きのこの山 Loser

あなたのチームです

Your team

Cumulative group points

グループ累計ポイント 47点

Today's group point

今日のグループポイント 5点

Group cohesion

グループの結束度 62%

Event overview

イベント概要

Event period

期間
6/13 0:00～6/19 23:59

Victory condition

勝敗判定
グループ累積ポイント

Prize

賞品
150円のクーポン券進呈

お願い：
スマホを振るのはやめましょう

Takenoko No Sato Points View

たけのこの里 Winner

ライバルチームです

Rival team

グループ累計ポイント 50点

今日のグループポイント 4点

グループの結束度 63%

Figure 17: Cumulative points and cohesion degrees of each group

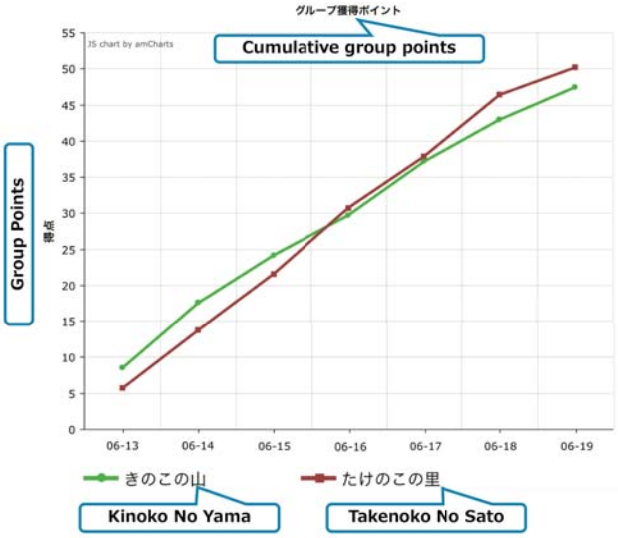


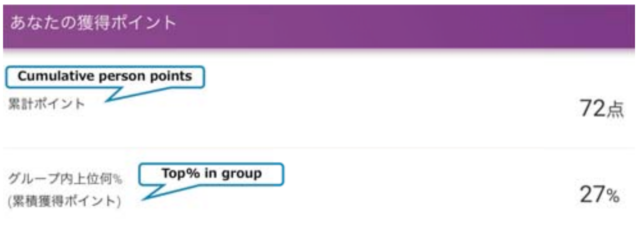
Figure 18: Group points graph

6 EVALUATION EXPERIMENT

In order to verify whether the motivation can actually be improved by the proposed method, we conducted an evaluation experiment in the real environment. The evaluation item is the presence or absence of increase in momentum.

6.1 Experiment Setting

Week 1 from June 13 to June 19 was the event week, followed by a week outside the event for a total of two weeks as the experiment period. The subject was 23 people, Aichi Institute of Technology Kaji Lab. As an evaluation method, increase / decrease of the average exercise amount for each subject is confirmed at the event week and after the event. As information presentation to the user, we conduct web page opening, signage display (Fig. 21) in the laboratory where participating members gather, and occasional distribution (Fig. 22) using communication tools. This is because it is important to confirm your position in the group one after another.



あなたの獲得ポイント

Cumulative person points

累計ポイント 72点

グループ内上位何%
(累積獲得ポイント)

Top% in group

27%

Figure 19: User points view

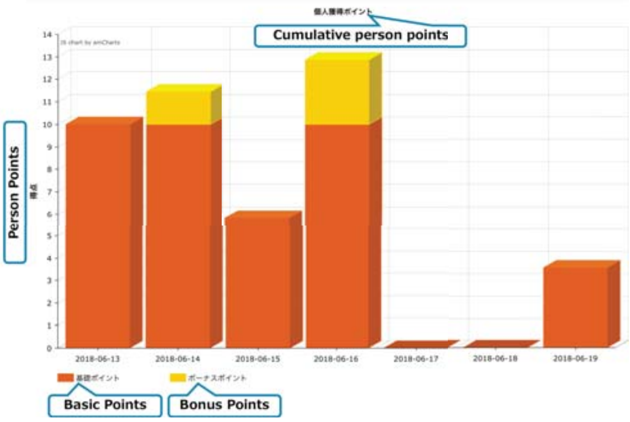


Figure 20: User points graph

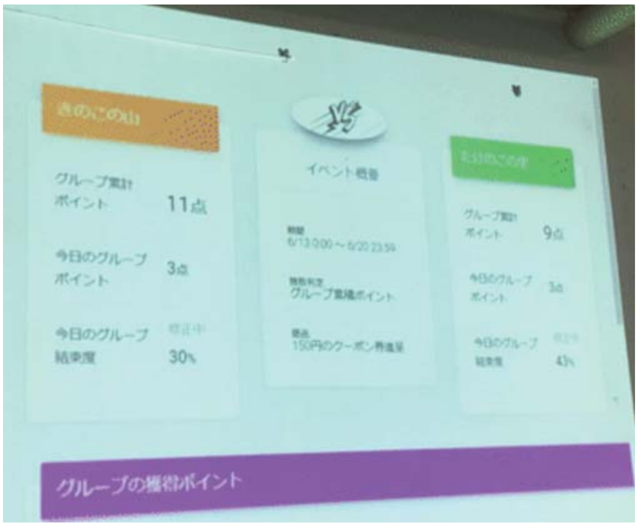
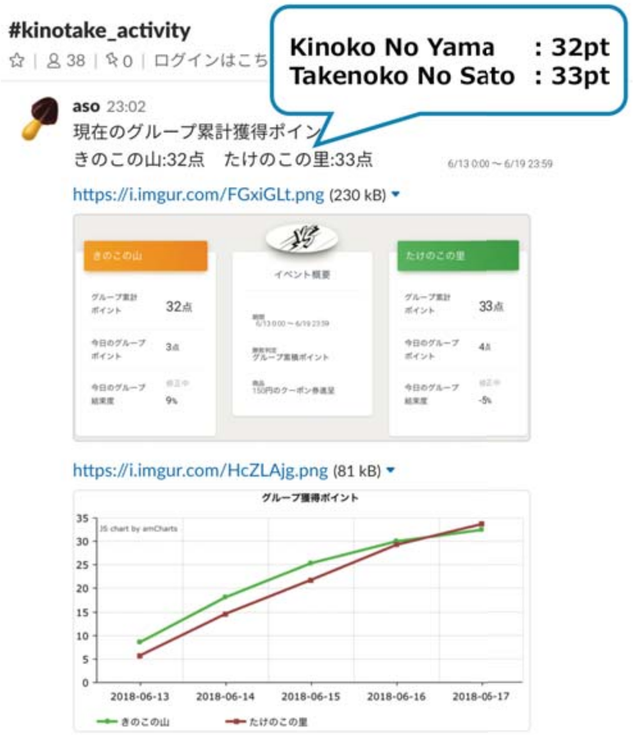


Figure 21: Information presentation in the laboratory



#kinotake_activity

☆ | 38 | 0 | ログインはこち

aso 23:02

現在のグループ累計獲得ポイン

きのこの山:32点　たけのこの里:33点

https://i.imgur.com/FGxiGLt.png (230 kB)

https://i.imgur.com/HcZLAjg.png (81 kB)

Figure 22: Distribution of interim progress at any time by communication tool

6.2 Experimental Results

Experimental results are shown in Fig. 17 and Fig. 18. During the event period, Takenoko No Sato was victorious with 47 points of Kinoko No Yama and 50 points of Takenoko No Sato. On May 13th of the event start date, Kinoko No Yama was dominant, but on May 16th there was a reversal. Also, the result of the exercise comparison between the event week from June 13th to June 19th and the subsequent week is shown in the Table 1. We recorded data even after the event ended. The week when the open campus was included was a good record.

Table 1: Average value per day during the event and after the event

Period	Average Daily Steps
Befor the event (6/20-6/27)	5559
Event (6/13-6/19)	5643
After the event (6/20-6/27)	5818
Include Open Campus (7/18-7/24)	6477

6.3 Discussion

Experimental results show that the amount of changes before and after the event are small. There is a marked increase in open campus compared to events. This makes it clear that temporary increase in activity amount was not confirmed. In addition about the open campus, we should take into consideration the fact that there is a duty to participate. Therefore, exercise by duty is more effective than exercise by motivation.

7 CONCLUSION

The purpose of this research is to build a visualization system of activity amount among multiple groups and to improve motivation to the movement of people belonging to each group. As a preliminary step for actual operation, we examined the method of improving motivation, building a visualization system. These systems are currently being implemented. Through actual operation for a long time, we are planning to promote research to make appropriate motivation stimulation and lead to continuous behavior change.

REFERENCES

[1] Ministry of Health, Labor and Welfare, Exercise Guidelines for Health Promotion 2006, <http://www.mhlw.go.jp/bunya/kenkou/undou01/pdf/data.pdf> (2006) (in Japanese).

[2] Ministry of Health, Labor and Welfare, Study Group Report on Revision of Exercise Standards and Guidelines, <https://www.mhlw.go.jp/stf/houdou/2r9852000002xple-att/2r9852000002xpqt.pdf> (2013) (in Japanese).

[3] Ministry of Health, Labor and Welfare, e-healthnet, Instruction Regarding Exercise and Physical Activities for Specific Medical Check-ups and Health Guidance, <https://www.e-healthnet.mhlw.go.jp/information/exercise/s-01-003.html/> (accessed 2018-7-24) (in Japanese).

[4] Workshop for the Management of Health on Company and Employee, What is Management of Health, <http://kenkokeiei.jp/whats/> (accessed 2018-7-24) (in Japanese).

[5] Ministry of Health, Labor and Welfare, Exercise Standards for Health Promotion 2006, <http://www.mhlw.go.jp/bunya/kenkou/undou02/pdf/data.pdf> (2006) (in Japanese).

[6] S. Shinkai, Y. Aoyagi, T. Suzuki, Activity Life Expectancy and Walking Ability of Senior Citizens, Research journal of walking (4), pp. 15–21 (2000) (in Japanese).

[7] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, reCAPTCHA: Human-Based Character Recognition via Web Security Measures, Science, Vol. 321, lssue 5895, pp. 1465-1468 (2008).

[8] Real World Games, BitHunters, <http://bithunters.jp/> (accessed 2018-7-24) (in Japanese).

[9] Y. Shirai, Health improvement activities by mini group, Safety and Health, Vol. 11, No. 3, pp. 294–296 (2010) (in Japanese).

[10] Fitbit Inc. : Fitbit, <http://www.fitbit.com/us> (accessed 2018-7-24).

[11] Y. Nishiyama, T. Okoshi, T. Yonezawa, J. Nakazawa, K. Takashio, and H. Tokuda, Encouraging Team Behavior Modification Using Life Log, The Special Interest Group Technical Reports of IPSJ, Vol. 2014-MBL-70, No. 9, pp. 1–8 (2014) (in Japanese).

[12] Moves, <https://moves-app.com> (accessed 2018-7-22).

[13] Google Fit, <https://www.google.com/fit/> (accessed 2018-7-24).

[14] T. Watanabe, M. Murase, K. Naito, N. Ito, K. Kaji, N. Chujo, and T. Mizuno, Prototype Implementation of RFID Based Health Management System with Low-power ARM Microcomputer, Consumer Communications & Networking Conference (CCNC), 2018 15th IEEE Annual, pp. 1–2 (2018).

Keynote Speech 2: Dr. Shigetoshi Sameshima (Deputy General Manager, Center for Technology Innovation / General Manager, Yokohama Research Laboratory, Hitachi,Ltd., Research & Development Group)

Digital transformation towards Society 5.0

Shigetoshi Sameshima
Hitachi,Ltd., Research & Development Group

Abstract:
Japan is sometimes referred to as "an advanced country with advanced issues" as we are the first among advanced nations to face the challenges of an increasingly aging population with a low-birthrate. Given this situation, both government and industry have a vested interested in resolving the various challenges that arise. The government of Japan has launched the Society 5.0 vision in "the 5th Science and Technology Basic Plan" on January 2016. Society 5.0 aims to realize a new social structure to match the needs of individual citizens by tackling societal challenges and surfing the digitalization wave. We will head for the active society promoting autonomous value creation to solve various problems and create value by wisdoms. Also, from a technology perspective, Japan has been contributing to various projects towards Smart Society. Hitachi solves challenges with IoT digital solutions. For example, to realize optimized transportation, we developed train traffic management system which automatically changes train schedule in accordance with the number of passengers. In this way, Hitachi realizes the fine-grained service for consumers and improves their Quality of Life with the IoT platform. Furthermore, Hitachi will co-create solutions with partners all over the world by creating wisdoms from high-quality IoT data, and by implementing the value in the society to build a new value chain.

Joined Hitachi, Ltd in 1993. He engaged in R&D of system technologies for public infrastructure systems like railway operation system, utility operation system, water supply system, especially in developing autonomous decentralized systems. Now he is directing Center for Technology Innovation in Systems Engineering and Production Engineering fields. Received Ph.D from University of Tokyo.

Session 7:
Trust and Risk

(Chair: Tomoki Yoshihisa)

A Study of Risk Management Method using Quantitative Process Management

Akihiro Hayashi[†]**, Nobuhiro Kataoka[‡], Yasunobu Kino*, Mikio Aoyama**

[†]Department of Information Design, Shizuoka Institute of Science and Technology, Japan
[‡]Interprise Laboratory, Japan
* Graduate School of Business Science, The University of Tsukuba, Japan
** Graduate School of Information Science and Engineering, Nanzan University, Japan
pixysbrain@gmail.com

Abstract - Risk management (RM) aims to lead system development projects to success by eliminating negative factors. It is expected that the number of failed projects can be reduced in organizations that have introduced RM successfully. However, we cannot confirm this expected result yet. In this study, we first analyze the RM process for the so-called successful system development projects conducted recently and then identify the factors where RM did not meet the expected criteria. Next, to eliminate the factors, we propose a quantitative RM method that could improve the RM process and project management (PM) process by using Earned Value Management (EVM) and Logistic Regression Analysis (LRA). When we apply the proposed method to a real system development projects, we concluded that the proposed method is effective.

Keywords: Project Risk Management, Logistic Regression Analysis,Quantitative Project Management

1 INTRODUCTION

”Software dependent Society” has arrived. Important functions such as organizational operation, home electronics, and automobile control are controlled using software. Therefore, many companies focus on system development.

However, according to reliable statistical information[1], only 27% of all projects succeed in all aspects of quality, cost, and delivery time (QCD) in domestic and foreign system development projects. Thus, three-fourths of the projects do not meet all the QCD criteria, which leads to the cancellation of 24% of software development projects[2].

To solve this problem, interest in introducing RM processes in system development has increased. It is believed that RM has the ability to lead projects to success by eliminating negative factors that may cause the project to fail.

In general, to introduce RM processes, an international “best practice model” has adopted the Project Management Body of Knowledge (PMBOK)[3], Program & Project Management for Enterprise Innovation (P2M) [4], the 2nd version of Projects in Controlled Environments (PRINCE2) [5] as reference models, and introduced specific practices for RM that all presented in these guides.

However, even the introduction of the “best practice process,” does not reduce the number of failed projects. The findings suggest that successful implementation of the RM process does not contribute toward the reduction of failed project occurrence.

We believe that there are two major factors that contribute to the event described above. One factor is that the project management standards and guides that have been proposed and developed overseas do not match practices that are in place in the domestic system development projects. The other factor is that even though the standards and guides are correct, they have not been successfully introduced to the system development site.

In both cases, it is necessary to establish an appropriate methodology to introduce RM processes to the management of standard-sized projects in Japanese industry.

To address this issue and decrease the incidence of failed projects, we first analyze four cases of a specific risk management process conducted recently and identify the factors that will create a bottleneck. Next, to solve the problem of failure, we proposed a method to introduce the RM process appropriately. The proposed method included quantitative risk management and the implementation of risk countermeasures. When we applied our proposed method to a real case for the RM of system development projects, a measurable effect was observed in the form of a reduction in the number of failed projects and a reduction in the contingency budget.

In most of the earlier research on risk management of system development, Boehm[6] and Williams el al[7] initially introduced the implementation methods of risk management practices that are commonly used presently, such as risk identification, evaluation, classification, and prioritization. When such a method is generally perceived, risk management process is also adopted widely for project management standards and guides such as PMBOK, P2M, PRINCE2 and process evaluation models such as ISO9001[8] and CMMI (Capability Maturity Model Integration)[9], Since then, improvement practice in each life cycle of risk management, risk identification using risk breakdown structure[10] and risk assessment method[11] have been reported. In addition, with the advancement of risk management technology, quantitative risk analysis[12] and the Six Sigma approach[13] have been proposed and carried out. Although such prior research explains the potential of success for the risk management process, it does not describe why such best practice does not successfully operate in the real world. To the best of our knowledge, no particular prior research discusses the appropriate method to introduce risk management process.

The remainder of this paper is organized as follows: In Section 2, we review the related works on PM, EVM and RM to confirm the originality of this study. In Section 3, we analyze a case on the implemented RM practice where we consult and identify the reason behind the failure of RM, and describe the problem to be solved. In Section 4, we propose a quantitative RM method by using a statistical tool called LRA. In Section 5, the effectiveness of the proposal is evaluated by applying the method to a real case of system development. In Section 6, we discuss the results of the case study. Finally, Section 7 presents the conclusion.

2 FACTORS CONTRIBUTING TO THE FAILURE OF RM

2.1 Case Analysis of RM Process

First, we aim to clarify the reasons due to which the number of failed projects has not decreased, even after best practice of the successful introduction of the RM process in organizations. We analyze the factors by taking up four organizations for whom we have provided management consultation to date. Below is a summary of the organizations to be analyzed:

Case 1 Electronic control of vehicle amenity
Case 2 Electronic notebook, which maintains a schedule, dictionary, calculator, and custom program
Case 3 Air conditioner system controlled by an internet-based remote control
Case 4 Derivative development of value-added of acoustic measurement calibration equipment

By analyzing the 4 aforementioned development cases, we found the following four problems in the RM process:

Problem 1 Since the triggering of alarm for the notification of risk was delayed by the project manager (PM), risk countermeasures could not be implemented on time. A similar problem occurred multiple times in a PM's tenure; he thought that the problem could be solved every time. Therefore, he did not report the emergence of risk to the higher-level managers.

Problem 2 Since the development project was originally planned the development period, it was biased toward keeping the delivery date. Therefore, the organizations were averse toward reworking due to RM activities and hence hesitated to report risk occurrence.

Problem 3 Despite the original plan to activate risk response measures in an event-driven manner, project members did not accurately understand the RM process and subsequently reported risks at weekly progress meetings that caused notification delays.

Problem 4 Despite a clear definition of the trigger and threshold for risk interpretation in the RM ledger, the roles and responsibilities were misunderstood, and the risks were not reported correctly.

The four problems listed above indicated that the RM practice was correct, but that risk countermeasure actions were not implemented on time.

2.2 Factor Analysis to show that RM was not Implemented on Time

Next, we analyze factors that contributed toward the delayed occurrence of RM, extracted from the documents and minutes of the meeting of the past RM assessment. As a result, the following four factors were clarified:

Factor 1 There was no timely and correct presentation of a risk report:
Even when the risks expanded and severe delays did not allow adequate management, organizations sometimes reported less risk or kept critical risks hidden since the project was at a stage wherein it would be evaluated by higher-level managers. Therefore, a correct risk report was not delivered on time. This is the cause of Problem 1.

Factor 2 Insufficient risk judgment skill of the PM:
Due to the PM's insufficient risk judgment skill, such as insufficient identification of risk and undistinguished critical risk, the organization failed to manage the risk properly. This led to Problems 1, 3, and 4.

Factor 3 Divided RM method:
Due to insufficient communication between the PM and higher-level managers, unclear terms and methods were used, and insufficient information was presented, which contributed to Problems 2 and 3.

Factor 4 Inadequate risk visualization:
The RM ledger included the PM's subjective evaluations. Higher-level managers were unable to monitor risk situations of the projects, which contributed to the Problem 4.

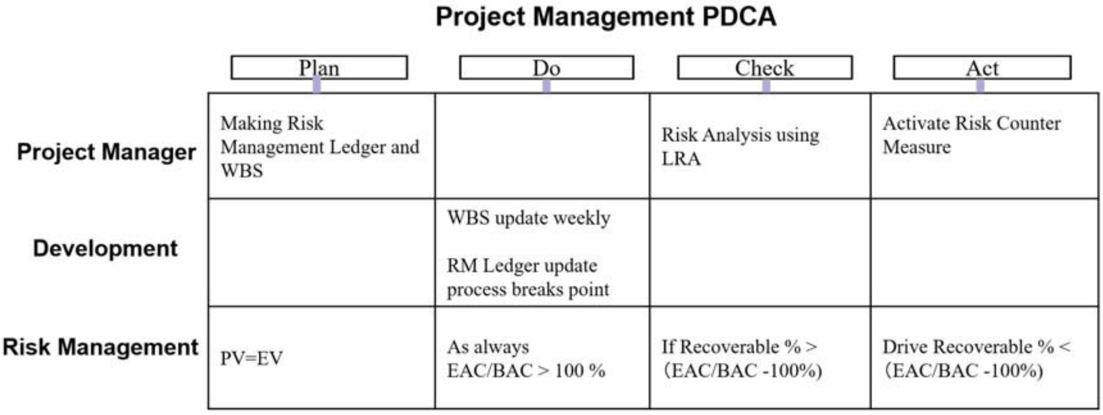
3 RM METHOD USING QUANTITATIVE PROCESS MANAGEMENT

3.1 Basic policy

A risk is a potentiality thing, and it does not necessarily become explicit. Therefore, the activation of risk countermeasures ahead of schedule will lead to the employment of unnecessary labor and costs.

For establishing a good RM process, it is important to define the risk of each project and judge them objectively by using quantitative data.

(1) Set a clear risk criterion
Objective judgment criteria are set for practicing each



Project Management PDCA

	Plan	Do	Check	Act
Project Manager	Making Risk Management Ledger and WBS		Risk Analysis using LRA	Activate Risk Counter Measure
Development		WBS update weekly RM Ledger update process breaks point		
Risk Management	PV=EV	As always EAC/BAC > 100 %	If Recoverable % > (EAC/BAC -100%)	Drive Recoverable % < (EAC/BAC -100%)

Figure 1: Procedure of this proposal

RM process, such as registering in the RM ledger, identifying risk explicitly, and using the quantification method.

(2) Process Performance Baseline (PPB)
With an emphasis on historical project data, not focusing on each project database, but the PPB focus on all historical projects' data accumulated.

(3) Evaluating the entire project status using a statistical method
The individual risk threshold is not evaluated, but whole project threshold is evaluated using a statistical method.

(4) Introduction of subject matter expert (SME) and quality assurance (QA): To solve the problem of skill shortage related to RM, an SME, a specialist of the development process who belongs to the engineering field, and a QA are introduced to discuss risks activities.

(5) Alignment Aligning the basic policies of (1) to (4) with the RM process.

3.2 Procedure

In order to solve the issue concerning a delay in the RM, which was identified in section 2.2, we will introduce the following RM procedures: making RM ledger and WBS, quantitative progress management using EVM, risk analysis using LRA, and risk counter strategy. These RM Procedures are shown in Figure 1. Hereafter, we will explain the RM procedures in detail.

3.2.1 Making RM Ledger and WBS

In Stage 1, we first make RM ledger and WBS. In the system development project, risks of the entire project are identified at the project planning stage. Mainly, risks are obtained as a result of awareness created during the process of analyzing customer need, creating customer requirement definition, creating project planning, and reviewing the QA Document.

The obtained risks are listed in the RM ledger of the project, and the properties of each risk are set. These properties include risk description, probability, impact, rework time, risk response strategy, the threshold of action taken, risk countermeasures, and priority. Priority is that the magnitude of the risk influence is sorted by order.

At the project planning stage, we also make WBS. WBS is a key project deliverable that organizes the team's work into manageable sections by hierarchical decomposition of the work to be executed by the project. At the project planning stage, we review all the tasks and set start and end dates of each task and efforts that are to be spent on the task.

When we identify all the tasks and make the WBS of a project, we automatically know the EVM value of the current status because the planned value (PV) and budget at completion (BAC) are calculated entirely.

In the RM process, the influence of risk is converted to "time" or "money." In this study, we convert the influence to "time" (minutes). The project stakeholders can understand the amount of influence quantitatively.

3.2.2 Quantitative Progress Management using EVM

In Stage 2, the project is managed on a weekly basis. Usually, a progress management meeting is conducted on a weekly basis. The project manager asks project members in charge to update WBS for the concerned week at the meeting. Subsequently, the actual value compared to the estimated value achieved in the project is known for a particular week. At the meeting, project members report the manager's current progress status by calculating these EVM values and problems that occurred, if any. This enables the managers to estimate the project's total cost at project completion, which is also referred to as estimate at completion (EAC)

In system development, by using a waterfall model, it has been empirically found that the risk is often explicit at process breaks. Therefore, at the progress meeting that is held at the end of the process, project members and SME conduct a risk review meeting. They reevaluate the risks according to the change of the environment at that time and update the RM ledger.

3.2.3 Risk analysis using LRA

At stage 3, we can predict whether the project will fail in the future, by using LRA. LRA is a statistical method that predicts the occurrence probability of an event from the size of accumulated data.

When we use a risk value as an explanatory variable, value that can take only a binary response (Yes / No), like the occurrence of project failure as a dependent variable, the probability of the influence on the occurrence of the failed project can be determined.

According to the basic policy (3), we decided that the whole project risk, instead of individual risk, should be set as the progress management threshold. Therefore, we decided that EAC/BAC should be set as the criteria for evaluating a project's success or failure. When the EAC/BAC exceed the specific trigger, the project manager should take appropriate action under the RM.

Organizations that conduct system development projects often have similar degrees of difficulty and similar scales. We have created repositories of PPB by accumulating project data over the past several years. We can also quantify the recoverable period for each construction period if the project is delayed.

For example, it is empirically known, "If you are projecting for 18 months, you will recover and meet the delivery date if the progress delay is less than 5% of the entire project period." We could calculate the recoverable period for each project by employing the LRA. It was found that if the period exceeds the value of (EAC/BAC- 100%), then it will lead to a delay in the delivery date. Subsequently, it will be necessary to take countermeasures that will not make the risk manifest in the schedule of recoverable limits.

3.2.4 Risk Counter Strategy

In stage 4, the project manager monitors the risk status at a progress meeting and takes appropriate risk actions if necessary. If the result of LRA exceeds the threshold, then the project manager can compensate for the project delay by immediately activating risk counter measures according to the priority measures set in the RM ledger until the EAC/BAC comes below the value EAC/BAC-100%. Then let the next PDCA cycle start.

4 EVALUATION OF THE APPLICATION IN ACTUAL DEVELOPMENT ORGANIZATION

In Section 4, a case study wherein the proposed RM method is applied at Company A and its effectiveness is evaluated. For the application of LRA, JMP®14 (SAS Institute Inc., Cary, NC, USA), which operates on Windows PC, was used.

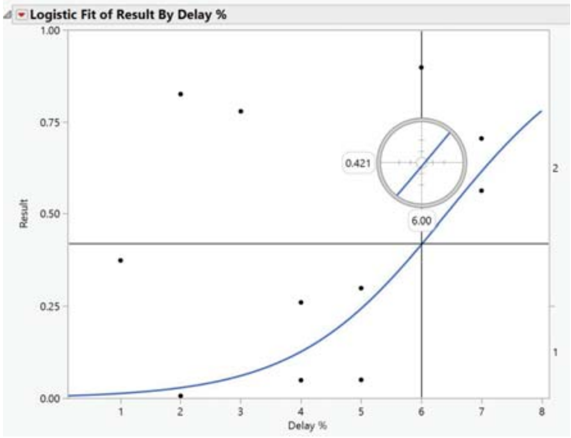
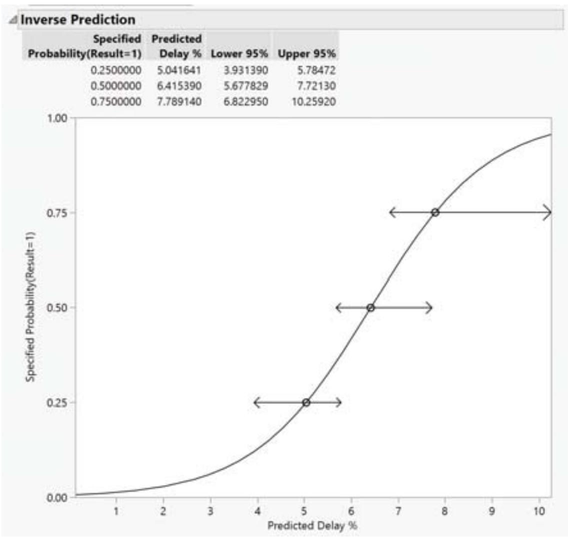


Figure 2: The Causes of Failure and its Proportion


Inverse Prediction

Specified Probability(Result=1)	Predicted Delay %	Lower 95%	Upper 95%
0.2500000	5.041641	3.931390	5.78472
0.5000000	6.415390	5.677829	7.72130
0.7500000	7.789140	6.822950	10.25920

Figure 3: LRA and Inverse Estimation

4.1 Case Study for Embedded System Development

Company A introduced project management using PMBOK for about 10 years. Development projects comprising the basic operation of the project management using the PDCA cycle, is well established.

However, in reality, even though it was called the RM process, its focus was on creating a risk matrix and completion record for the preparation of assessment evidence. This is a situation that does not lead to effective RM.

Company A set the development process standard, based on the waterfall model that considered the entire organization. Additionally, at the time of the introduction of the PMBOK, the RM plan, the creation of the RM ledger, and the risk assessment checklist were managed within the company. The company mainly undertook derived development. The company focuses on delivering in a timely manner in a short cycle. In the cases where the delivery is delayed among QCDs, the project is labeled as a "failed project." The policy of the organization is to prevent delayed delivery.

Company A's customers do not present their requirements clearly. Therefore, the project manager in Company A must analyze customer needs and make a feasible completion plan, and formulate the budget and set the construction duration accordingly.

Subsequently, to implement the requirement definition, the company undertakes project planning and conducts a QA review for implementing the requirement definition and project plan.

After a discussion with the PM, SME, QA, and project members, the company rules out the related risks and prepares the RM ledger based on the risks.

The project manager holds a progress meeting on a weekly basis. The project manager asks all the project members in charge to update the WBS. Subsequently, the project manager calculate the projected EAC and EAC/BAC. Then they update the RM ledger with latest data and entire EVM value is converted into "time."

The development period at Company A is set at around 6-18 months, depending on the scale of development. We calculate the probability of project failure by using the LRA.

For example, when a project construction term is 18 months, the predicted probability of failure would be as shown in Figure 2. The cross-hair tool indicates that the prediction probability is 0.421 at the time of 6% delay. In other words, if the project is delayed by 6%, then the delivery date will be delayed with a probability of 42%.

A 42% probability is difficult to employ as a psychological milestone to activate risk countermeasures. A value at which the prediction probability becomes 50% was calculated using an inverse estimation of LRA, and it was 6.41%. Conversely, it means that "If the project is delayed by 6.41%, then there can be a 50% chance of a recovery." This state is shown in Figure 3.

Actually, if the recoverable range exceeded the threshold (like 6.41%), they activated risk countermeasures to reduce the project delay in descending order of influence until the value fell below the threshold to ensure that the value fit within the recoverable range.

4.2 Evaluation of Effectiveness

Figure 4 shows the three-year trend of delayed project delivery in Company A. Although the duration for which the project delays were observed is small, the number of projects subjected to delivery time delay has definitely been reduced. Meanwhile, Company A did not introduce any other measures during this time, but the proposed method has been introduced. Company A considered the transition shown in Figure 4 which this can be regarded as an improvement that is achieved through the proposed method.

Figure 5 shows the transition of contingency for 3 years, after the introduction of the proposed method in Company A. Contingency is a reserve expenditure fund that can be drawn on to prevent project settlement deficit. It is preferable not to use contingency funds because it is recorded as a profit if not used. In the first year, after the introduction of our proposed method, we consumed nearly 30% of the contingency. It was suppressed to 20% or less in the third year. Since countermeasures are given priority in the order of the risk of damage due to anticipated risk, we believe that it contributed to the prevention of major deficit in projects. We confirmed that the proposed method also improves cost.

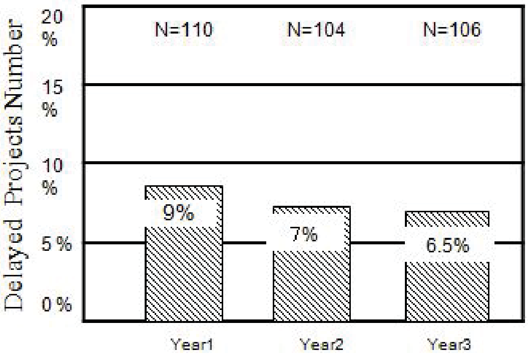


Figure 4: No. of Delivery Delayed Projects

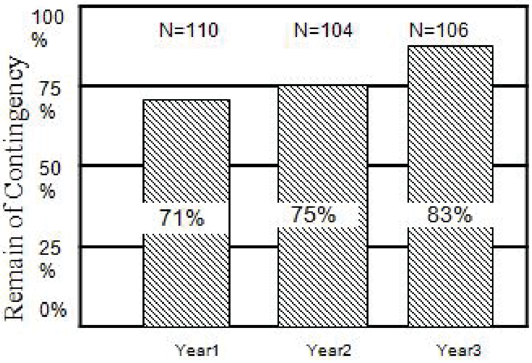


Figure 5: Transition of Contingency Budget

5 DISCUSSION

In this section, we discuss whether the factors presented in section 2.2 have been resolved.

5.1 Timely and Correct Risk Report

In this study, project risks at each process break are identified at the project planning stage and reviewed by PM, SME, and QA during the RM meeting. The RM ledger and WBS are updated. Subsequently, we obtain the EVM value and objectively calculate appropriate parameters for the project. At these meetings, the overall project risk and progress status are reported in a timely manner. The project manager can judge the status promptly.

Thus, nobody can reduce the number of risks or hide critical risks while reporting them; the risks are reported in an accurate and timely manner, thereby resolving the issue of delayed and erroneous risk reports.

5.2 PM's Risk Judgment Skill

In this study, to identify and judge risks, an SME and a QA reviewer are assigned to the project. Instead of a project manager, they support and perform RM processes, including risk identification and the activation of risk countermeasure.

In addition, risk is evaluated objectively by using parametric data to avoid the biases of project managers that might result from their assumed expertise over all the technical fields. It implies the resolution of the issues concerning skill shortage in project managers and their lack of risk-judgement skill.

5.3 RM Method in an Organization

In this study, the organization's historical project data are accumulated through the PPB. Recoverable range of each project period calculated by LRA method are accumulated.

Project managers or higher-level managers can judge risks objectively and activate risk countermeasures. The proposed RM method facilitated the unification of the organization's RM method. Thus, this method contributed toward the effective implementation of the RM method in an organization.

5.4 Full Visualization of Risk

In this study, after identifying the risk through an upstream process, we monitored the risks through weekly meetings and visualized the magnitude of an allowable recovery range by using statistical methods. Furthermore, the magnitude of the risk effect was converted into "time." We visualized the possibility of leadtime to delivery delay. Thus, the issue of full visualization of risk was resolved.

6 CONCLUSIONS

In this research study, we analyzed four actual RM processes carried out at the system development site. Additionally, the study identified the factors that contributed to delayed RM, despite the introduction of the correct RM process at the organization. Subsequently, we proposed the RM method using a quantitative process management approach that included PM, EVM, and LRA.

When we applied the method to the actual embedded development projects, we could verify and confirm the improvement effect on the reduction of the number of projects with delayed delivery times and a decrease in contingency. Thus, the proposed method is considered potentially effective.

This proposed method can be introduced easily in any organization implementing process improvement. In the future, it is necessary to increase case examples, evaluate effectiveness, and make improvements to the existing RM process.

REFERENCES

[1] Nikkei Computer, Survey on Information Actual Condition (2003)

[2] Standish Group International, Inc,"CHAOS Summary 2009," http://www.standishgroup.com, (2009)

[3] K.H. Rose, "A guide to the project management body of knowledge (pmbok guide) fifth edition," Project management journal，vol.44，no.3，(2013)

[4] Project Management Association Japan, P2M Program,Project Management Standard guidebook (2014)

[5] Martin Tomanek and Jan Juricek , "Project RM model based on prince2 and scrum frameworks," International Journal of Software Engineering Applications (IJSEA), Vol.6, No.1 (2015)

[6] B.W. Boehm, "Software risk management: principles and practices," IEEE software, vol.8, no.1, pp.32-41(1991)

[7] R.C.Williams, J.A.Walker, and A.J.Dorofee, "Putting risk management into practice," IEEE software, vol.14, no.3, pp.75-82,(1997)

[8] ISO9001(2015).Quality Management System Requirement. Japanese Standards Association.

[9] Chrissis,M.B. Konrad,M. and Shrum,S. CMMI for development: guidelines for process integration and product improvement, Pearson Education.(2011)

[10] Rasool,M, Franck,T. Denys,B. and Halidou,N. Methodology and tools for risk evaluation in construction projects using risk breakdown structure, European J. of Environmental and Civil Engineering, vol.16, pp.78-98.(2012).

[11] Menezes,J, Gusm o,Jr,C. and Moura,H Defining indicators for risk assessment in software development projects, Clei electronic J., vol.16, no.1, pp.17-21.(2013)

[12] Galway,L. Quantitative risk analysis for project management, a critical review, (2004).

[13] Zafar. K Software project risk management by using Six Sigma approach, Int.J. of Engineering Research and General Science, vol.3, no.4, pp.17-21 (2015)

Trust Representation under Confusion and Ignorance

Kazuhiko Ohkubo[1], Tetsuhisa Oda[2], Yuki Koizumi[3], Tetsushi Ohki[4],
Masakatsu Nishigaki[4], Toru Hasegawa[3] and Yoshinobu Kawabe[5]

[1]NTT Secure Platform Laboratories, NTT Corporation
3-9-11 Midori-cho, Musashino-shi, Tokyo 180-0012, Japan
E-mail : ohkubo.kazuhiko@lab.ntt.co.jp

[2]Department of Business Administration, Aichi Institute of Technology
2-49-2 Jiyugaoka, Chikusa-ku, Nagoya, Aichi 464-0044, Japan
E-mail : oda@aitech.ac.jp

[3]Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
E-mail : { ykoizumi, t-hasegawa }@ist.osaka-u.ac.jp

[4]Graduate School of Science and Technology, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan
E-mail : { ohki, nishigaki }@inf.shizuoka.ac.jp

[5]Department of Information Science, Aichi Institute of Technology
Yachigusa 1247, Yakusa-cho, Toyota, Aichi 470-0392, Japan
E-mail : kawabe@aitech.ac.jp

Abstract - Handling trust among agents/users is an important issue in the Internet, and many researchers have tackled on describing and analyzing trust values. Marsh and Dibben used a real value in the range of $[-1, 1)$ to express a trust value, where value 1 corresponds to a state for an agent fully trusted and value -1 corresponds to a state of a complete distrust. However, there are cases where such a one-dimensional representation of trust is not sufficient; for example, we cannot give a proper trust value on an agent if the agent is both trusted and distrusted — this kind of confusion may happen since trust is a property closely related to human's impressions. Also, ignorance on an agent may affect giving a trust value. To deal with trust values for cases where some confusion or ignorance is allowed, this study employs a multi-dimensional representation of Fuzzy-set Concurrent Rating method, FCR for short. The FCR method is originally developed for the study of mathematical psychology, and it can handle an answer like ”basically good, but somewhat bad simultaneously”. With the FCR method, we can describe trust values in a two-dimensional way, and this enables a proper description of trust states. In this study we introduce how to represent a trust value, and we discuss how the trust value corresponds to a one-dimensional trust value defined by Marsh and Dibben.

Keywords: On-line Trust, FCR Method, Fuzzy Logic, I/O-automaton

1 Introduction

Nowadays a huge amount of sensors and devices are connected to the Internet and various messages which may contain private information are massively exchanged. Also, people often exchange important messages actively; for example, in the major disaster of the Great East Japan Earthquake of 2011, many messages on human life and relief are exchanged through social media. For such situations, it is an important requirement to deal with the trustworthiness on participants/messages for communications.

Marsh and Dibben [1] have formalized and classified some trust notions as follows:

- Trust: a measure of how much an agent believes a trustee,
- Distrust: a measure of how much an agent believes that the trustee will actively work against the agent in a given situation,
- Untrust: a measure of how little the trustee is actually trusted, and
- Mistrust: a misplaced trust.

In their study the degree of trust is modeled with a score in $[-1, 1)$, and the notions of trust, distrust and untrust are formally defined as follows:

- Trust: a state where the trustee's trust value is more than a threshold value and enough to cooperate,
- Distrust: a state where the trustee's trust value is negative, and
- Untrust: a state where the trustee's trust value is positive but not enough to cooperate.

For the notions of trust and distrust, Lewicki et.al [2] suggest that trust and distrust are entirely separate dimensions, and as described in [1][2], low distrust is not same as high trust, and high distrust is not same as low trust. From this discussion, we can see that there are cases where one-dimensional expression is not sufficient for trust values.

We describe an example situation where one trusts and distrusts a message simultaneously. Suppose that in a large scale disaster you have received a message that describes the current situation of your friend. The content of the message was enough consistent, and you understood that the message was probably true. However, someone later told you that the message was not sent from a correct sender. In this case, you may trust the message from the viewpoint of message's consistency, but simultaneously you may not trust the message from the viewpoint of sender's integrity.

Trust is a property closely related to human's impressions. Hence, we consider that a mathematical psychology's method for impression formation is well applicable to describing trust values. Oda [3][4][5] developed such a method in the filed of fuzzy set theory. The Fuzzy-set Concurrent Rating method, FCR for short, enables us to measure and analyze human's impressions. By applying the FCR method to trust representation, we can describe trust and distrust notions properly. Furthermore, we can introduce some "finer" untrust notions. That is, it is possible to define two sorts of untrust notions:

- Untrust under confusion: a state where the trustee is both trusted and distrusted, and
- Untrust under ignorance: a state where the trustee is ignored; that is, the trustee is both little trusted and little distrusted.

The original untrust notion by Marsh and Dibben corresponds the untrust under ignorance. In this paper, we give a FCR-based semantics for the trust notions.

This paper is organized as follows. Section 2.1 describes an overview of the FCR method. In Section 3, we describe the trust notions shown in [1], and we apply the FCR method to represent a trust value. Finally, we discuss the validity of our trust representation in Section 4; also, an analysis for time-related trust notions such as mistrust or swift trust is discussed.

2 Preliminaries

2.1 Fuzzy-Set Concurrent Rating Method

The rating scale method (Fig. 1) is often used for questionnaire, where there are items such as "poor", "fair", "average", "good", and "excellent" and a respondent chooses one from them. This is a simple way, but this method has a problem that the respondent tends to choose a "middle" item of a scale, which makes an analysis difficult. There are two cases where this problem occurs. The first case is that the respondent has multiple candidates for an answer. Especially, if the candidates are at both extremities, then the respondent may choose one of them (forcibly) or a middle item. The chosen middle item is not a true answer of the respondent; moreover, the middle item usually has a label such as "average", "neutral" or "I do not know", and this makes the analysis more difficult. Another case is that the respondent does not have enough knowledge/interest for judgment. In any case, we cannot determine if the chosen middle item is the true answer or not.

In the FCR method, a respondent is requested to answer the confidence for each item (Fig. 2); that is, the respondent answers how much s/he thinks so. The range of each confidence value is from 0 to 1. Then, we calculate an integration value from a set of the confidence values by applying the fuzzy inference. From the theoretical point of view we have no restriction with regard to the dimension (i.e. the number of items), but from the application standpoint dimensions of about 2 or 3 are considered practical; for example, two dimensions by "true" and "false", three dimensions by "agree", "disagree" and "I do not know", and so on. In the rest of this paper, we have discussions for the case of two dimensions.

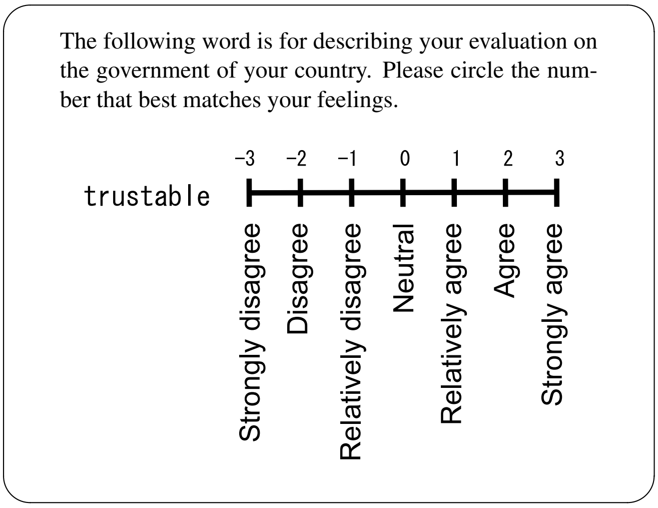

The following word is for describing your evaluation on the government of your country. Please circle the number that best matches your feelings.

Figure 1: Conventional questionnaire

2.1.1 HLS Model

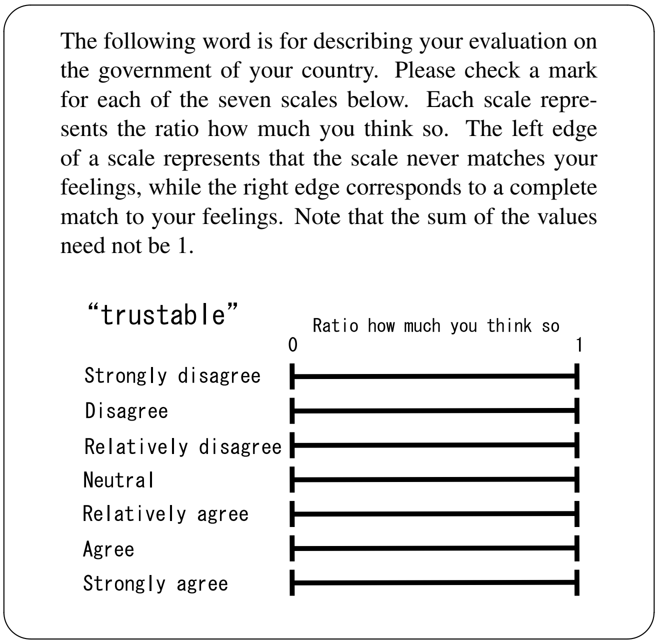
The FCR method employs the Hyper Logic Space model, HLS for short, as a logic space of the multiple-dimensional multiple-valued logic. Figure 3 shows a two-dimensional space based on two metrics of $true$ (T) and $false$ (F). An observation (t, f), or an observed point, is a point in $T \times F$, where we have $t, f \in [0, 1]$. Note that t and f are independent; that is, we do not introduce restrictions such as $t+f = 1$ or $t+f \leq 1$. A domain $\{(t, f) \,|\, t, f \in [0, 1] \wedge t + f > 1\}$ is called the region of contradiction. Also, a domain $\{(t, f) \,|\, t, f \in [0, 1] \wedge t + f < 1\}$ is called the region of ignorance, or the region of irrelevance. A domain $\{(t, f) \,|\, t, f \in [0, 1] \wedge t + f = 1\}$ is called the numerical truth value space, or the consistent region. If we regard observations $A = (t_A, f_A)$ and $B = (t_B, f_B)$ as truth values, logical operations over HLS are defined as:

$$\begin{array}{rcl} A \vee B & = & (max(t_A, t_B), min(f_A, f_B)), \\ A \wedge B & = & (min(t_A, t_B), max(f_A, f_B)), \text{ and} \\ \neg A & = & (1 - t_A, 1 - f_A). \end{array}$$

For details on the logical operations, see [6].

2.1.2 Integration Value and Degree of Contradiction

For observation (t, f), truth value t and falsity value f are given independently. From these values, we need to calculate an "actual" truth value. This special truth value is called an



The following word is for describing your evaluation on the government of your country. Please check a mark for each of the seven scales below. Each scale represents the ratio how much you think so. The left edge of a scale represents that the scale never matches your feelings, while the right edge corresponds to a complete match to your feelings. Note that the sum of the values need not be 1.

“trustable”

Ratio how much you think so
0 1

Strongly disagree
Disagree
Relatively disagree
Neutral
Relatively agree
Agree
Strongly agree

Figure 2: Rating with the FCR method

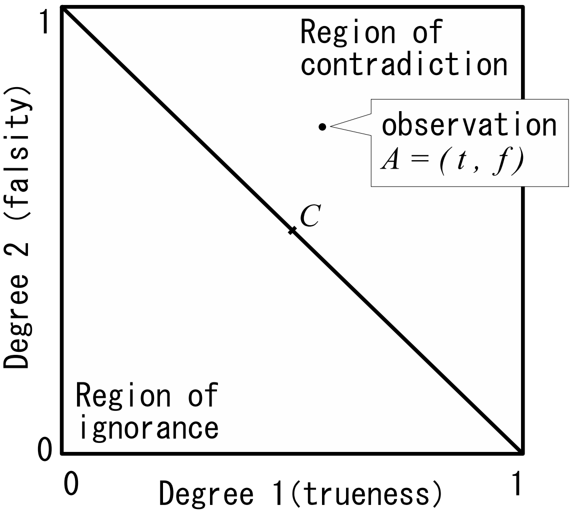


Figure 3: Two-dimensional HLS model

integration value. There are several ways to calculate integration values, and it is chosen properly according to the analysis target. Mainly, the following methods are used:

- Simple scoring method: In this case, integration value I_1 is defined as:

$$I_1(t, f) = \begin{cases} 0.5 & \text{if } t + f = 0, \\ \dfrac{t}{t+f} & \text{otherwise;} \end{cases}$$

- Reverse-item averaging method: Integration value I_2 is given by calculating the average of t and $1-f$, where t is a positive measure and f is a negative measure. That is, we have:

$$I_2(t, f) = \frac{t + (1 - f)}{2};$$

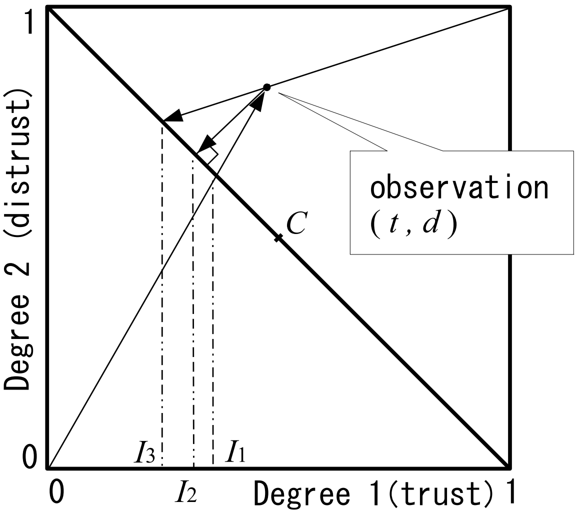


Figure 4: Graphical calculation for integration values

- Inverse scoring method: For measures t and f, we employ their negations $f' = 1 - t$ and $t' = 1 - f$. Then, we apply the simple scoring method for t' and f'. That is, integration value I_3 is given as:

$$I_3(t, f) = \begin{cases} 0.5 & \text{if } t + f = 2, \\ \dfrac{1-f}{2-t-f} & \text{otherwise.} \end{cases}$$

These methods are sometimes combined; for example, to calculate an integration value we employ $I_1(t, f)$ and $I_3(t, f)$ for the region of irrelevance and for the region of contradiction, respectively. Note that we have $I_1(t, f) = I_3(t, f)$ for the consistent region. In the FCR method, we can calculate the integration values $I_1(t, d)$, $I_2(t, d)$ and $I_3(t, d)$ in a graphical manner (see Fig. 4). For $I_2(t, d)$, we draw a perpendicular line from the observation (t, d) to Fig. 4's diagonal line, and we read the value of "Degree 1"[1].

Another important special value, which is called the degree of contradiction, has been introduced in the FCR method. In the field of personality psychology, we allow situations like "I like it but I don't like it" or "I don't like it but I don't dislike it." This kind of confusion is formulated and measured with a value called the degree of contradiction [3]; we employ this value to handle situations like "This message is trustable but simultaneously not trustable." The degree of contradiction C, which is more precisely the degree of contradiction-irrelevance, has the value of $C = 1$ for the case of total contradiction, $C = 0$ for the case of $t + f = 1$, and $C = -1$ for the case of complete ignorance, respectively. We often use the definition for the degree of contradiction given by:

$$C(t, f) = t + f - 1$$

which represents the distance between the observation (t, f) and the consistent region.

[1] For $I_1(t, d)$, we draw a line from $(0, 0)$ to (t, d); $I_1(t, d)$ is the value of Degree 1 at the crossing point with the diagonal line. For $I_3(t, d)$, we draw a line from $(1, 1)$ and we read the value of Degree 1 at the intersection with the diagonal line.

2.2 Four Trust Notions

Marsh and Dibben[1] introduced four trust notions: trust, distrust, untrust, and mistrust, and they formulated the condition for each of the trust notions. To formalize the trust notions, Marsh and Dibben have used the following functions with range $[-1, 1)$:

- General trust $T_x(y)$: how much x trusts y; and
- Situational trust $T_x(y, \alpha)$: how much x trusts y in situation α.

The situational trust $T_x(y, \alpha)$ is specifically defined as:

$$T_x(y, \alpha) = U_x(\alpha) \times I_x(\alpha) \times \widehat{T_x(y)}$$

where $U_x(\alpha)$ is the utility function to agent x for situation α, $I_x(\alpha)$ is the importance of α to x, and $\widehat{T_x(y)}$ is x's estimation on the trust value for trustee y [7]. Note that $U_x(\alpha)$, $I_x(\alpha)$ and $\widehat{T_x(y)}$ range over $[0, 1]$, $[0, 1]$ and $[-1, 1)$, respectively. For example, let us consider the "two agents and two pieces of furniture" example shown in [7]. There are two agents (A and B) in a room, each with the task of carrying out one piece of furniture. Each agent cannot move his piece of furniture alone, and each piece of furniture needs two agents to move it. Each agent must consider to cooperate with the other in order to get the job done. However, an agent may choose not to cooperate, and in this case there are no sanctions. This situation is labeled with α. Suppose agent B is relatively trustable for agent A, and this is estimated with $\widehat{T_A(B)} = 0.63$. Agent A considers the importance on the situation is fair; i.e. $I_A(\alpha) = 0.5$. The furniture of agent A is his own, and he believes that the utility to move it from the room is very high. This is estimated with $U_A(\alpha) = 0.8$. Summarizing, the estimated situational trust is $T_A(B, \alpha) = 0.8 \times 0.5 \times 0.63 = 0.252$. If $T_x(y)$ is greater than a value called a cooperation threshold, then y is trusted by x; similarly, we say y is trusted by x in situation α if $T_x(y, \alpha)$ is greater than the threshold value.

Distrust is a measure of how much an agent believes that the trustee will actively work against the agent in a given situation. That is, if agent x distrusts y in situation α, the agent x expects that y will work to make sure the worst (or at least not the best) will happen in situation α. In other words, distrust is an active judgment in the negative intentions of the other. This is defined as:

$$T_x(y, \alpha) < 0 \implies Distrust(x, y, \alpha).$$

We can see that the notion of distrust is defined as a case where $T_x(y, \alpha)$ is negative.

Untrust is a measure of how little a trustee is actually trusted. If we say a trustee is untrusted, then the truster has little confidence in the trustee acting in their best interests in a situation. In other words, the truster cannot determine if the trustee is trustable. In such a situation, it is not enough for a truster to cooperate with a trustee. Formally, a state of untrust is defined with:

$$(T_x(y, \alpha) > 0 \land T_x(y, \alpha) < CT_x(y, \alpha)) \\ \implies Untrust(x, y, \alpha).$$

In this formula, $CT_x(y, \alpha)$ is a cooperation threshold which is given by:

$$CT_x(y, \alpha) = \frac{Risk_x(\alpha)}{Comp_x(y, \alpha) + \widehat{T_x(y)}} \times I_x(\alpha)$$

where $Risk_x(\alpha)$ is a value on x's perceived risk in situation α, and $Comp_x(y, \alpha)$ is a x's evaluation value on how much y has a competence for α. From this definition, we can see that the cooperation threshold for trust is set high if the values of risk $Risk_x(\alpha)$ and importance $I_x(\alpha)$ are high. Also, the cooperation value becomes low if the value of $Comp_x(y, \alpha)$ is high, that is, if the truster x believes that the trustee y can deal with things in situation α. The cooperation value also becomes low if the value of $\widehat{T_x(\alpha)}$ is high.

Finally, the notion of mistrust is introduced in [1]. This is a misplaced trust, and it is discussed in Section 4.2.

3 FCR-based Trust Representation

In the previous section we have shown the functions such as $T_x(y)$ and $T_x(y, \alpha)$ to explain a formalization of three trust notions (except the notion of mistrust). A trust value is an element in $[-1, 1)$, and the notions of trust, distrust and untrust are defined one-dimensionally. However, as described in [1], low distrust is not same as high trust, and high distrust is not same as low trust. In this section we assume that trust and distrust are entirely separate dimensions, and we use the FCR method in Section 2.1 for trust representations.

In this paper we use the set of trust values $Trust$ and the set of distrust values $DisTrust$, which are defined as:

$$Trust = DisTrust = \{\, v \mid 0 \leq v \leq 1\}.$$

An observation is an element of $Trust \times DisTrust$. For any observation $o \in Trust \times DisTrust$, we can see that

1. If o is around $(1, 0) \in Trust \times DisTrust$, then the observation has a high trust value and a low distrust value;
2. If o is around $(0, 1)$, then the observation has a low trust value and a high distrust value;
3. If o is around $(0, 0)$, then both of the trust and distrust values are low;
4. If o is around $(1, 1)$, then both of the trust and distrust values are high; and
5. If o is around $(0.5, 0.5)$, the both kind of values are moderate.

Here we have five cases, and it is natural to understand that the first case corresponds to Section 2.2's trust state while the second case corresponds to a distrust state. For cases 3, 4 and 5, we cannot determine whether a trustee is actually trustable; hence, these cases basically correspond to a state of untrust. However, for cases 3, 4 and 5, if the distrust value is greater than the trust value, it is considered appropriate that the observation belongs to a state of distrust.

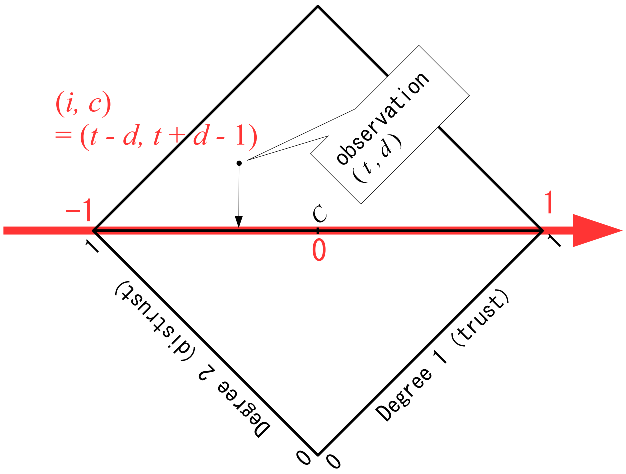


Figure 5: From 2D representation to 1D representation

Points $(1, 0)$ and $(0, 1)$ are both ends of the diagonal line shown in Fig. 3. Also, we can see that every point in the diagonal line is ideal in the sense that the degree of trust t and the degree of distrust d follow the consonance law; that is, $t + d = 1$. In this study we transform the diagonal line to the measure by Marsh and Dibben. Specifically, we move points $(1, 0)$, $(0, 1)$ and $(0.5, 0.5)$ to $(1, 0)$, $(-1, 0)$ and $(0, 0)$, respectively. Note that $(1, 0)$, $(-1, 0)$ and $(0, 0)$ correspond to trust values 1, -1 and 0, respectively. Observation $(t, d) \in Trust \times DisTrust$ is mapped to $(t - d,\ t + d - 1)$ by

$$\left[\begin{pmatrix}\cos\frac{\pi}{4} & -\sin\frac{\pi}{4}\\ \sin\frac{\pi}{4} & \cos\frac{\pi}{4}\end{pmatrix}\left\{\begin{pmatrix}t\\ d\end{pmatrix}-\begin{pmatrix}1\\ 0\end{pmatrix}\right\}+\begin{pmatrix}\frac{\sqrt{2}}{2}\\ 0\end{pmatrix}\right]\times\frac{1}{\frac{\sqrt{2}}{2}}$$
$$=\begin{pmatrix}t-d\\ t+d-1\end{pmatrix}$$

and we call the resulting point (i, c); see Fig. 5. Note that we have $c = 0$ if the original point (t, d) is in the consistent region (i.e. $t + d - 1 = 0$).

For the resulting point (i, c), the first element $i = t - d$ represents the "actual" trust value given by Marsh and Dibben, and this is calculated with the reverse-item averaging method in Section 2.1.2; actually, the range of integration value $I_2(t, d)$ is $[0, 1]$ and the value of i is calculated by normalizing $I_2(t, d)$ to be a value in region $[-1, 1]$. The second element $c = t + d - 1$ coincides the degree of contradiction-irrelevance in the FCR method.

As described in Section 2.1.2, we can calculate the integration values in a graphical manner. Observing the graphical calculation, points in the same perpendicular line have the same integration value with regard to I_2. For example, observation $A = (t, d)$ and its nearest point on the diagonal line

$$A' = (\frac{t + (1 - d)}{2}, 1 - \frac{t + (1 - d)}{2})$$

have the same integration value, which means that $i = t - d$ and $i' = \frac{t + (1 - d)}{2} - (1 - \frac{t + (1 - d)}{2})$ are equivalent. However, for A and A', the distance from the diagonal line is different. The distance between the observation and the diagonal line is given by $|t+d-1|$, which is the absolute value of the second element of the resulting point (i, c). This value coincides (the absolute value of) the degree of contradiction-irrelevance in the FCR method.

4 Discussions

4.1 Validity of Two-Dimensional Trust Representation

Let $(t, d) \in Trust \times DisTrust$ be an observation and (i, c) be its corresponding resulting point by the transformation of the previous section. We can see that the "actual" trust value $i = t - d$ is calculated by subtracting the degree of distrust from the degree of trust; we believe that this is consistent with the intuition. Furthermore, we can explain the trust notions in Section 2.2 properly by defining:

- A state of trust is a state with $i \geq CT$, where CT is a cooperation threshold; and
- A state of distrust is a state with $i < 0$.

The second element $c = t + d - 1$ of (i, c) is the degree of contradiction-irrelevance in the FCR method, and the value enables us to introduce a new sort of untrust notion. In the field of fuzzy logic, it is considered that a state of contradiction, where the value of c is high, is caused by an information overload. Also, a state of irrelevance, where the value of c is negative, is considered due to a lack of information. This study considers that an information overload is closely related to an excessive interest in a trustee. Actually, if you are excessively interested in a trustee, you may have many evidences for judgment. Some of the evidences might increase the trust value on the trustee, but simultaneously there may exist another evidence that increases the distrust value. Also, this study considers that a lack of information is caused by an ignorance. If you are not interested in the trustee and you do not know anything, you cannot discuss whether the trustee is trustable or not. At least, in such a case, both of the trust value and the distrust value remain low, since there exist no evidences that increase the trust/distrust value.

With the degree of contradiction-irrelevance, we can further classify the notion of untrust. This study introduces the following two types of untrust notions:

- Untrust under confusion: This is a state where a trustee is both trusted and distrusted. Formally, this is a state with $0 < i < CT$ and $c \geq 0$; and
- Untrust under ignorance: This is a state where the trustee is ignored; in other words, the trustee is both little trusted and little distrusted. Formally, this is a state with $0 < i < CT$ and $c < 0$.

The original untrust notion in [1] is considered as the notion of untrust under ignorance. On the other hand, this paper introduced a new kind of untrust notion from the viewpoint of "contradiction", and this result is obtained based on the FCR method.

```
automaton mistrust(id, alpha)
  signature
    output move(i: Nat,
                pt: VL, pd: VL,
                ev:Event,
                dt: VL, dd: VL)
  states
    t: Array[Nat, VL],   % Sort VL is the set of
    d: Array[Nat, VL],   % real values [-1, 1]
    step: Nat := 0,
    stateOfId: State := InitState(id, alpha)
      so that (\A i:Nat (t[i] = 0 /\ d[i] = 0))

  transitions
    output move(i, pt, pd, ev, dt, dd)
      pre    i = step
          /\ pt = t[i]
          /\ pd = d[i]
          /\ (0 <= pt + dt /\ pt + dt <= 1)
          /\ (0 <= pd + dd /\ pd + dd <= 1)
          /\ condition(id, alpha,
                       i, ev, pt, pd, dt, dd,
                       stateOfId)
      eff t[i+1] := t[i] + dt;
          d[i+1] := d[i] + dd;
          step := step+1;
          stateOfId := change(id, alpha,
                              stateOfId, i, ev)
```

Figure 6: IOA specification

4.2 Mistrust

Mistrust is a state in which initial trust has been betrayed; more precisely, the notion of mistrust can be considered as "either a former trust destroyed, or former distrust healed" since the trustee may not have had bad intentions and it is not always "betrayed". This is considered as a property with regard to a change of trust value over time.

We consider that it is possible to model a change of trust value with IOA [8], which is a formal specification language based on I/O-automaton theory [9][10]. For example, the automaton in Fig. 6 models the change of actor `id`'s trust value in situation `alpha`. In this specification, `step` is a clock, and `t[i]` and `d[i]` are the trust value and the distrust value at clock `i`, respectively. The range of `t[i]` and `d[i]` is $[0, 1]$. The variable `stateOfId` is actor `id`'s state and its initial value is given by `InitState(id, alpha)`. The only action is `move(i, pt, pd, ev, dt, dd)`, and an occurrence of this action means:

> Actor `id`'s trust value is $(\mathtt{pt}, \mathtt{pd})$ at clock `i`, and there is an occurrence of event `ev` which changes the trust value to $(\mathtt{pt} + \mathtt{dt}, \mathtt{pd} + \mathtt{dd})$ at clock `i+1`.

To analyze time-related trust properties such as mistrust or swift trust [11][12], a verifier should appropriately describe the predicate `condition` in the precondition part of action `move`. For example, in verifying swift trust, the verifier should define `condition` to satisfy that $\mathtt{t[step]} - \mathtt{d[step]}$ becomes greater than a cooperation threshold within a given short period. We consider that it can be modeled as a liveness property defined with a trace set. Swift trust is attracting attention for disaster situations [13][14][15], and its importance is getting larger recently. It is a future work to formalize the various trust notions for disaster situations.

5 Conclusion

In this paper, we introduced a trust representation based on a theory for impression formation — the FCR method. A trust value was given as an observation in the two-dimensional HLS, where we employed two metrics of "trust" and "distrust". We discussed the validity of this representation by showing a mapping to the one-dimensional representation by Marsh and Dibben. For the untrust notion, we introduced a new classification; that is, there are two sorts of untrust notions called the untrust under confusion and the untrust under ignorance.

In this study we discussed how to represent a single trust state, but we are actually interested in trust transitions. For future work, we are planning to formalize and analyze time-related trust notions such as swift trust. We believe that results from theory of distributed algorithms, such as analysis methods for safety properties or liveness properties, are applicable to analyze trust notions. In this paper we described an automaton (in Fig. 6) as a first step to define and analyze the trust properties. It gives one idea for the analysis of time-related trust, but we do not say that it is a complete definition; it should be refined and sophisticated.

It is also crucial to conduct experiments to apply our representation of trust to real situations. We consider that dealing with a disaster situation is especially important, since adversaries and malicious actors often seek to exploit the vulnerable situation. For such situations it is important to evaluate the trust of messages and participants with our trust representation.

Acknowledgment

This work has been supported by the National Institute of Information and Communications Technology (NICT) in Japan (Contract No. 193).

REFERENCES

[1] S. Marsh and M. R. Dibben, "Trust, untrust, distrust and mistrust – an exploration of the dark(er) side," in *Proceedings of the Third International Conference on Trust Management*, iTrust'05, (Berlin, Heidelberg), pp. 17–33, Springer-Verlag, 2005.

[2] R. J. Lewicki, D. J. B. McAllister, and R. J. Bies, "Trust and distrust: New relationships and realities," *Academy of Management Review*, vol. 23, pp. 438–458, 1998.

[3] T. Oda, "Fundamental characterestics of fuzzy-set concurrent rating method," *Journal of Japan Association for Management Systems*, vol. 12, no. 1, pp. 23–32, 1995. In Japanese.

[4] T. Oda, "Fuzzy set theoretical approach for improving the rating scale method : Proposing and introducing the FCR-method and the IR-method as novel rating methods," *Japanese Psychological Review*, vol. 56, no. 1, pp. 67–83, 2013. In Japanese.

[5] T. Oda, "Measurement technique for ergonomics, section 3: Psychological measurements and analyses (3) measurements and analyses by kansei evaluation,"

The Japanese Journal of Ergonomics, vol. 51, no. 5, pp. 293–303, 2015. In Japanese.

[6] T. Oda, “A proposal for multidimensional multivalued logic space: Hyper logic space — Extended fuzzy-logic for the fuzzy-set concurrent rating method —,” *Japanese Journal of Industrial Management Association*, vol. 49, no. 3, pp. 135–145, 1998. In Japanese.

[7] S. P. Marsh, “Formalising trust as a computational concept,” tech. rep., University of Stirling, 1994.

[8] A. Bogdanov, “Formal verification of simulations between I/O-automata,” Master’s thesis, Massachusetts Institute of Technology, 2000.

[9] N. A. Lynch, *Distributed Algorithms*. Morgan Kaufmann Publishers, 1996.

[10] N. Lynch and F. Vaandrager, “Forward and backward simulations — part I: Untimed systems,” *Information and Computation*, vol. 121, pp. 214–233, Sept. 1995.

[11] D. Meyerson, K. E. Weick, and R. M. Kramer, *Swift Trust and Temporary Groups in Trust in Organizations: Frontiers of Theory and Research*. SAGE, 1995.

[12] J. Wildman, M. Shuffler, E. Lazzara, S. Fiore, and S. Burke, “Trust development in swift starting action teams: A multilevel framework,” *Group & organization management*, vol. 37, no. 2, pp. 137–170, 2012.

[13] Y. Murayama, “Issues in disaster communications,” *Journal of Information Processing*, vol. 22, no. 4, pp. 558–565, 2014.

[14] M. G. Busa, M. T. Musacchio, S. Finan, and C. Fennell, “Trust-building through social media communications in disaster management,” in *Proceedings of the 24th International Conference on World Wide Web*, WWW ’15 Companion, (New York, NY, USA), pp. 1179–1184, ACM, 2015.

[15] F. Lemieux, “The impact of a natural disaster on altruistic behaviour and crime,” *Disasters*, vol. 38, pp. 483–499, July 2014.

Elliptic Curves Suitable for Elliptic Curve Signature

Masaaki Shirase[†]

[†]School of Systems Information Science, Future University Hakodate, Japan
shirase@fun.ac.jp

Abstract - In an elliptic curve signature using 256-bit prime p, thousands of modular multiplications $X \cdot Y \bmod p$ performed according to a signature algorithm are dominant. Therefore, how to speed up multiplication and $\bmod p$ computations is one of the objectives of researches on elliptic curve signature implementations. One of speeding up method of reduction $\bmod p$ is to use a special form of prime called pseudo Mersenne prime such that $p = 2^n - k$, where k is a small value. However, in an elliptic curve signature, computation of $\bmod l$ with another integer l, which is the order of a base point, also requires although the number is a few.

In this paper, the authors give a program to construct elliptic curves such that reduction $\bmod l$ can be computed as mod a pseudo Mersenne prime and give an example of such curves.

Keywords: Elliptic Curve, Elliptic Curve Signature, Modular Multiplication, Pseudo Mersenne Prime

1 INTRODUCTION

Recently, elliptic curve signatures as ECDSA are often used in the TLS communication and block chains. In Europe, it was decided to use an elliptic curve signature in the V2X communication [5]. However, signature verification in the V2X communication requires further speeding up.

Elliptic curve is a cubic curve given by Weierstrass form ($y^2 = x^3+ax+b$) or Montgomery curve ($By^2 = x^3+Ax^2+x$) [9]. A remarkable feature of elliptic curve is that an operation $+$ is defined [14]. Dominant processes of the operation $+$, which is explained in Sec.2.2, are modular multiplications $X \cdot Y \bmod p$ for $X, Y \in \mathbb{F}_p = \{0, 1, 2, \cdots, p-1\}$, where p is typically a 256-bit prime. The calculation of it is divide into

$$Z \leftarrow \underbrace{X}_{256 \text{ bit}} \cdot \underbrace{Y}_{256 \text{ bit}} \text{ and } W \leftarrow \underbrace{Z}_{512 \text{ bit}} \bmod \underbrace{p}_{256 \text{ bit}}. \quad (1)$$

Dominant processes of elliptic curve cryptosystems (ECCs) including elliptic curve signatures are thousands of modular multiplications. Therefore, it is important to speed up (1) to speed up processes of elliptic curve signatures. As explained in Sec.2.3, using Montgomery curve rather than Weierstrass form reduces the number of modular multiplications required for signature generation and verification. Moreover, when a coefficient A of Montgomery curve is 6, 10, 14, and 18, the number of modular multiplications is further reduced.

Montgomery reduction [8] that can be applied to arbitrary odd number p is a famous method for reduction $\bmod p$. Also, when p is a pseudo Mersenne prime written as $p = 2^n - k,\ k < 2^{n/2}$, reduction $\bmod p$ can be very efficient.

In public key encryptions as ECElGamal and key agreements as ECDH using elliptic curves, required reduction is $\bmod p$ only. On the contrary, in elliptic curve signatures as ECDSA, required reductions are not only $\bmod p$ but also $\bmod l$, where l is the order of a base point. As far as the authors know, the efficiency of $\bmod p$ is taken into account in the construction of an elliptic curve for ECCs, but the efficiency of $\bmod l$ is not considered.

The purpose of this paper is to make a program to find Montgomery curves such that $\bmod l$ can be computed by same way of $\bmod p$ for pseudo Mersenne prime and $A = 6$, 10, 14, and 18, and to give examples of such curves.

Sec.2 explains the definition of elliptic curve, operation $+$, scalar multiplication, coordinate system, secure elliptic curve, and Curve25519. Sec.3 introduces ECDSA that is the most popular elliptic curve signature. Sec.4 introduces efficient reduction methods. Sec.5 is the contribution of this paper. Sec.5 proposes a requirement for elliptic curves to be suitable for ECDSA, makes a program to find elliptic curves that meet the requirement, and gives examples of such curves. Sec.6 concludes this paper and gives future works.

2 ELLIPTIC CURVE

2.1 Definition of Elliptic Curve

Elliptic curve is a cubic curve given by

$$E : y^2 = x^3 + ax + b \text{ (Weierstrass form)} \quad (2)$$

or

$$E : By^2 = x^3 + Ax^2 + x \text{ (Montgomery curve)} \quad (3)$$

with variables x, y. When used in cryptosystems, Montgomery curve (3) is often selected because it can reduce the cost of cryptographic processes.

For a prime p, the set $\mathbb{F}_p$ is defined as $\mathbb{F}_p = \{0, 1, 2, \cdots, p-1\}$. The set $E(\mathbb{F}_p)$ is of points on E be considered on $\mathbb{F}_p$, where $E(\mathbb{F}_p)$ includes a special point $\mathcal{O}$ called the point at infinity.

The order of $E(\mathbb{F}_p)$, $\#E(\mathbb{F}_p)$, is defined as the number of points in $E(\mathbb{F}_p)$. The trace of $E(\mathbb{F}_p)$ is defined as an integer such that

$$\#E(\mathbb{F}_p) = p + 1 - t.$$

When a Montgomery curve $E : By^2 = x^3 + Ax^2 + x$ on $\mathbb{F}_p$ has the trace t, another Montgomery curve $E' : B'y^2 = x^3 + Ax^2 + x$ has the trace t or $-t$. In other words, we have $\#E'(\mathbb{F}_p) = \#E(\mathbb{F}_p)$ or $2p+2-\#E(\mathbb{F}_p)$. When $\#E'(\mathbb{F}_p) = 2p + 2 - \#E(\mathbb{F}_p)$, E' is called the twist of E.

2.2 Operation $+$

Let E be an elliptic curve given by Weierstrass form (2) or a Montgomery curve (3). Then, the operation $+$ in $E(\mathbb{F}_p)$ is defined as follows.

1. For any $P \in E(\mathbb{F}_p)$, $P + \mathcal{O} = \mathcal{O} + P$.

2. In the case of $P = (x_1, y_1), Q = (-x_1, y_1) \in E(\mathbb{F}_p)$, $P + Q = \mathcal{O}$.

3. In the case of $P = (x_1, y_1), Q = (x_2, y_2) \in E(\mathbb{F}_p), x_1 \neq x_2$, and $P \neq Q$, $P + Q = (x_3, y_3)$ is computed as

$$\left.\begin{aligned} \lambda &= \frac{y_1 - y_2}{x_1 - x_2}\ , \\ x_3 &= \lambda^2 - x_1 - x_2, \\ y_3 &= \lambda(x_1 - x_3) - y_1. \end{aligned}\right\} \tag{4}$$

4. In the case of $P = Q = (x_1, y_1) \in E(\mathbb{F}_p)$, $P + Q = (x_3, y_3)$ is computed as

$$\left.\begin{aligned} \lambda &= \begin{cases} \dfrac{3x_1^2 + a}{2y_1} & E\text{: Weierstrass} \\ \dfrac{3x_1^2 + Ax_1 + 1}{2By_1} & E\text{: Montgomery,} \end{cases} \\ x_3 &= \lambda^2 - 2x_1, \\ y_3 &= \lambda(x_1 - x_3) - y_1. \end{aligned}\right\} \tag{5}$$

Eqs.(4) and (5) are called addition formula and doubling formula, respectively.

For $P - Q = (x_4, y_4)$ on Montgomery curve, x_3 can be computed as

$$x_3 = \frac{(x_1 x_2 - 1)^2}{x_4 (x_1 - x_2)^2} \tag{6}$$

when $x_1 \neq x_2$. Eq.(6) is an important property of Montgomery curve.

2.3 Scalar Multiplication

For a base point $P \in E(\mathbb{F}_p)$ and a natural number n, additions of n terms of P,

$$nP = P + P + \cdots + P$$

is called scalar multiplication. For $P \in E(\mathbb{F}_p)$, the order of P is defined as the smallest positive integer l such that $lP = \mathcal{O}$.

For the order L of $E(\mathbb{F}_p)$ and the order l of $P \in E(\mathbb{F}_p)$, the followings are held (Lagrange's theorem).

1. l is a divisor of L.

2. $LP = \mathcal{O}$.

Algorithm 1 (Binary method) and Algorithm 2 (Montgomery ladder) are algorithms for computing a scalar multiplication. Let n be a k-bit integer, and

$$n = (n_{k-1}, n_{k-2}, \cdots, n_0)_2$$

be the binary representation of n. Then, Algorithm 1 takes k doubling formulas and $k/2$ addition formulas on average, and Algorithm 2 takes k doubling formulas and k addition formulas.

Table 1: Cost of scalar multiplication of nP, where n is k-bit

Used curve	Cost
Weierstrass	$18kM + kM_a + 18k\ add$
Montgomery	$10kM + kM_{A'} + 10k\ add$

Algorithm 1 (Binary method)
Input: $P \in E(\mathbb{F}_p), n = (n_{k-1}n_{k-2} \cdots n_0)_2 \in \mathbb{N}$
Output: $nP \in E(\mathbb{F}_p)$

```
Q ← P
for i = k − 2 down to 0
  Q ← 2Q
  if n_i = 1 then Q ← Q + P
end for
return Q
```

Algorithm 2 (Montgomery ladder)
Input: $P \in E(\mathbb{F}_p), n = (n_{k-1}n_{k-2} \cdots n_0)_2 \in \mathbb{N}$
Output: $nP \in E(\mathbb{F}_p)$

```
Q_0 ← O, Q_1 ← P
for i = k − 1 down to 0
  if n_i = 0 then Q_1 ← Q_0 + Q_1, Q_0 ← 2Q_0
  if n_i = 1 then Q_0 ← Q_0 + Q_1, Q_1 ← 2Q_1
end for
return Q_0
```

We will estimate the cost of theses algorithms. Let M, M_a, $M_{A'}$, and add denote a modular multiplication in $\mathbb{F}_p$, a modular multiplication in $\mathbb{F}_p$ with a constant a, a modular multiplication in $\mathbb{F}_p$ with a constant A', and a modular addition/subtraction in $\mathbb{F}_p$, respectively, where a is of (2), and $A' = (A + 2)/4$ for A of (3). The cost of a scalar multiplication is given by Table 1, where `add-1998-cmo-2` and `dbl-2007-b1` for Weierstrass, and `dadd-1987-m-3` and `dbl-1987-m-3` for Montgomery are used as algorithms for addition and doubling [2].

2.4 Secure Elliptic Curve

The security of ECCs including digital signature depends on the maximum prime factor l of the order $L = \#E(\mathbb{F}_p)$, not the size of p [12]. Attack time against ECCs is roughly proportional to $\sqrt{l}$ and then the larger l is, the more secure ECCs are. Therefore, we have to select elliptic curve E such that

$$L = \#E(\mathbb{F}_p) \text{ has a big prime factor} \tag{7}$$

for ECCs. Also, it is desirable that

$$\begin{gathered} 2p + 2 - L \text{ that is the order of the twist of } E \\ \text{has a big prime factor} \end{gathered} \tag{8}$$

according to [3]. Moreover, we have to select E such that

$$L \neq p, p \pm 1 \tag{9}$$

according to [6], [13].

Curve25519 [1] is a Montgomery curve

$$E_{25519} : y^2 = x^3 + 486662x^2 + x$$

with $p = 2^{255} - 19$. The order $L = \#E_{25519}(\mathbb{F}_p)$ is

$$L = 2^2 \cdot l,$$
$$l = 2^{252} + 27742317777372353535851937790883648493,$$

where l is a 253-bit prime. Curve25519 meets the security requirement in Sec,2.4.

Curve 25519 has been applied in many cryptographic libraries such as NaCl [10], and Curve 25519 was added to Special Publication 800-186, which specifies the approved elliptic curve used by the US federal government by NIST in 2017.

3 ECDSA

ECDSA is a digital signature using an elliptic curve. ECDSA consists of *system parameter* held by all users, *key generation* for generating each user's (private key, public key), *signature generation* for generating using a user A's secret key, and *signature verification* for verifying the signature using A's public key.

System parameter
A sufficiently large (e.g. 256-bit) prime p, an elliptic curve E such that $E(\mathbb{F}_p)$ meets the security requirements (7), (8), and (9), and a base point $G \in E(\mathbb{F}_p)$ of which the order is l are selected. Also a hash function $H : \{0,1\}^* \rightarrow \{0,1,2,...,l-1\}$ is selected. (p, l, E, G, H) is the system parameter.

Key generation
User A choose $s \in [1, l-1]$ at random, and computes $Y = sG$ (scalar multiplication) in $E(\mathbb{F}_p)$. Then, s and Y are A's private key and public key, respectively.

Signature generation
User A generates a signature of a message $m \in \{0,1\}^*$ as follows.

1. Computing $m' = H(m)$.
2. Choosing $r \in [1, l-1]$ at random, and compute
$$U = \underbrace{rG}_{\text{scalar mul. on } E(\mathbb{F}_p)} = (u_x, u_y),$$
$$u = u_x \bmod l.$$
3. Using the secret key s to compute
$$v = r^{-1}(m' + su) \bmod l.$$
4. The pair (u, v) is the signature of m.

Signature verification
A recipient of the message m with signature (u, v) verifies the signature as follows.

1. Computing $m' = H(m)$.
2. Computing $d = v^{-1} \bmod l$.
3. Computing $U' = \underbrace{(m'd)G}_{\text{scalar mul. on } E(\mathbb{F}_p)} + \underbrace{(ud)Y}_{\text{scalar mul. on } E(\mathbb{F}_p)}$.
4. Computing $u' = $ (the x coordinate of U') $\bmod l$.
5. If $u = u'$ then the signature is accepted, and if $u \neq u'$ then it is rejected.

Algorithm 3 (Montgomery Reduction)

Input: Odd number N of n bits, $R = 2^n$,
$N' = (-N^{-1}) \bmod R$, natural number $u < RN$
Output: $u \bmod N$ in Montgomery representation

```
$t \leftarrow u$
$k \leftarrow TN' \bmod 2^n$
$t \leftarrow t + NR$
$t \leftarrow t/R$
if $t \geq N$then $t \leftarrow t - N$
return $t$
```

Thus, the dominant processes of ECDSA is scalar multiplications in $E(\mathbb{F}_p)$. Signature generation of ECDSA takes one scalar multiplication and signature verification of ECDSA takes two scalar multiplications. Therefore, we see that

in order to speed up processes of ECDSA,
it is important to speed up scalar multiplication.

As seen Table 1, using not Weierstrass form but Montgomery curve reduces the number of modular multiplications required for a scalar multiplication. Moreover, using a Montgomery curve with appropriate coefficient A such as $A = 6, 10, 14, 18$ further reduces the number of modular multiplications.

4 MODULAR REDUCTION

In order to speed up ECCs including ECDSA, it is important not only to reduce the number of modular multiplications but also to reduce the cost of one modular multiplication. This section introduces efficient reduction methods.

The Montgomery reduction (Algorithm 3) [8] is a method for efficiently calculating $X \bmod N$ for general odd number N and X given in Montgomery representation[1].

When a prime p is written as

$$p = 2^n - k, \ k < 2^{n/2},$$

it is called pseudo Mersenne prime. For pseudo Mersenne prime p, reduction $\bmod p$ can be computed at high speed using by Algorithm 4 [7]. Notice that $u/2^n$ in step 1 and $v/2^n$ in step 3 are integer divisions and then they are performed by shift operations.

5 CONTRIBUTIONS

5.1 Program to Search Elliptic Curve Suitable for ECDSA

The purpose of this paper is to make a program to search for elliptic curves that is secure and suitable for high-speed

[1] For Montgomery representation, refer to [8] or [4].

Algorithm 4 (Reduction mod pseudo Mersenne prime)

```
Input: prime $p = 2^n - k$ $(k < 2^{n/2})$,
       integer $0 \leq u \leq (p-1)^2$
Output: $u \bmod p$
$u_0 \leftarrow u \bmod 2^n,\ u_1 \leftarrow u/2^n$
$v \leftarrow u_1 k + u_0$
$v_0 \leftarrow v \bmod 2^n,\ v_1 \leftarrow v/2^n$
$w \leftarrow v_1 k + v_0$
if $w \geq p$ then $w \leftarrow w - N$
return $w$
```

implementation of ECDSA (especially by hardware implementation), and to give examples of such an elliptic curves. Specifically, we will search curves that meet the following requirements.

Elliptic Curve Requirements to Search

1. Montgomery curve is selected because a scalar multiplication takes fewer modular multiplications (Table 1).

2. Moreover, Montgomery curve with $A' = 1, 2, 3, 4$, that is, $A = 6, 10, 14, 18$ is selected according to Table 1.

3. Prime p is is a pseudo Mersenne prime $p = 2^n - k$ of 256-bit. For convenience of execution time, set the range of k to $k \leq 2^{20}$.

4. To meet the security requirement (7), the order $L = \#E(\mathbb{F}_p)$, which is always a multiple of 4, is as $L = 4l, 8l, 16l$, where l is a prime.

5. To meet the security requirement (8), $L' = 2p + 2 - L$ is as $L' = 4l', 8l', 16l'$, where l' is a prime.

6. To meet the security requirement (9), curves such as $L = p, p \pm 1$ is removed.

7. L is written as $L = 2^n - k',\ k' < 2^{n/2}$. Then, a reduction $\bmod\, l$ is computed by an algorithm, which is as efficient as reduction mod a pseudo Mersenne prime.

Note Curve25519 also meets the requirements 1, 4, 5, and 7, and Curve25519 adopts not 256-bit prime but 254-bit for a prime field. Curve25519 does not consider the requirements 2 and 7.

The authors made a program as Fig.1 in PARI/GP [11] to search elliptic curves meeting the requirements. The program is straightforward and then it may be easy for some readers to make a similar program. But, giving the program makes all readers (especially PARI/GP users) generate good curves. Notice that the program output only a prime p, a coefficient A of Montgomery curve, the order L of $E(\mathbb{F}_p)$, and the order l of a base point. At the moment, it is necessary to manually find another coefficient B of Montgomery curve, generate a base point whose order is l, and check $L \neq p, p-1$.

This program is briefly explained. The line

```
e=ellinit([0,A,0,1,0]);
```

sets (Montgomery) elliptic curve $E : y^2 = x^3 + Ax^2 + x$ to `e`. The function `ellap(e,p)` outputs the trace of $E(\mathbb{F}_p)$. Thus, `Num1` is the order of $E(\mathbb{F}_p)$, and `Num2` is the order of the twist of $E(\mathbb{F}_p)$. `isprime` is a prime decision function. `write` is an output function to text.

By this program, the following elliptic curves are found.

```
\\Checking pseudo Mersenne prime
check_mer(p)={
  local(n,c,k);
  n=0;
  c=0;
  while(c==0,
    n++;
    if(2^(n-1)<=p && p<2^n,c=1);
  );
  if(2^n-p < sqrt(2^n),
    k=floor(log(2^n-p)/log(2)+1);
    return([n,k]),
    return(0));
}

\\Main program
{
  count=0;
  for(a=0,3,
  A=4*a+6;
  e=ellinit([0,A,0,1,0]);
  for(k=1,2^20,
    print([a,k,count]);
    p=2^256-k;
    if(isprime(p)==1,
      t=ellap(e,p);
      num1=p+1-t;
      num2=p+1+t;
      Num1=num1;
      Num2=num2;
      check1=0;
      check2=0;
      if(num1%2==0,num1=num1/2;check1=1);
      if(num1%2==0,num1=num1/2;check1=2);
      if(num1%2==0,num1=num1/2;check1=3);
      if(num1%2==0,num1=num1/2;check1=4);
      if(num2%2==0,num2=num2/2;check2=1);
      if(num2%2==0,num2=num2/2;check2=2);
      if(num2%2==0,num2=num2/2;check2=3);
      if(num2%2==0,num2=num2/2;check2=4);
      isprime_num1=isprime(num1);
      isprime_num2=isprime(num2);
      if(t!=0 && isprime_num1==1
              && isprime_num2==1
              && (check_mer(Num1)!=0
              || check_mer(Num2)!=0),
        if(check_mer(Num1)!=0,
          count++;
          write("iwin.txt",k","A","Num1","num1);
        );
        if(check_mer(Num2)!=0,
          count++;
          print("A="A);
          write("iwin.txt",k","A","Num2","num2);
        );
      );
    );
  );
  );
}
```

Figure 1: Proposed program to find elliptic curves suitable for ECDSA

1. $p = 2^{256} - 58097$,
   $E_{S1} : 638y^2 = x^3 + 10x^2 + x$,
   base point $P = (11, 2)$,
   $L = 2^{256} - k'$, where k' is 125-bit integer
   $k' = 25181363380428710453079967399017869328$,
   $l = L/16$, which is 252-bit prime.

2. $p = 2^{256} - 507225$,
   $E_{S2} : 82y^2 = x^3 + 18x^2 + x$,
   base point $P = (2, 1)$,
   $L = 2^{256} - k'$, where k' is 127-bit integer
   $k' = 134184981501621384111934924743103436264$,
   $l = L/8$, which is 253-bit prime.

3. $p = 2^{256} - 979077$,
   $E_{S3} : 3805y^2 = x^3 + 18x^2 + x$,
   base point $P = (20, 2)$,
   $L = 2^{256} - k'$, where k' is 126-bit integer
   $k' = 67240641251824776802983670794157366424$,
   $l = L/8$, which is 253-bit prime.

For the convenience of time, the authors set the search range to $k < 2^{20}$, however, if the search range is expanded, more appropriate elliptic curve may be found.

5.2 Proposed Modular Reduction

This section proposes an algorithm (Algorithm 5) [2] similar to Algorithm 8 for computing a reduction $\bmod l$ for a prime l such that $L = 2^m l$ is written as $L = 2^n - k,\ k < 2^{n/2}$.

Notice v, v_0 and k are multiples of 2^m. Thus, w is also a multiple of 2^m. As well,

$$x \text{ is a multiple of } 2^m. \tag{10}$$

Steps from 2 to 7 are same as Algorithm 8 and then we see

$$x = 2^m u \bmod 2^m l. \tag{11}$$

By (10) and (11), we see $x/2^m = u \bmod l$.

Note that $v/2^n$ in step 2, $w/2^n$ in step 4, and $x/2^n$ in step 7 are integer divisions and then they are performed by shift operations, and $2^m u$ in step 1 is also performed by a shift operation.

6 Conclusion

This paper searched three elliptic curves suitable for ECDSA. In these curves, not only the reduction $\bmod p$ but also the reduction $\bmod l$ can be computed at high speed, where p is of $\mathbb{F}_p$ and l is the order of a base point, and doubling is faster because of a coefficient of curves $A = 10$ or 18. ECDSA adopting the searched curves has the same security as ECDSA adopting Curve25519, and it can process faster.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 16K00188.

Proposed Algorithm 5

Input : integer l such that $2^m l = 2^n - k$ $(k < 2^{n/2})$,
integer $0 \le u \le (l-1)^2$
Output : $u \bmod l$

```
1. $v \leftarrow 2^m u$
2. $v_0 \leftarrow v \bmod 2^n,\ v_1 \leftarrow v/2^n$
3. $w \leftarrow v_1 k + v_0$
4. $w_0 \leftarrow w \bmod 2^n,\ w_1 \leftarrow w/2^n$
5. $x \leftarrow w_1 k + w_0$
6. **if** $x \ge 2^m l$ **then** $x \leftarrow x - 2^m l$
7. $y \leftarrow x/2^m$
8. **return** $y$
```

REFERENCES

[1] Daniel J. Bernstein. Curve25519: New Diffie-Hellman Speed Records. *Public Key Cryptography - PKC 2006*, LNCS 3958, pp. 207–228. Springer (2006).

[2] Daniel J. Bernstein and Tanja Lange. Explicit-Formulas Database. `https://hyperelliptic.org/EFD/`.

[3] Daniel J. Bernstein and Tanja Lange. Safecurves: Choosing Safe Curves for Elliptic-Curve Cryptography. `https://safecurves.cr.yp.to`.

[4] Henri Cohen and Gerhard Frey. *Handbook of Elliptic and Hyperelliptic Curve Cryptography*. Chapman and Hall/CRC (2005).

[5] ETSI. ETSI TS 103 097 v1.1.1 Intelligent Transport Systems (ITS); Security; Security Header and Certificate Formats (2013).

[6] Gerhard Frey and Hans-Georg Rück. A Remark Concerning M-Divisibility and the Discrete Logarithm in the Divisor Class Group of Curves. *Mathematics of Computation*, Vol. 62, No. 206, pp. 865–874 (1994).

[7] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone. *Handbook of Applied Cryptography*. CRC Press (1997).

[8] Peter L. Montgomery. Modular Multiplication Without Trial Division. *Mathematics of Computation*, Vol. 44, No. 170, pp. 519–521 (1985).

[9] Peter L. Montgomery. Speeding the Pollard and Elliptic Curve Methods of Factorization. *Mathematics of Computation*, Vol. 48, No. 177, pp. 243–264 (1987).

[10] NaCl: Networking and Cryptography Library. `https://pari.math.u-bordeaux.fr`.

[11] PARI/GP. `https://pari.math.u-bordeaux.fr`.

[12] S. Pohlig and M. Hellman. An Improved Algorithm for Computing Logarithms Over $GF(p)$ and Its Cryptographic Significance. *IEEE Transactions on Information Theory*, Vol. 24, No. 1, pp. 106–110 (1978).

[13] Hans-Georg Rück. On the Discrete Logarithm in the Divisor Class Group of Curves. *Mathematics of Computation*, Vol. 68, No. 226, pp. 805–806 (1999).

[14] Joseph. H. Silverman. *The Arithmetic of Elliptic Curves*. Springer-Verlag New York (1985).

[2] Although the authors do not know whether Algorithm 5 is already known, it may be already known because Algorithm 5 is almost same as Algorithm 8.