IWIN2016

International Workshop on Informatics

Proceedings of
International Workshop on Informatics

August 28-31, 2016
Riga, Latvia

Sponsored by

Informatics Society and Riga Technical University

IWIN2016

International Workshop on Informatics

Proceedings of
International Workshop onInformatics

August 28-31, 2016
Riga, Latvia

Informatics Society

1862
RIGA TECHNICAL
UNIVERSITY

Sponsored by

Informatics Society and Riga Technical University

Publication office:
Informatics Laboratory
3-41, Tsujimachi, Kitaku, Nagoya 462-0032, Japan
Publisher:
Tadanori Mizuno, President of Informatics Society
ISBN:
978-4-902523-39-3
Printed in Japan

Table of Contents

Keynote Speech 1

(9:10 - 9:55, Aug. 29)

Session 1: Life Systems

(Chair: Ryozo Kiyohara) (9:55 - 10:45, Aug. 29)

Session 2: Sensor Networks

(Chair: Yoshitaka Nakamura) (10:55 - 12:10, Aug. 29)

Keynote Speech 2

(13:30 - 14:15, Aug. 29)

Session 3: Intelligent Transportattion Systems

(Chair: Yoshitaka Nakamura) (14:25 - 16:05, Aug. 29)

Keynote Speech 3

(9:00 - 9:45, Aug. 30)

Session 4: Robots Application

(Chair: Yoshia Saito) (9:45 - 10:35, Aug. 30)

Session 5: Data Analysis

(Chair: Tomoya Kitani) (10:45 - 12:25, Aug. 30)

Panel Session: Next Generation Distributed System ー IoT/ M2M and Its Application ー

(13:30 - 14:15, Aug. 30)

Chair

- Prof. Norio Shiratori, Tohoku University/Waseda University

Panelists

- Prof. Katsuhiko Kaji, Aichi Institute of Technology
- Prof. Takuya Yoshihiro, Wakayama University
- Prof. Tomoki Yoshihisa, Osaka University

Session 6: Systems and Applications

(Chair: Takuya Yoshihiro) (14:15 - 15:30, Aug. 30)

Session 7: Data Models

(Chair: Tomoki Yoshihisa) (9:00 - 10:40, Aug. 31)

Session 8: Multimedia Systems

(Chair: Katsuhiko Kaji) (10:50 - 12:30, Aug. 31)

A Message from the

General Co-Chairs

It is our great pleasure to welcome all of you to Riga, Republic of Latvia., for the Tenth International Workshop on Informatics (IWIN 2016). This workshop has been held annually and sponsored by the Informatics Society. The first, second, third, fourth, fifth, sixth, seventh, eighth, and ninth workshops were held in Napoli, Italy, Wien, Austria, Hawaii, USA, Edinburgh, Scotland, Venice, Italy, Chamonix France, Stockholm, Sweden, Prague, Czech Republic, and Amsterdam, Netherlands, respectively. The first workshop was held in 2007. All of workshops were held in September.

In IWIN 2016, 26 papers have been accepted. Based on the papers, eight technical sessions have been organized in a single track format, which highlight the latest results in research areas such as mobile computing, networking, information system, and groupware and education systems. In addition, IWIN 2016 has three invited sessions from Dr. Tetsuo Nakakawaji of Mitsubishi Electric Corp Information Technology R & D Center, from Prof. Agris Nikitenko of Riga Technical University, and from Prof. Fusako Kusunoki of Tama Art University. We really appreciate the participation of the three invited speakers in this workshop.

IWIN 2016 is co-sponsored by Riga Technical University.

We would like to thank all of 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 all will experience a great and enjoyable meeting in Riga.

Toru Hasegawa

Toru Hasegawa, and Kozo Okano
General Co-Chairs
The International Workshop on Informatics 2016

Organizing Committee

General Chair

Toru Hasegawa (Osaka University, Japan)
Kozo Okano (Shinshu University, Japan)

Steering Committee

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 College, Japan)
Ken-ichi Okada (Keio University, Japan)
Norio Shiratori (Waseda University, Japan)
Osamu Takahashi (Future University Hakodate, Japan)

Program Chair

Tomoki Yoshihisa (Osaka University, Japan)

Financial Chair

Tomoya Kitani (Shizuoka University, Japan)

Publicity Chair

Yoshitaka Nakamura (Future University Hakodate, Japan)

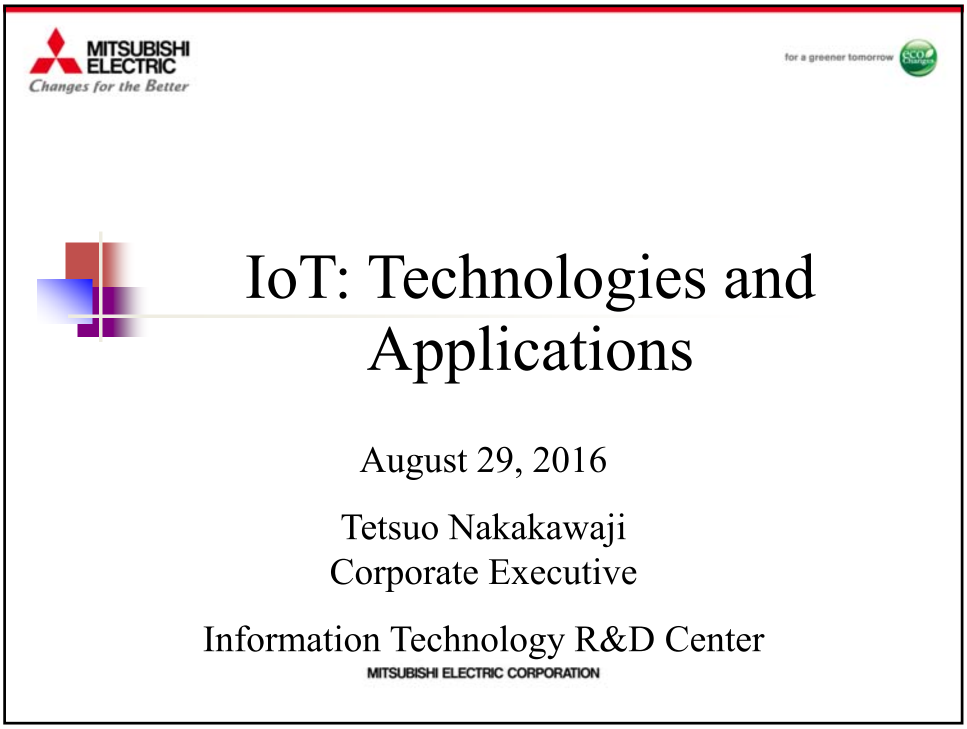
Program Committee

Keiichi Abe (Kanagawa Institute of Technology, Japan)
Chiaki Doi (NTT DOCOMO, Inc., Japan)
Teruyuki Hasegawa (KDDI R&D Laboratories, Japan)
Hiroshi Horikawa (Mitsubishi Electric Information Network Corporation, Japan)
Tomoo Inoue (University of Tsukuba, Japan)
Katsuhiko Kaji (Aichi Institute of Technology, Japan)
Tomoya Kitani (Shizuoka University, Japan)
Minoru Kobayashi (Meiji University, Japan)
Hiroshi Mineno (Shizuoka University, Japan)
Shinichiro Mori (Chiba Institute of Technology, Japan)
Yoshitoshi Murata (Iwate Prefectural University, Japan)
Yoshitaka Nakamura (Future University Hakodate, Japan)
Kozo Okano (Shinshu University, Japan)
Masashi Saito (Kanazawa Institute of Technology, Japan)
Tetsuya Shigeyasu (Prefectural University of Hiroshima, Japan)
Hirosato Tsuji (Information-technology Promotion Agency, Japan)
Hirozumi Yamaguchi (Osaka University, Japan)
Tomoki Yoshihisa (Osaka University, Japan)
Naoya Chujo (Aichi Institute of Technology, Japan)
Yu Enokibori (Nagoya University, Japan)
Takaaki Hishida (Aichi Institute of Technology, Japan)
Yusuke Ichikawa (NTT Corporation, Japan)
Hiroshi Inamura (Future University Hakodate, Japan)
Hiroshi Ishikawa (Tokyo Metropolitan University, Japan)
Yoshinobu Kawabe (Aichi Institute of Technology, Japan)
Ryozo Kiyohara (Kanagawa Institute of Technology, Japan)
Tsukasa Kudo (Shizuoka Institute of Science and Technology, Japan)
Hiroaki Morino (Shibaura Institute of Technology, Japan)
Katsuhiro Naito (Aichi Institute of Technology, Japan)
Ken Ohta (NTT DOCOMO, Inc., Japan)
Yoshia Saito (Iwate Prefectural University, Japan)
Fumiaki Sato (Toho University, Japan)
Hideyuki Takahashi (Tohoku University, Japan)
Takaaki Umedu (Shiga University, Japan)
Yuji Wada (Tokyo Denki University, Japan)
Takuya Yoshihiro (Wakayama University, Japan)
Takaya Yuizono (Japan Advanced Institute of Science and Technology, Japan)

Keynote Speech 1：

Dr. Tetsuo Nakakawaji

（Corporate Executive / Senior General Manager of Mitsubishi Electric Corporation Information Technology R&D Center）



IoT: Technologies and Applications

August 29, 2016

Tetsuo Nakakawaji
Corporate Executive

Information Technology R&D Center
MITSUBISHI ELECTRIC CORPORATION

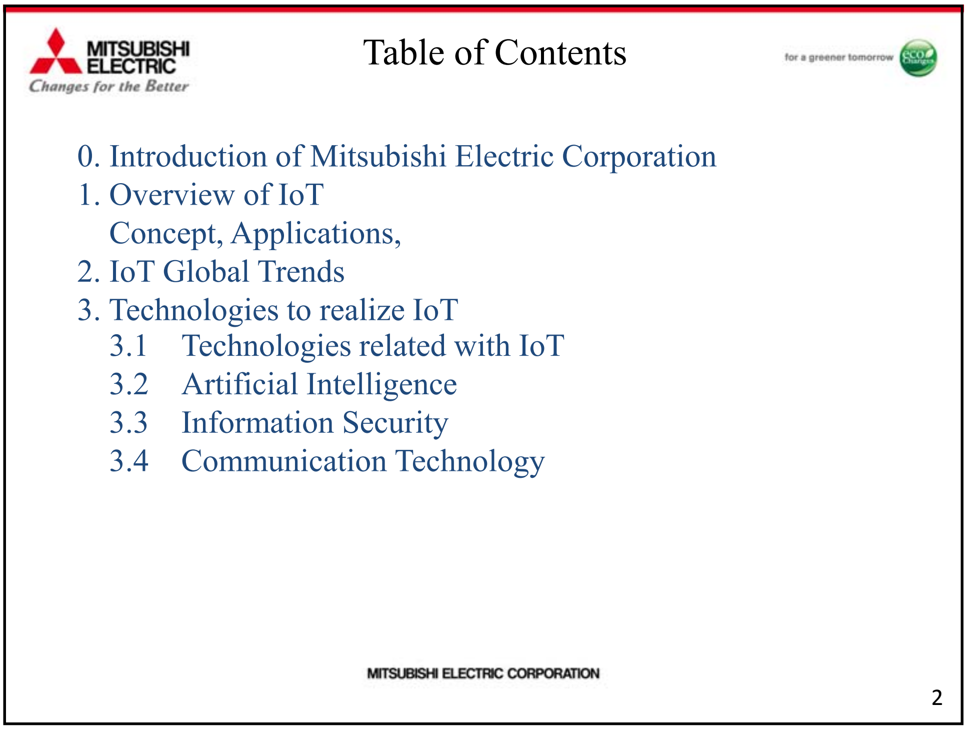
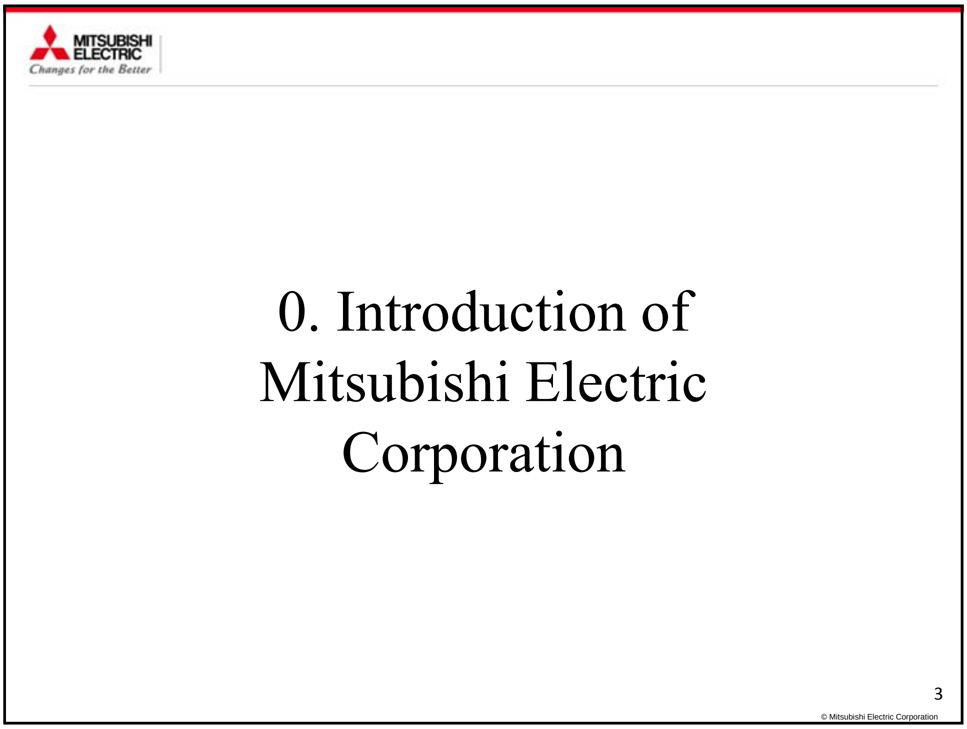


Table of Contents

0. Introduction of Mitsubishi Electric Corporation
1. Overview of IoT
   Concept, Applications,
2. IoT Global Trends
3. Technologies to realize IoT
   3.1 Technologies related with IoT
   3.2 Artificial Intelligence
   3.3 Information Security
   3.4 Communication Technology

MITSUBISHI ELECTRIC CORPORATION

2



0. Introduction of Mitsubishi Electric Corporation

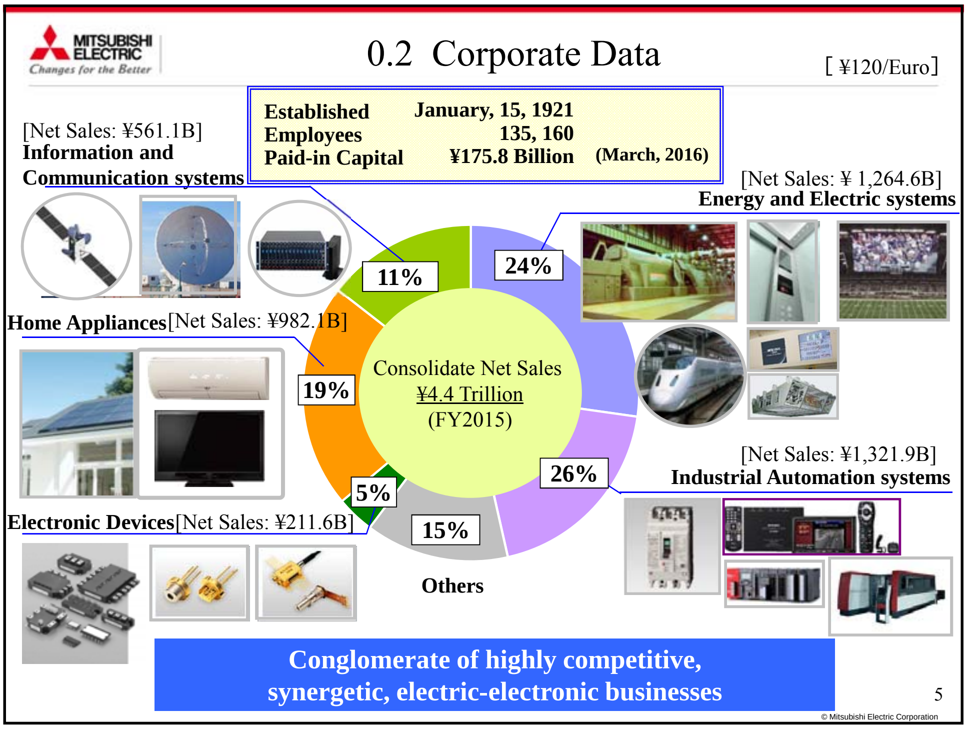
3

© Mitsubishi Electric Corporation

0.1 Introduction～12 business areas～

- Space systems
- Public systems
- Energy systems
- Information & Comm. Systems
- Factory Automation Systems
- Building Systems
- Transportation Systems
- Home Products
- IT Solution
- Semiconductors & Devices
- Air Conditioning Systems
- Automotive Equipment

© Mitsubishi Electric Corporation



MITSUBISHI ELECTRIC *Changes for the Better*

0.2 Corporate Data

[¥120/Euro]

Established **January, 15, 1921**
Employees **135, 160**
Paid-in Capital **¥175.8 Billion** **(March, 2016)**

[Net Sales: ¥561.1B]
Information and Communication systems

[Net Sales: ¥ 1,264.6B]
Energy and Electric systems

Home Appliances[Net Sales: ¥982.1B]

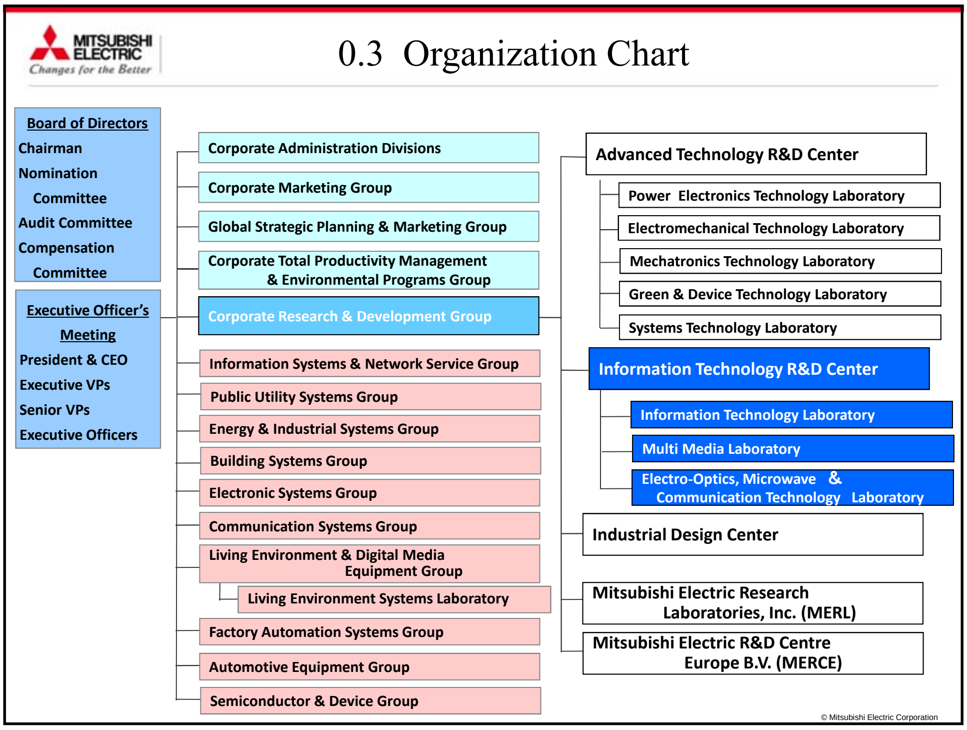
[Net Sales: ¥1,321.9B]
Industrial Automation systems

Electronic Devices[Net Sales: ¥211.6B]

Conglomerate of highly competitive, synergetic, electric-electronic businesses

5

© Mitsubishi Electric Corporation



MITSUBISHI ELECTRIC *Changes for the Better*

0.3 Organization Chart

Board of Directors
Chairman
Nomination Committee
Audit Committee
Compensation Committee

Executive Officer's Meeting
President & CEO
Executive VPs
Senior VPs
Executive Officers

Corporate Administration Divisions
Corporate Marketing Group
Global Strategic Planning & Marketing Group
Corporate Total Productivity Management & Environmental Programs Group
Corporate Research & Development Group
Information Systems & Network Service Group
Public Utility Systems Group
Energy & Industrial Systems Group
Building Systems Group
Electronic Systems Group
Communication Systems Group
Living Environment & Digital Media Equipment Group
Living Environment Systems Laboratory
Factory Automation Systems Group
Automotive Equipment Group
Semiconductor & Device Group

Advanced Technology R&D Center
Power Electronics Technology Laboratory
Electromechanical Technology Laboratory
Mechatronics Technology Laboratory
Green & Device Technology Laboratory
Systems Technology Laboratory

Information Technology R&D Center
Information Technology Laboratory
Multi Media Laboratory
Electro-Optics, Microwave & Communication Technology Laboratory

Industrial Design Center

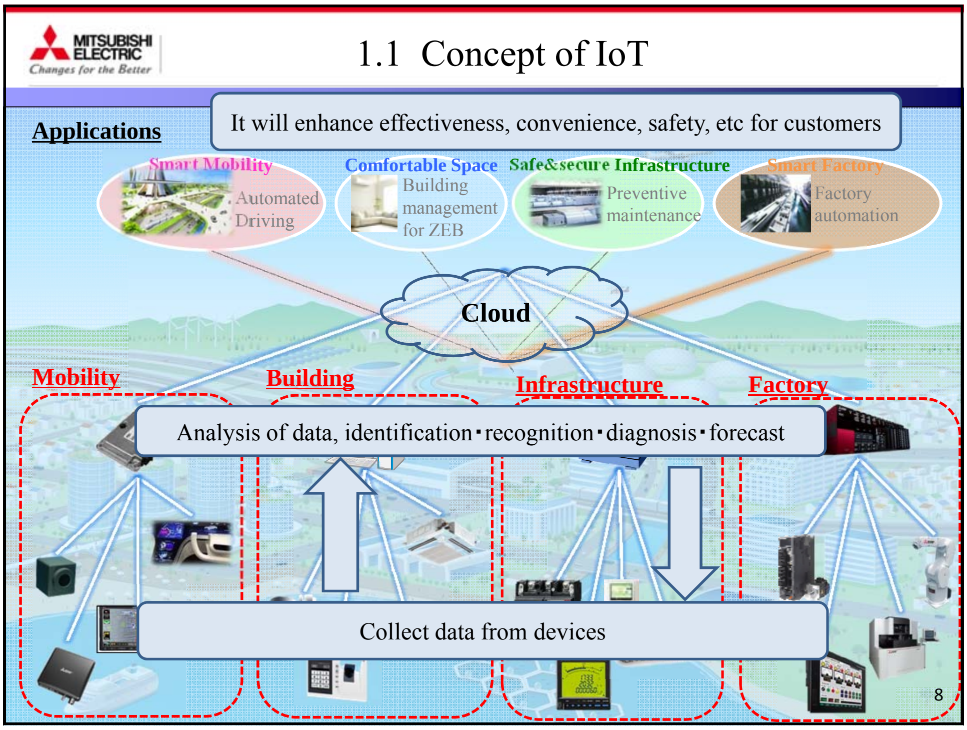
Mitsubishi Electric Research Laboratories, Inc. (MERL)

Mitsubishi Electric R&D Centre Europe B.V. (MERCE)

© Mitsubishi Electric Corporation

1. Overview of IoT (Internet of Things)

7



1.1 Concept of IoT

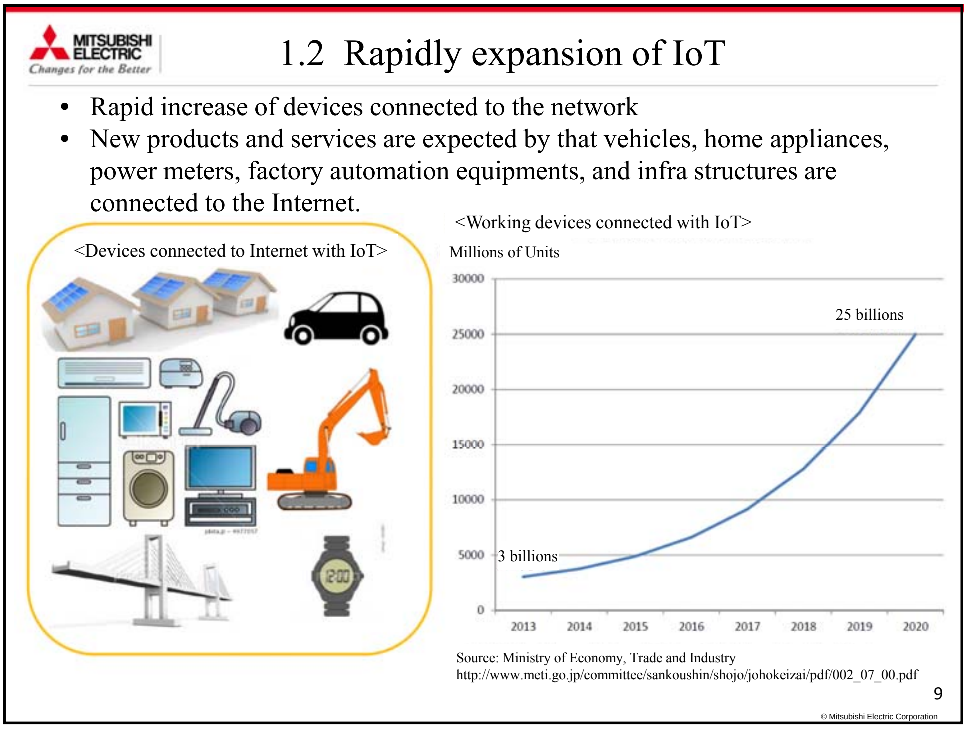
Applications

It will enhance effectiveness, convenience, safety, etc for customers

Analysis of data, identification･recognition･diagnosis･forecast

Collect data from devices

8



1.2 Rapidly expansion of IoT

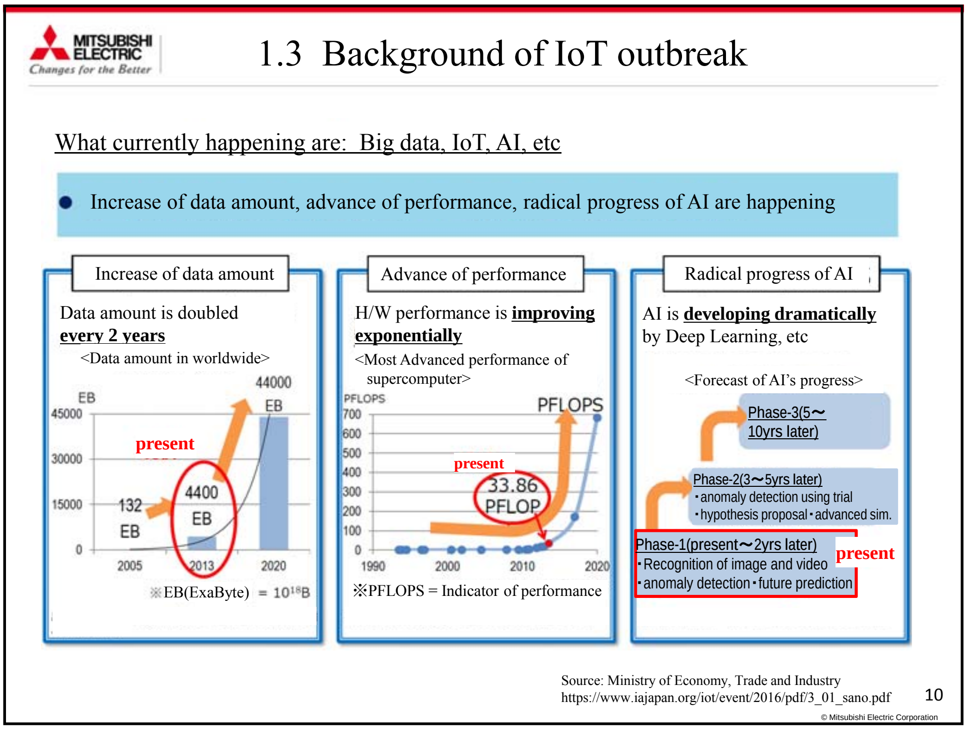
- Rapid increase of devices connected to the network
- New products and services are expected by that vehicles, home appliances, power meters, factory automation equipments, and infra structures are connected to the Internet.

<Devices connected to Internet with IoT>

<Working devices connected with IoT>

Source: Ministry of Economy, Trade and Industry
http://www.meti.go.jp/committee/sankoushin/shojo/johokeizai/pdf/002_07_00.pdf

9



1.3 Background of IoT outbreak

What currently happening are: Big data, IoT, AI, etc

- Increase of data amount, advance of performance, radical progress of AI are happening

Increase of data amount

Data amount is doubled **every 2 years**

<Data amount in worldwide>

※EB(ExaByte) = 10^{18}B

Advance of performance

H/W performance is **improving exponentially**

<Most Advanced performance of supercomputer>

※PFLOPS = Indicator of performance

Radical progress of AI

AI is **developing dramatically** by Deep Learning, etc

<Forecast of AI's progress>

Phase-3(5～10yrs later)

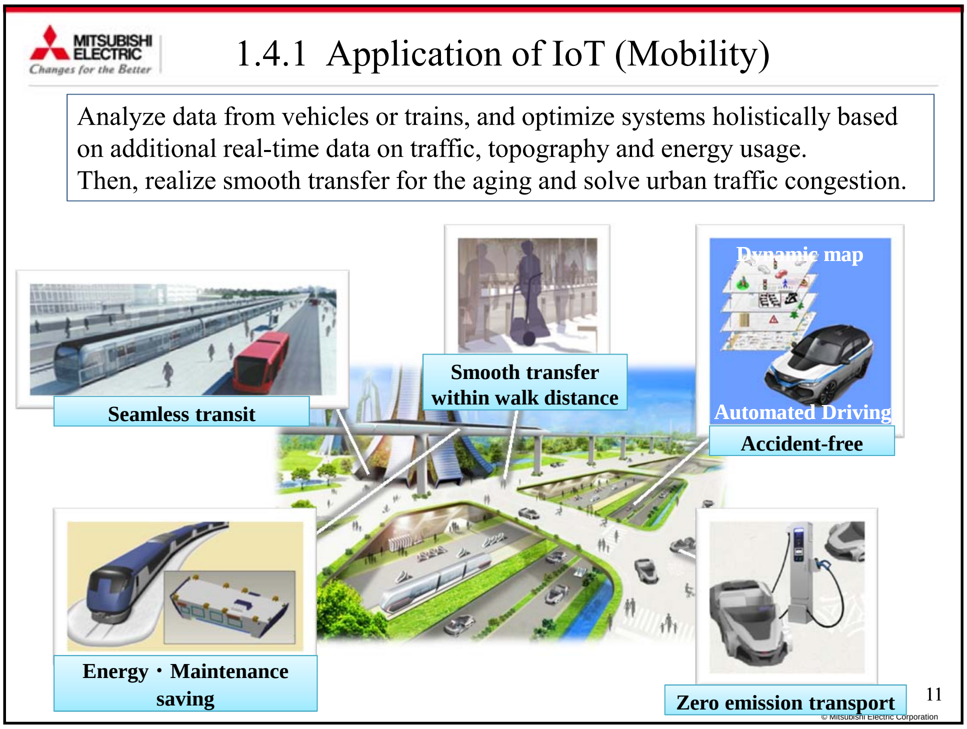
Phase-2(3～5yrs later)
・anomaly detection using trial
・hypothesis proposal・advanced sim.

Phase-1(present～2yrs later)
・Recognition of image and video
・anomaly detection・future prediction

present

Source: Ministry of Economy, Trade and Industry
https://www.iajapan.org/iot/event/2016/pdf/3_01_sano.pdf

10



1.4.1 Application of IoT (Mobility)

Analyze data from vehicles or trains, and optimize systems holistically based on additional real-time data on traffic, topography and energy usage.
Then, realize smooth transfer for the aging and solve urban traffic congestion.

Seamless transit

Smooth transfer within walk distance

Dynamic map

Automated Driving

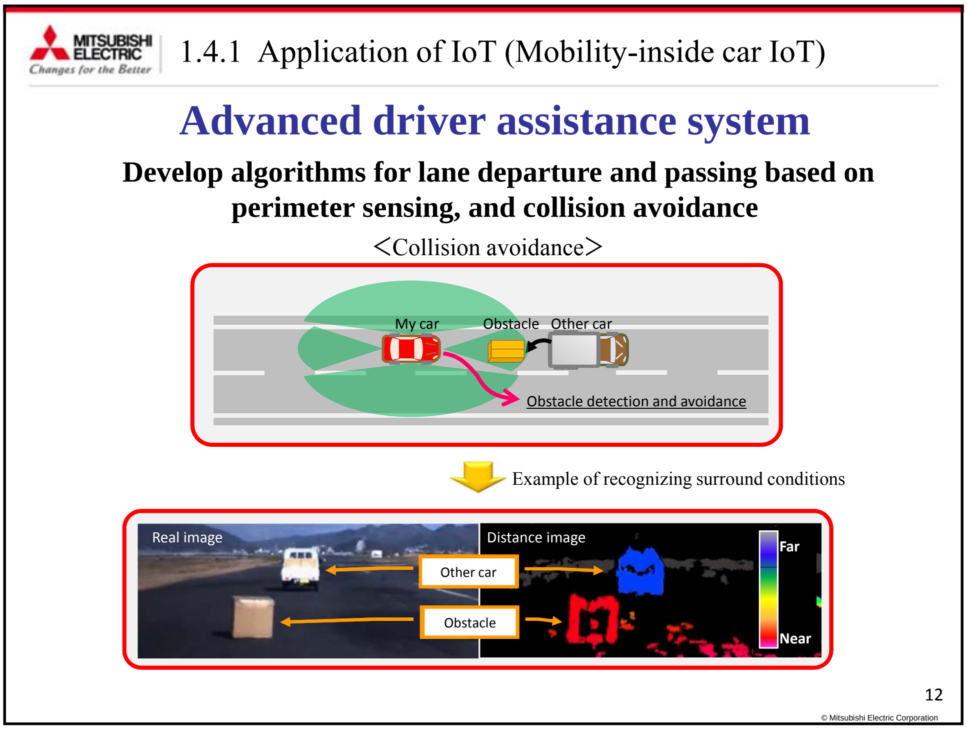
Accident-free

Energy・Maintenance saving

Zero emission transport

11

© Mitsubishi Electric Corporation



1.4.1 Application of IoT (Mobility-inside car IoT)

Advanced driver assistance system

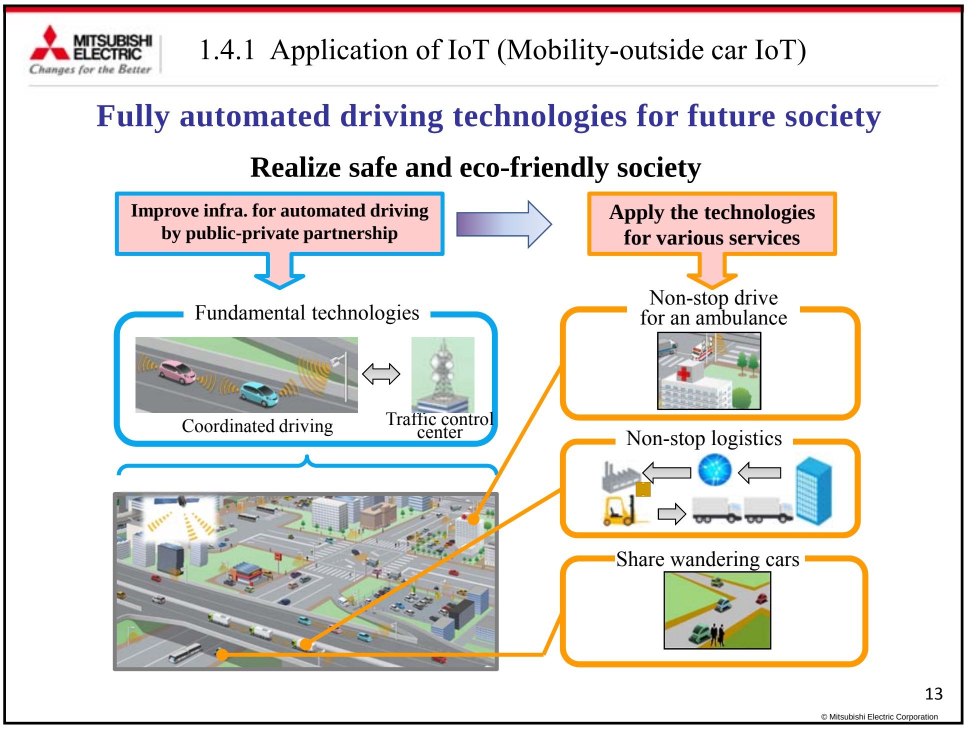
Develop algorithms for lane departure and passing based on perimeter sensing, and collision avoidance

＜Collision avoidance＞

Example of recognizing surround conditions

12

© Mitsubishi Electric Corporation



1.4.1 Application of IoT (Mobility-outside car IoT)

Fully automated driving technologies for future society

Realize safe and eco-friendly society

Improve infra. for automated driving by public-private partnership

Apply the technologies for various services

Fundamental technologies

Coordinated driving

Traffic control center

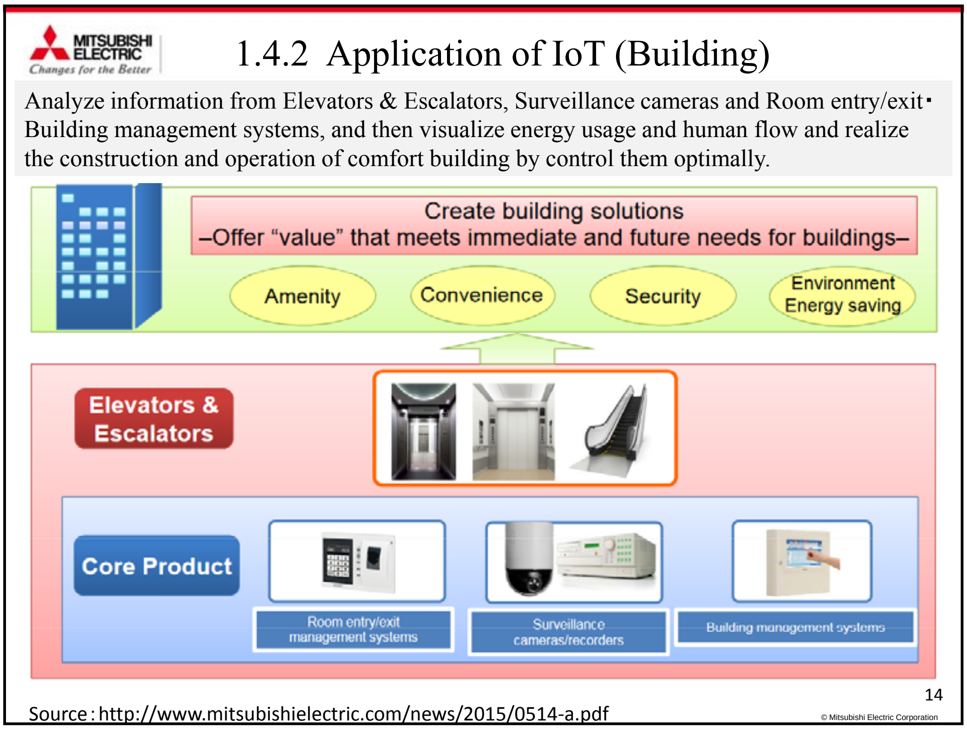
Non-stop drive for an ambulance

Non-stop logistics

Share wandering cars

13

© Mitsubishi Electric Corporation



1.4.2 Application of IoT (Building)

Analyze information from Elevators & Escalators, Surveillance cameras and Room entry/exit･Building management systems, and then visualize energy usage and human flow and realize the construction and operation of comfort building by control them optimally.

Create building solutions
–Offer “value” that meets immediate and future needs for buildings–

Amenity

Convenience

Security

Environment Energy saving

Elevators & Escalators

Core Product

Room entry/exit management systems

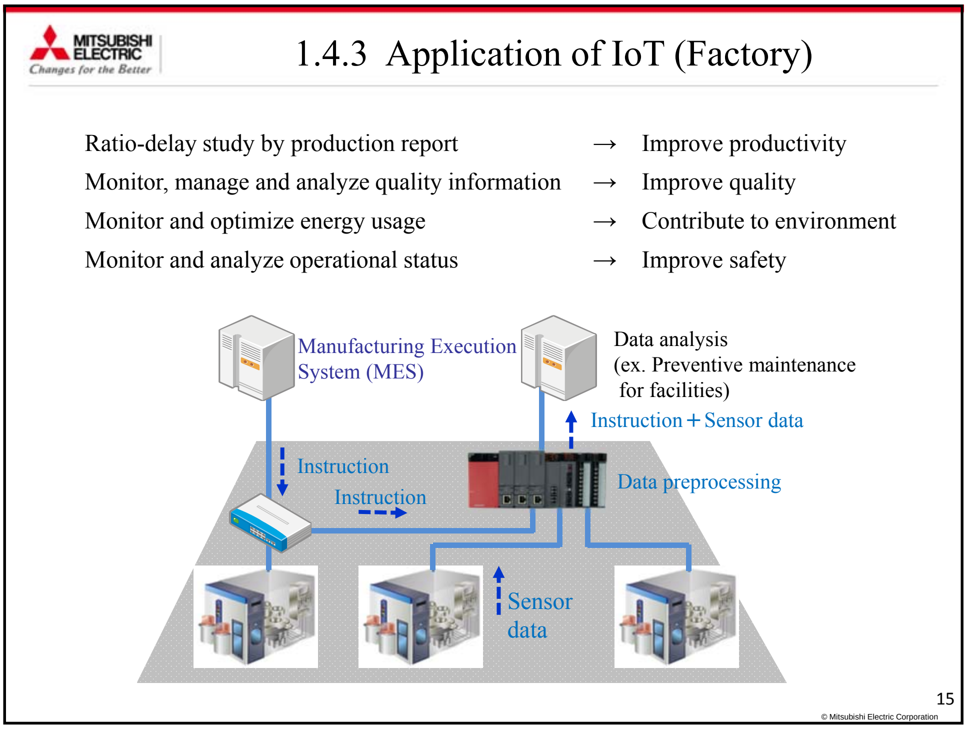
Surveillance cameras/recorders

Building management systems

14

Source：http://www.mitsubishielectric.com/news/2015/0514-a.pdf

© Mitsubishi Electric Corporation



1.4.3 Application of IoT (Factory)

Ratio-delay study by production report → Improve productivity

Monitor, manage and analyze quality information → Improve quality

Monitor and optimize energy usage → Contribute to environment

Monitor and analyze operational status → Improve safety

Manufacturing Execution System (MES)

Data analysis
(ex. Preventive maintenance
for facilities)

Instruction＋Sensor data

Instruction

Instruction

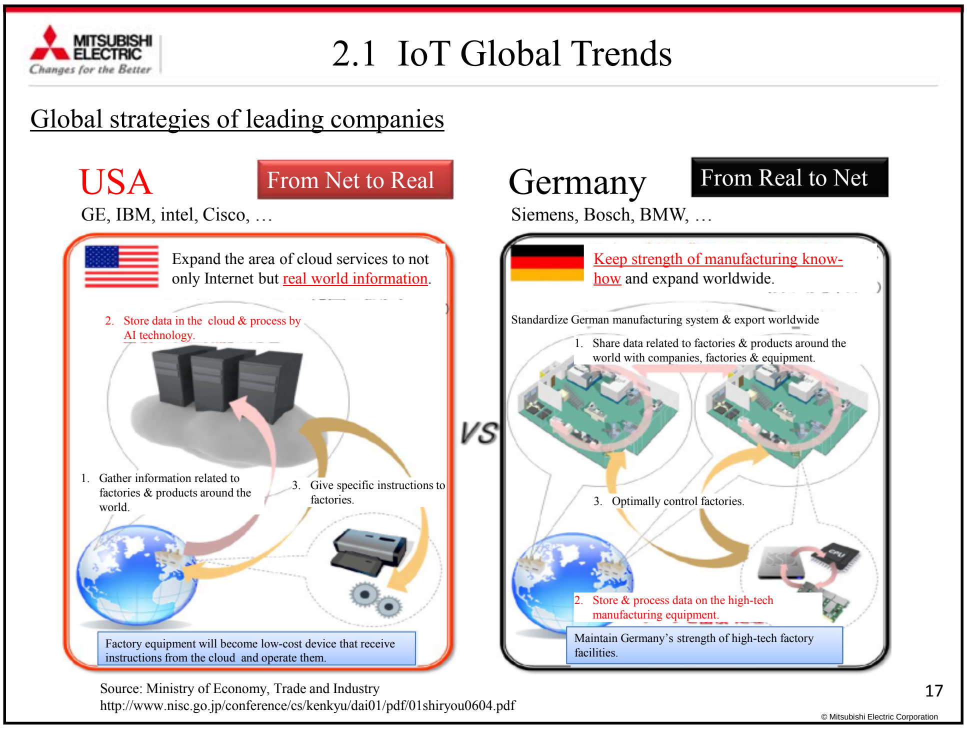
Data preprocessing

Sensor data

15

2. IoT Global Trends

16



MITSUBISHI ELECTRIC
Changes for the Better

2.1 IoT Global Trends

Global strategies of leading companies

USA — From Net to Real

GE, IBM, intel, Cisco, …

Expand the area of cloud services to not only Internet but real world information.

2. Store data in the cloud & process by AI technology.

1. Gather information related to factories & products around the world.

3. Give specific instructions to factories.

Factory equipment will become low-cost device that receive instructions from the cloud and operate them.

VS

Germany — From Real to Net

Siemens, Bosch, BMW, …

Keep strength of manufacturing know-how and expand worldwide.

Standardize German manufacturing system & export worldwide

1. Share data related to factories & products around the world with companies, factories & equipment.

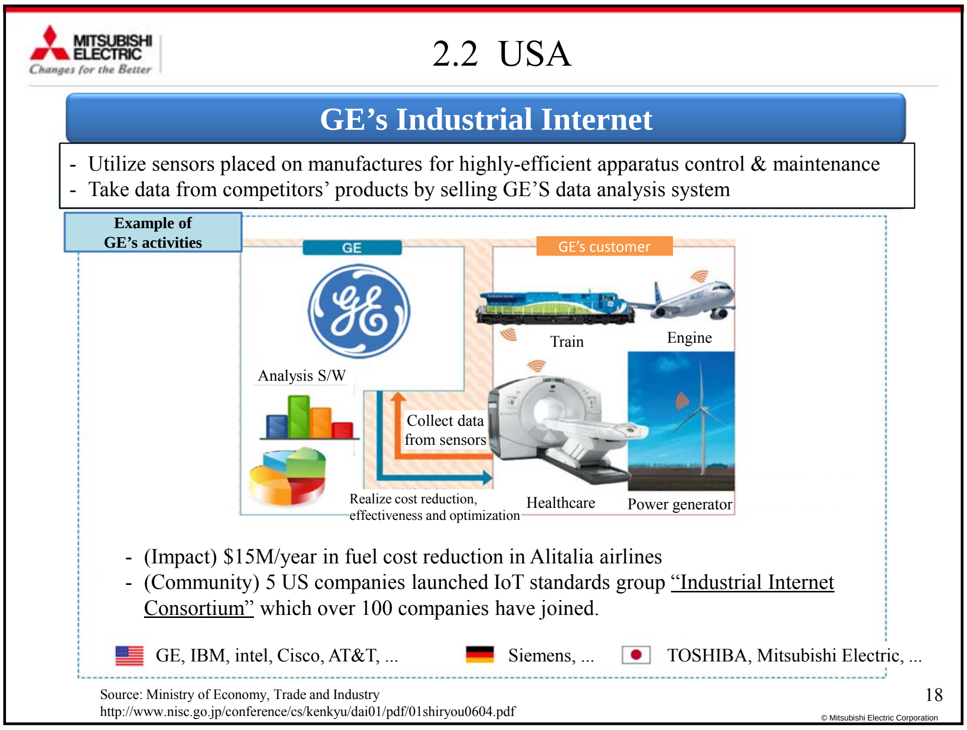
3. Optimally control factories.

2. Store & process data on the high-tech manufacturing equipment.

Maintain Germany's strength of high-tech factory facilities.

Source: Ministry of Economy, Trade and Industry
http://www.nisc.go.jp/conference/cs/kenkyu/dai01/pdf/01shiryou0604.pdf

17



MITSUBISHI ELECTRIC
Changes for the Better

2.2 USA

GE's Industrial Internet

- Utilize sensors placed on manufactures for highly-efficient apparatus control & maintenance
- Take data from competitors' products by selling GE'S data analysis system

Example of GE's activities

GE

Analysis S/W

Collect data from sensors

Realize cost reduction, effectiveness and optimization

GE's customer

Train

Engine

Healthcare

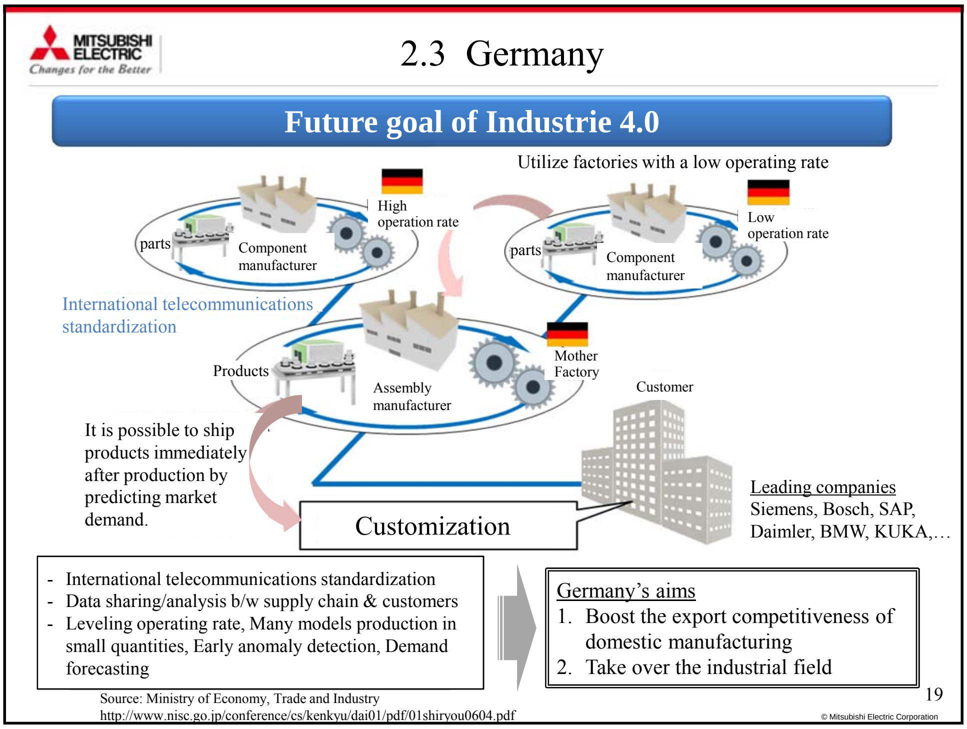
Power generator

- (Impact) $15M/year in fuel cost reduction in Alitalia airlines
- (Community) 5 US companies launched IoT standards group "Industrial Internet Consortium" which over 100 companies have joined.

GE, IBM, intel, Cisco, AT&T, ... Siemens, ... TOSHIBA, Mitsubishi Electric, ...

Source: Ministry of Economy, Trade and Industry
http://www.nisc.go.jp/conference/cs/kenkyu/dai01/pdf/01shiryou0604.pdf

18



MITSUBISHI ELECTRIC Changes for the Better

2.3 Germany

Future goal of Industrie 4.0

It is possible to ship products immediately after production by predicting market demand.

Customization

Leading companies
Siemens, Bosch, SAP, Daimler, BMW, KUKA,…

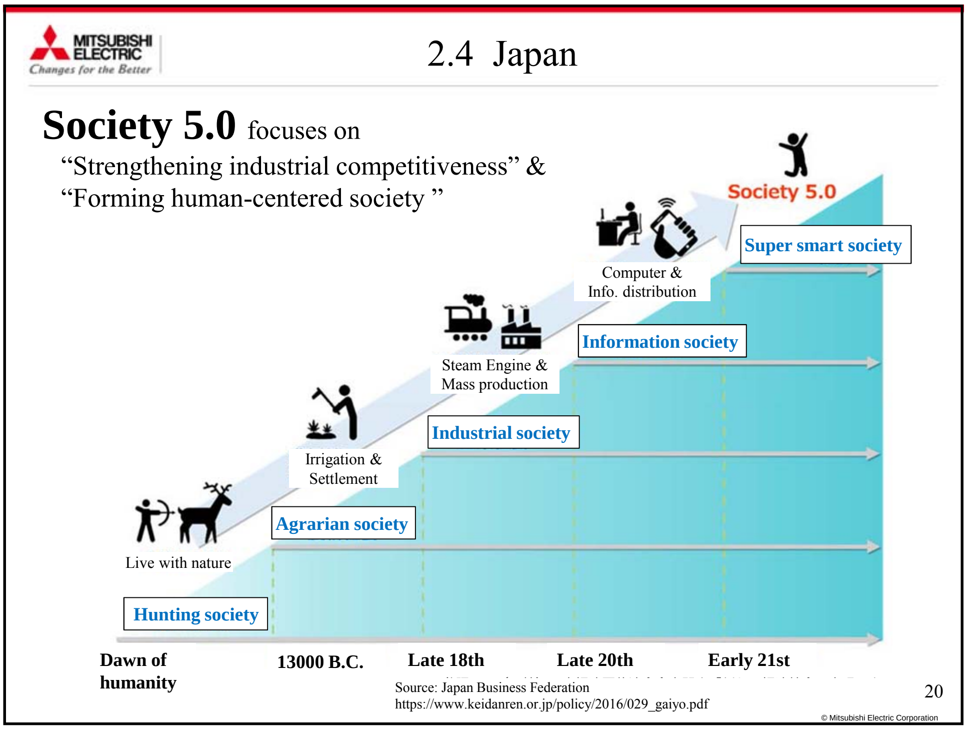
- International telecommunications standardization
- Data sharing/analysis b/w supply chain & customers
- Leveling operating rate, Many models production in small quantities, Early anomaly detection, Demand forecasting

Germany's aims
1. Boost the export competitiveness of domestic manufacturing
2. Take over the industrial field

Source: Ministry of Economy, Trade and Industry
http://www.nisc.go.jp/conference/cs/kenkyu/dai01/pdf/01shiryou0604.pdf

19

© Mitsubishi Electric Corporation



MITSUBISHI ELECTRIC Changes for the Better

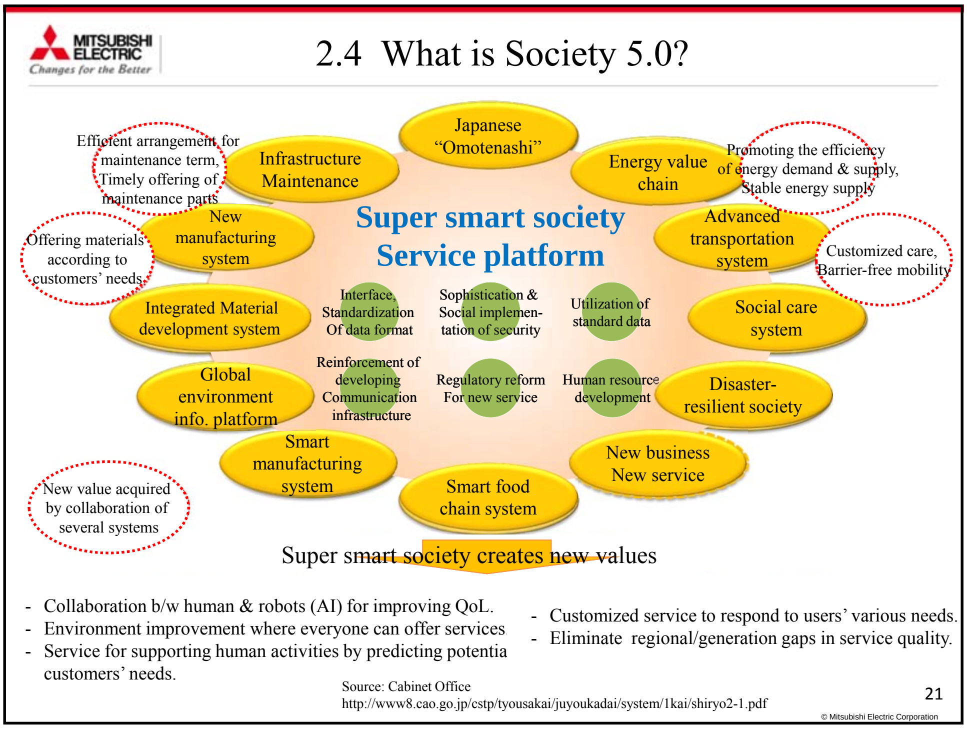
2.4 Japan

Society 5.0 focuses on
"Strengthening industrial competitiveness" &
"Forming human-centered society "

Source: Japan Business Federation
https://www.keidanren.or.jp/policy/2016/029_gaiyo.pdf

20

© Mitsubishi Electric Corporation



2.4 What is Society 5.0?

Super smart society
Service platform

Japanese "Omotenashi"

Infrastructure Maintenance

Energy value chain

New manufacturing system

Advanced transportation system

Integrated Material development system

Social care system

Global environment info. platform

Disaster-resilient society

Smart manufacturing system

Smart food chain system

New business New service

Interface, Standardization Of data format

Sophistication & Social implemen-tation of security

Utilization of standard data

Reinforcement of developing Communication infrastructure

Regulatory reform For new service

Human resource development

Efficient arrangement for maintenance term, Timely offering of maintenance parts

Promoting the efficiency of energy demand & supply, Stable energy supply

Offering materials according to customers' needs

Customized care, Barrier-free mobility

New value acquired by collaboration of several systems

Super smart society creates new values

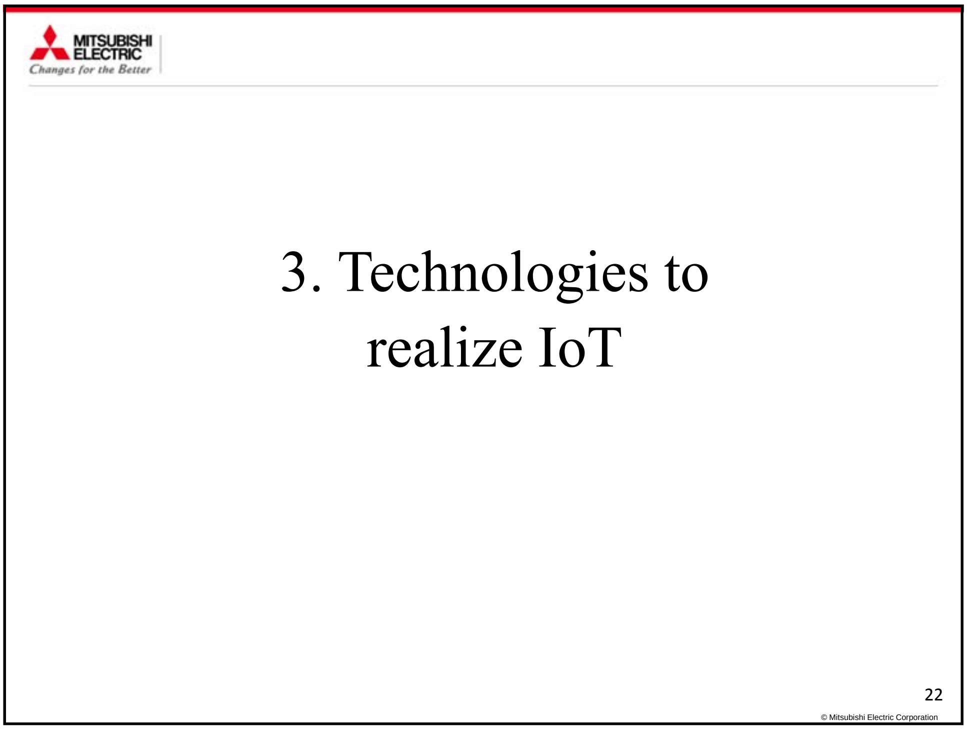
- Collaboration b/w human & robots (AI) for improving QoL.
- Environment improvement where everyone can offer services
- Service for supporting human activities by predicting potentia customers' needs.

- Customized service to respond to users' various needs.
- Eliminate regional/generation gaps in service quality.

Source: Cabinet Office
http://www8.cao.go.jp/cstp/tyousakai/juyoukadai/system/1kai/shiryo2-1.pdf

21

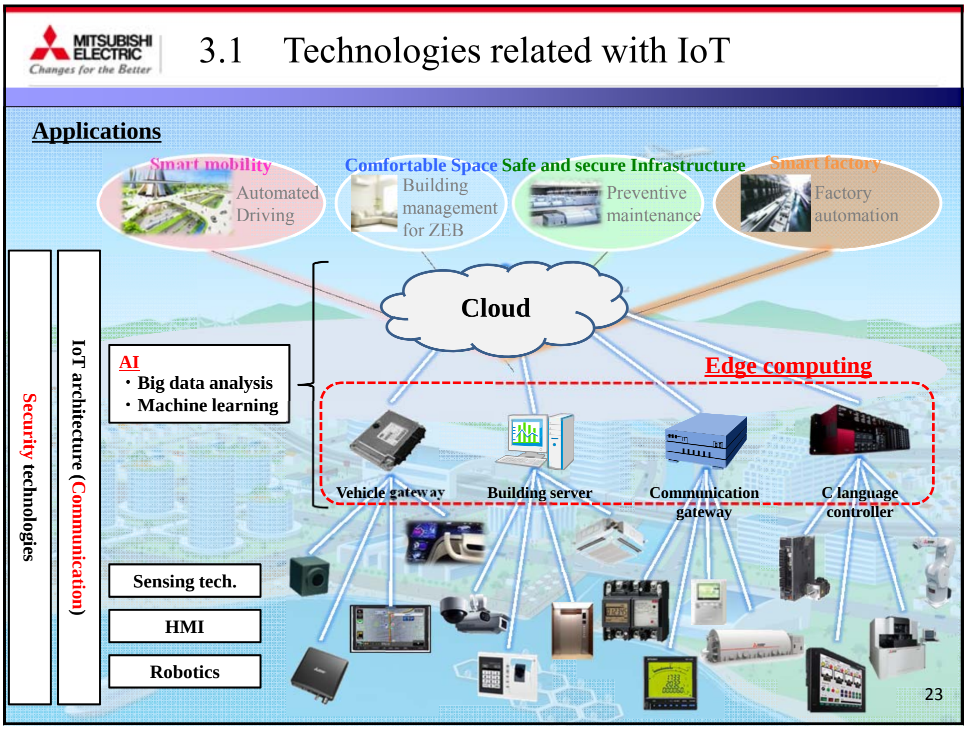
© Mitsubishi Electric Corporation



3. Technologies to realize IoT

22

© Mitsubishi Electric Corporation



3.1 Technologies related with IoT

Applications

Smart mobility: Automated Driving

Comfortable Space: Building management for ZEB

Safe and secure Infrastructure: Preventive maintenance

Smart factory: Factory automation

Cloud

Edge computing

AI
・Big data analysis
・Machine learning

Vehicle gateway, Building server, Communication gateway, C language controller

Security technologies

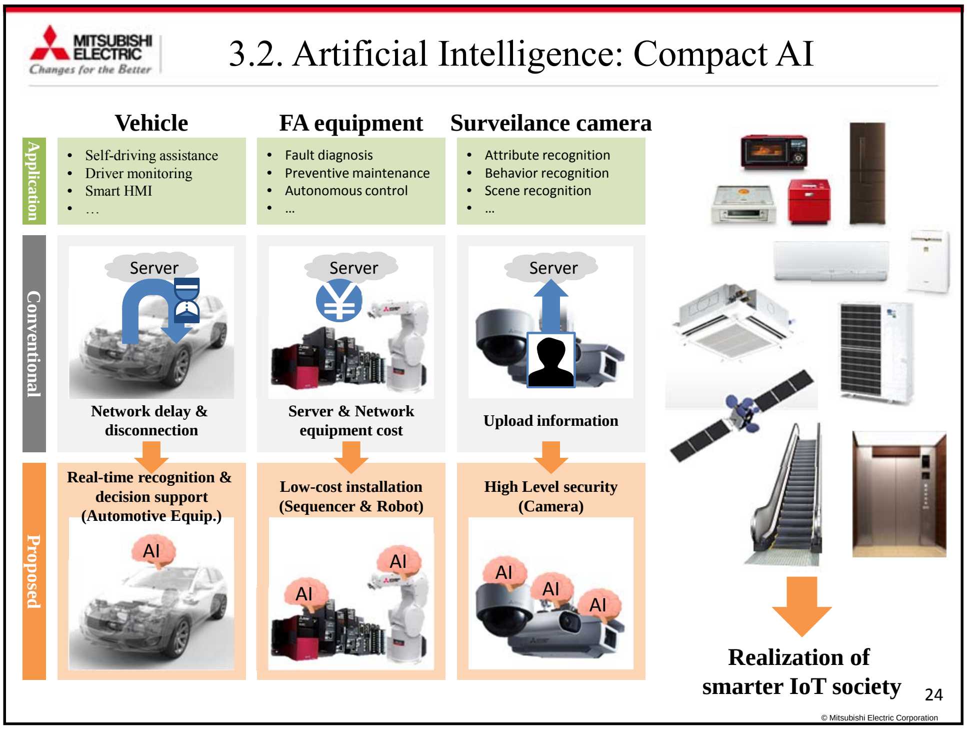
IoT architecture (Communication)

Sensing tech.

HMI

Robotics

23



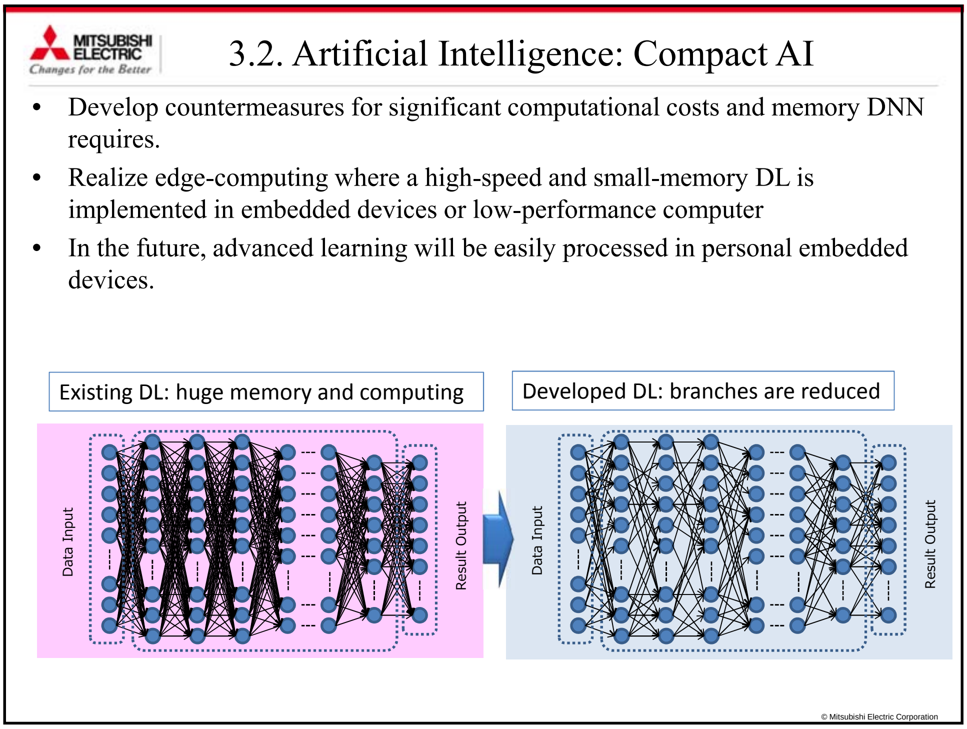
3.2. Artificial Intelligence: Compact AI

	Vehicle	FA equipment	Surveilance camera
Application	Self-driving assistance Driver monitoring Smart HMI …	Fault diagnosis Preventive maintenance Autonomous control …	Attribute recognition Behavior recognition Scene recognition …
Conventional	Server Network delay & disconnection	Server Server & Network equipment cost	Server Upload information
Proposed	Real-time recognition & decision support (Automotive Equip.) AI	Low-cost installation (Sequencer & Robot) AI AI	High Level security (Camera) AI AI AI

Realization of smarter IoT society

24

© Mitsubishi Electric Corporation



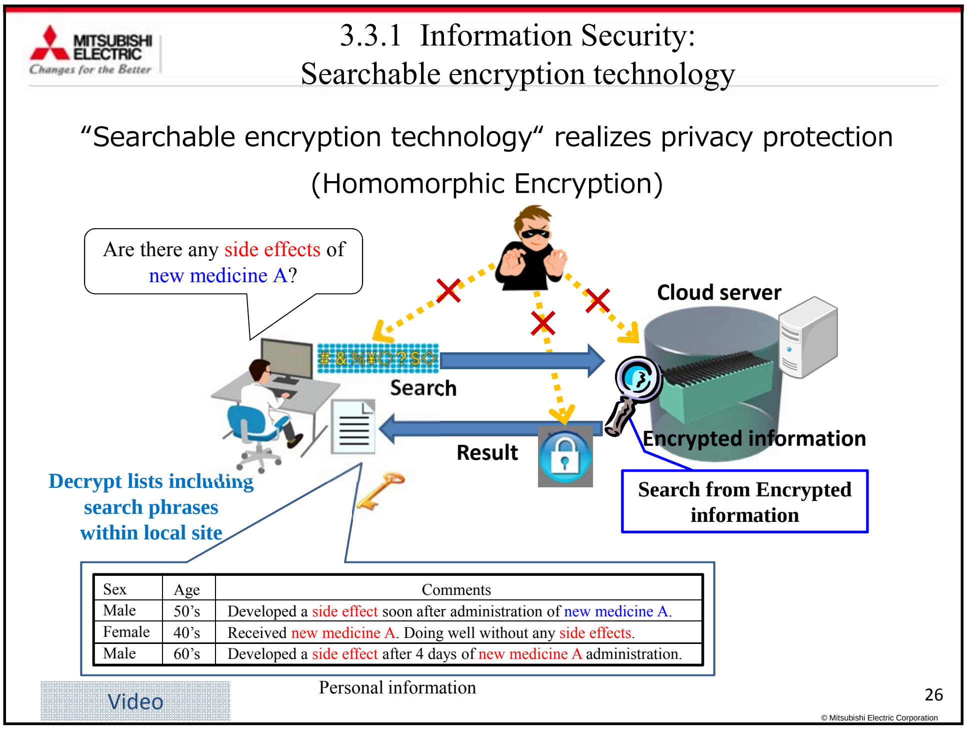
3.2. Artificial Intelligence: Compact AI

- Develop countermeasures for significant computational costs and memory DNN requires.
- Realize edge-computing where a high-speed and small-memory DL is implemented in embedded devices or low-performance computer
- In the future, advanced learning will be easily processed in personal embedded devices.

Existing DL: huge memory and computing

Developed DL: branches are reduced

© Mitsubishi Electric Corporation



3.3.1 Information Security: Searchable encryption technology

“Searchable encryption technology“ realizes privacy protection (Homomorphic Encryption)

Are there any side effects of new medicine A?

Search from Encrypted information

Decrypt lists including search phrases within local site

Sex	Age	Comments
Male	50’s	Developed a side effect soon after administration of new medicine A.
Female	40’s	Received new medicine A. Doing well without any side effects.
Male	60’s	Developed a side effect after 4 days of new medicine A administration.

Personal information

Video

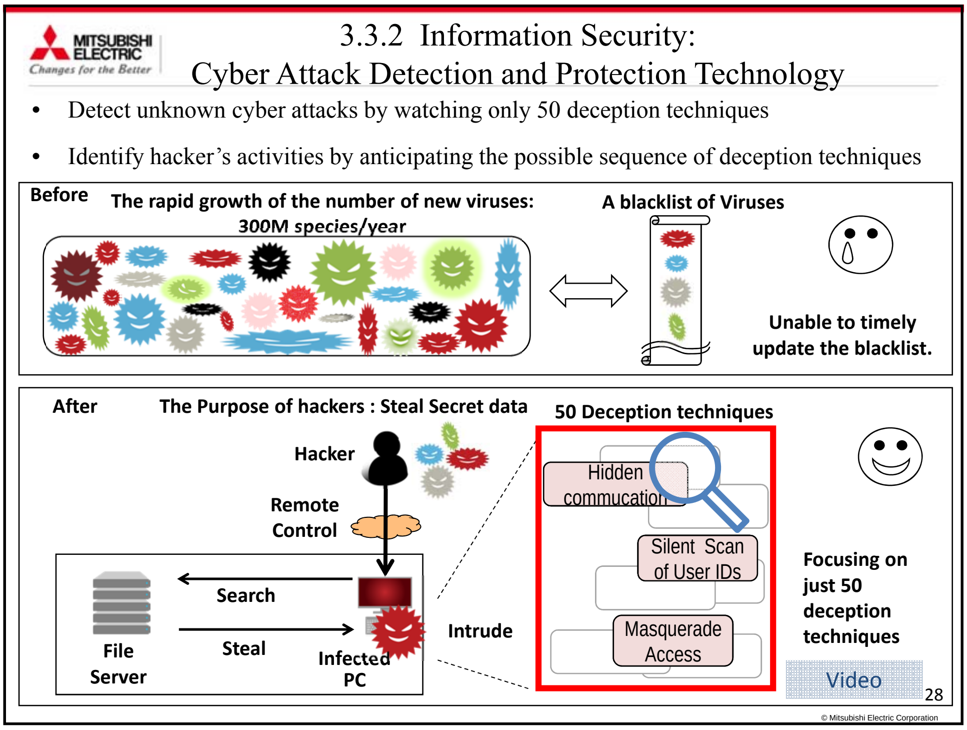
26

© Mitsubishi Electric Corporation

3.3.1 Searchable encryption technology

Demonstration video:
Searchable encryption technology

27

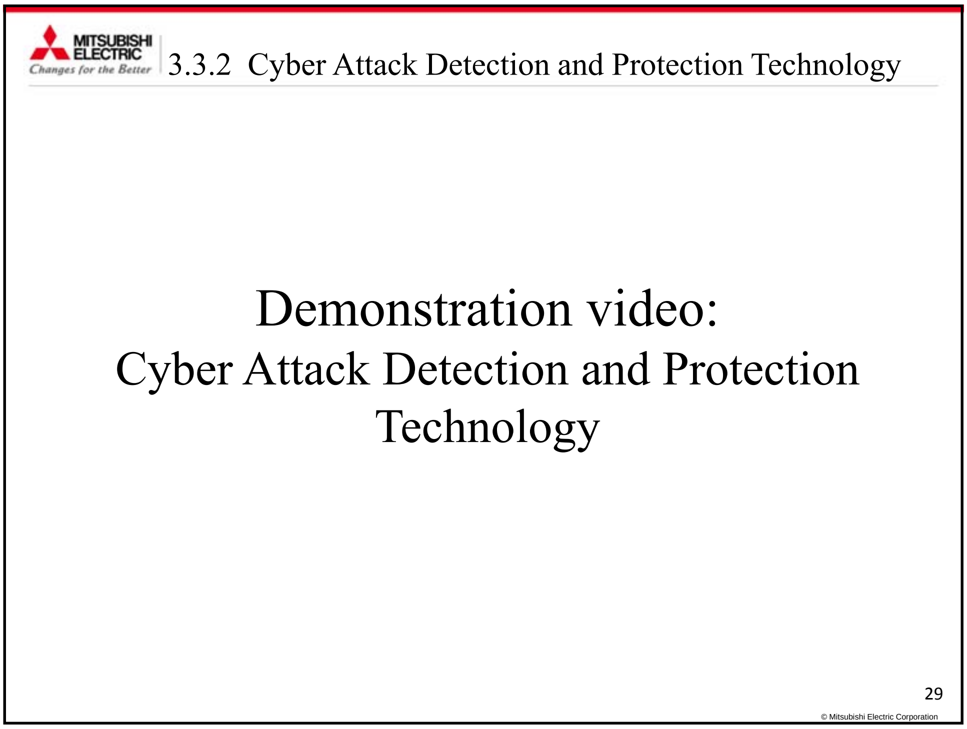


3.3.2 Information Security: Cyber Attack Detection and Protection Technology

- Detect unknown cyber attacks by watching only 50 deception techniques
- Identify hacker's activities by anticipating the possible sequence of deception techniques

Before
The rapid growth of the number of new viruses: 300M species/year
A blacklist of Viruses
Unable to timely update the blacklist.
After
The Purpose of hackers : Steal Secret data
Hacker
Remote Control
File Server
Search
Steal
Infected PC
Intrude
50 Deception techniques
Hidden commucation
Silent Scan of User IDs
Masquerade Access
Focusing on just 50 deception techniques
Video

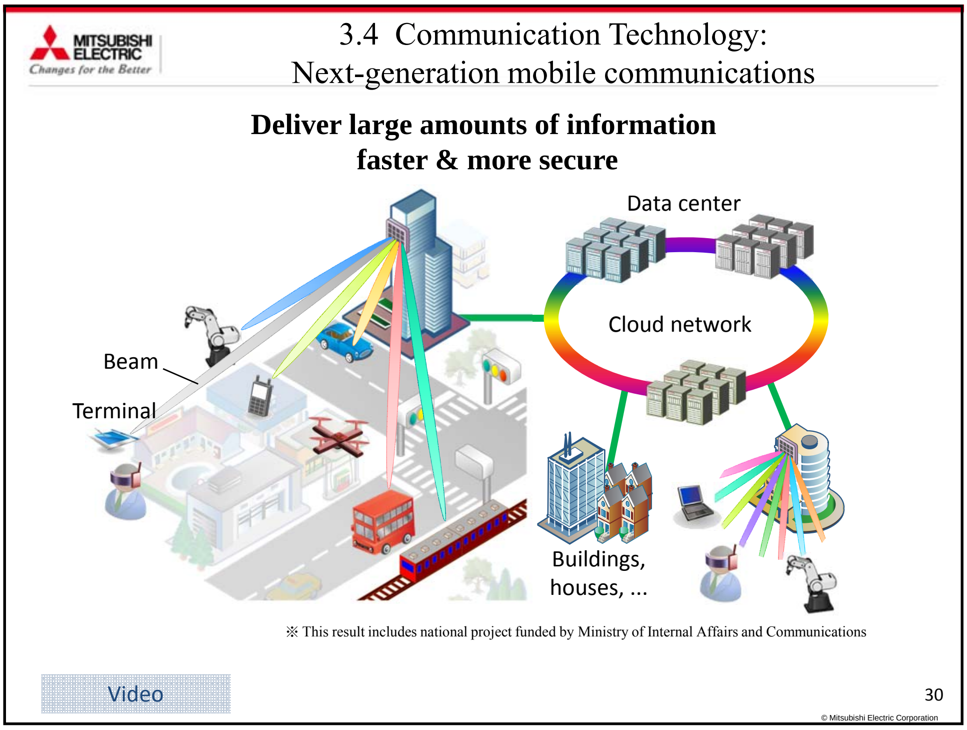
28



3.3.2 Cyber Attack Detection and Protection Technology

Demonstration video:
Cyber Attack Detection and Protection Technology

29



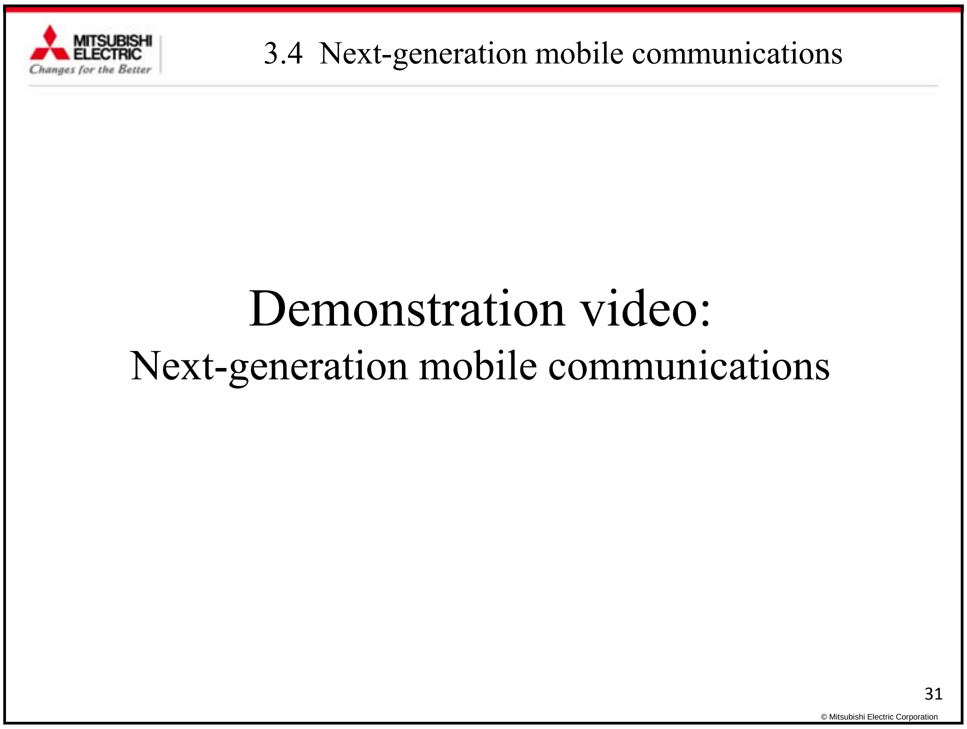
3.4 Communication Technology: Next-generation mobile communications

Deliver large amounts of information faster & more secure

※ This result includes national project funded by Ministry of Internal Affairs and Communications

Video

30



MITSUBISHI ELECTRIC *Changes for the Better*

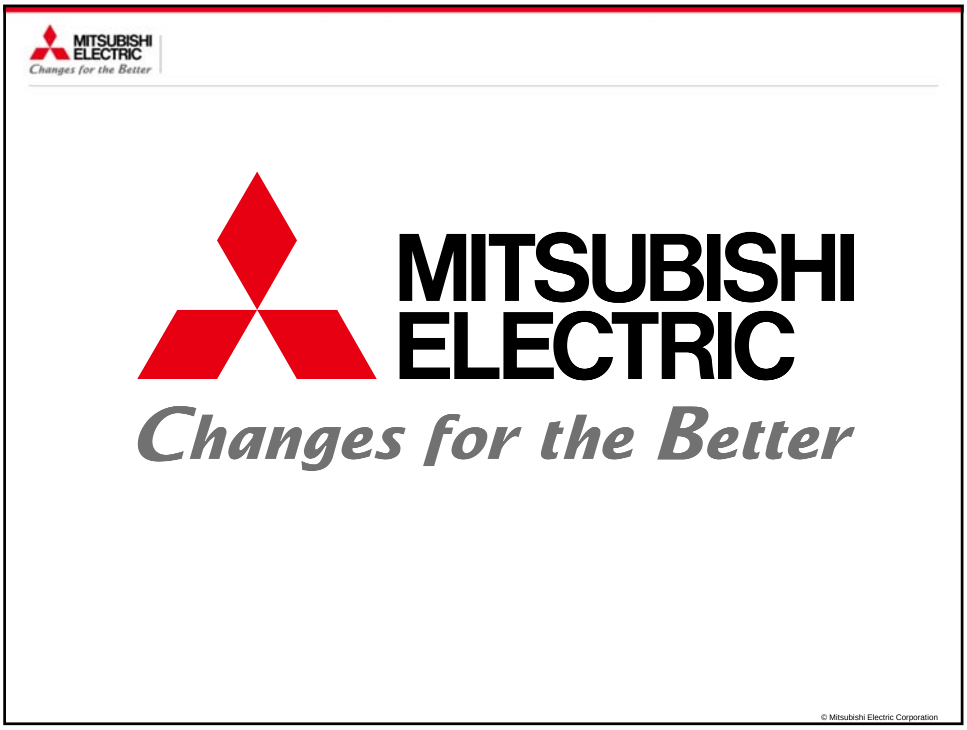
3.4 Next-generation mobile communications

Demonstration video:

Next-generation mobile communications

31

© Mitsubishi Electric Corporation



MITSUBISHI ELECTRIC *Changes for the Better*

MITSUBISHI ELECTRIC

Changes for the Better

© Mitsubishi Electric Corporation

Session 1:
Life Systems

(Chair: Ryozo Kiyohara)

An Access Prediction Method for a Safety Confirmation System Using a Lognormal Distribution Model

Masaki Nagata[*], Yusuke Abe[**], Misato Fukui[**] , and Hiroshi Mineno[*]

[*]Graduate School of Science and Technology, Shizuoka University, Japan
[**]AvanceSystem Corporation, Japan
{nagata, yu-abe, m_fukui}@avancesys.co.jp
mineno@inf.shizuoka.ac.jp

Abstract – A safety confirmation system presents a way to share information with victims during disasters. A web system used for safety confirmation should be highly reliable with good stability when suddenly accessed by many users after large-scale disaster. Therefore, the ability of the architecture to expand to accommodate additional resources during disasters is desirable. In order to expand the appropriate resources during times of disaster necessitates the use of access prediction based on the access distribution during past disasters. Therefore, we propose an access prediction method using lognormal distribution for the purpose of predicting access to the safety confirmation system during disaster. However, depending on the disaster situation, access prediction is a difficult case of a single lognormal distribution. In this paper, we propose an access prediction method using a plurality lognormal distribution depending on the disaster situation. The results of the evaluation showed that the proposed method was able to allocate the appropriate resources to access distribution at the time of disaster.

Keywords: access prediction, lognormal distribution, load balancing, safety confirmation system

1 INTRODUCTION

The ability to share safety information with victims during disasters that result in serious damage and life-threatening danger, such as the Great East Japan Earthquake of 2011 and the 2016 Kumamoto Earthquake, is important, because the early collection and disclosure of victim safety information could save many lives. A safety confirmation system offers a way to share information with victims during disasters [1]. A safety confirmation system is web system that collects and discloses safety information during disasters between the user that is preliminarily registered in the system. For example, the disaster bulletin board of the telecommunications carrier and Google Person Finder [2], and J-anpi [3] are able to crossover and collectively search from the safety information of each of these company holdings. These systems are suitable for implementation using a web system, because it is accessible by PC and smartphone for reporting and discloses of safety information. Generally, the process followed by a safety confirmation system is as follows (Figure 1). Firstly, the meteorological information service provides information about the occurring disaster to the safety confirmation system.

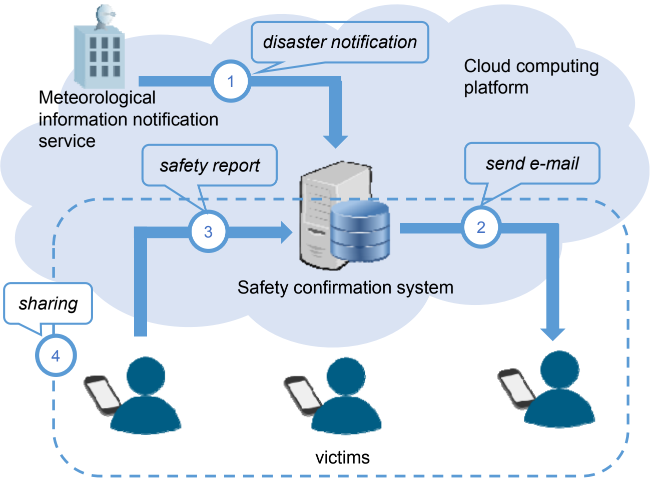


Figure 1: Flow of the safety confirmation system

Secondly, during the disaster, the safety confirmation system sends an e-mail to prompt users' for safety information. Thirdly, users who received the e-mail, report their safety information to the safety confirmation system. Finally, users share their safety information among each other. Recently developed safety confirmation systems are mainly based on cloud computing, which provides a sustainable service by using scalable platform redundancy and load balancing. Cloud computing offers a scalable solution in that it can be used on an as needed when needed basis without the need to acquire hardware assets, such as servers and switches. Additionally, cloud computing contributes to a reduction in both hardware and labor costs.

A safety confirmation system for the purpose of safety information sharing during disasters is required to continue operating reliably during disaster. Previous research [4] proposed a method to improve the availability by using a redundant server to configure the system, and we also proposed global redundancy by distributing the servers in multiple regions, including overseas as well as domestic. Global redundancy would involve achieving high availability using server failover to another area by configuring the system by locating servers around the world, even if the datacenter in the disaster area were destroyed.

Another important aspect of the safety confirmation system is the difference between the amount of access traffic at times of disaster and peace. Because of this characteristic, it would reduce operating costs by using a number of servers that are suitable for access traffic. Access to the safety confirmation system increases at the time of a disaster; hence, the ability to determine the number of servers suitable to accommodate the access traffic during a disaster is important. We obtained an understanding of the tendency of access traffic during disaster by analyzing access distribution during past disasters. As a result, real access traffic was found to exhibit lognormal distribution. Therefore, we proposed an access prediction model using lognormal distribution for the purpose of predicting access to the safety confirmation system at the time of disaster [5]. The access prediction model showed that the cost of using additional servers can be reduced by allocating an appropriate number of servers, for access distribution that varies with time during a disaster. However, access prediction is problematic in that it depends on the disaster situation. The modeling of access prediction is difficult with a single lognormal distribution, because the trend of the distribution of access to the system differs according to the disaster occurrence time. Our solution to this problem is to propose an access prediction method using a plurality lognormal distribution depending on the time at which a disaster occurs.

2 RELATED WORK AND RESEARCH PROBLEM

2.1 Safety confirmation system

Work related to safety confirmation systems has been reported in various fields, e.g., information collection and sharing, network communication, and web systems. In terms of information collection and sharing [6], a safety information system was proposed to gather refugees' information on evacuation centers. The system is intended to gather and share refugee information between the different evacuation centers set up by each local government during a disaster. Moreover, the registration of refugees' information, which is done by holding the IC card to the reader, with consideration for children and older people inexperienced with the operation of ICT equipment, has been carried out. In terms of network communication [7], a proposal based on the use of the AODV protocol was made to enable communication between users using smartphones. The system enables reliable transmission of safety information using node-to-node communication, when the communication infrastructure is damaged or usage of communication resources is restricted. In terms of web systems [8] is a proposal to relocate the safety confirmation system running on equipment on the premises to the cloud environment. Here relocation to the cloud is intended to improve the service availability to ensure sustained operation should the on-premises environment be adversely affected during the disaster. Moreover, other researchers working in the field of distributed web systems [9] have proposed load balancing and redundancy by mirroring using multiple servers, for the purpose of robustness improvement. Thus, web systems have generally been used for information management in this communication and information gathering related work. Therefore, sustainable operation of the web system infrastructure is important to achieve effective overall safety information management at times of disaster.

2.2 Problem: The number of servers in accordance with the situation

The cost of safety confirmation systems that differ depending on the number of users accessing the system during peacetime and disaster can be reduced by operating the number of servers in accordance with the access situation. The most simplistic resource management measure is to continue running the system on a large number of servers, regardless of the situation. However, the smaller amount of access traffic during peacetime means that the continuous operation of many servers at all times results in surplus resources, and, consequently, surplus costs. Therefore, if the required number of servers can be ensured to be in accordance with the access situation, this would be ideal for the resource management of the safety confirmation system, it would reduce the cost during peacetime when the amount of access traffic is small. Moreover, it is desirable to calculate and allocate the appropriate number of servers before access concentration, to avoid impairing user convenience when the response performance decreases. Calculation of a suitable number of servers in accordance with the access situation necessitates the prediction of the access distribution to the system at times of disaster. In our previous study, we proposed an access prediction model using a lognormal distribution for the purpose of predicting access to the safety confirmation system at the time of disaster. However, this approach is problematic in that the use of a single lognormal distribution to model access prediction is difficult. This is because the access distribution trend to the system was found to differ according to the time at which a disaster occurs. Therefore, it is necessary to calculate the number of servers by selecting the appropriate access prediction model according to the disaster occurrence time.

3 PROPOSED SYSTEM

3.1 System overview

The proposed system consists of a control server and the safety confirmation system (Figure 2). The control server has the function of using the access prediction model to calculate a suitable number of servers required during times of disaster, and to scale out the server for the safety confirmation system. The control server and the web server (Web) and database (DB) server for safety confirmation use Amazon Web Services (AWS) [10], EC2 [11]. AWS has Availability-Zones (AZs) [12] at different locations, and these services are equivalent to a general data center. The safety confirmation system aims to improve availability by distributed deployment in two AZs with web and DB servers, and implement load balancing by using Elastic Load Balancing (AWS ELB) [13] at the front of each AZ.

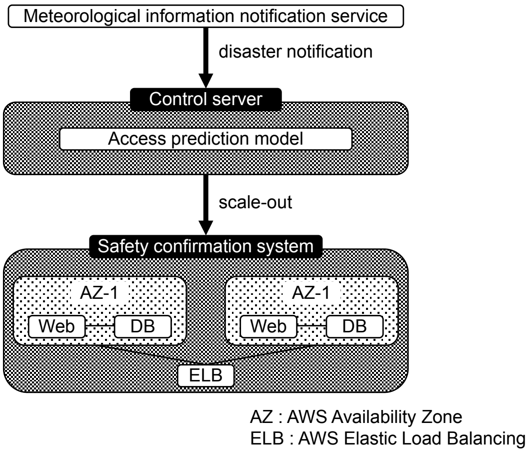

Meteorological information notification service
disaster notification
Control server
Access prediction model
scale-out
Safety confirmation system
AZ-1
Web
DB
AZ-1
Web
DB
ELB
AZ : AWS Availability Zone
ELB : AWS Elastic Load Balancing

Figure 2: Architecture of the safety confirmation system

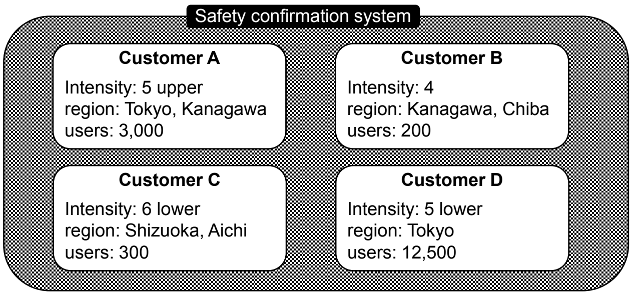


Figure 3: Multiple customers operation

It should be noted that, in our previous study, we implemented global redundancy for the proposed system by using distributed deployment across the world; however, because the subject of this paper was access prediction and load balancing, redundancy was not mentioned.

An overview of the safety confirmation system is shown in Figure 3, which depicts operation by multiple customers to share the server resources and usage by pre-system registered users. A summary of the operation after the occurrence of the disaster that corresponds to the region and the earthquake intensity threshold set by the customer unit is sent by e-mail to promote the safety report to the target users. A user who have received do safety report access to the system. Figure 3, for example, shows that an earthquake of intensity 5 upper occurred in Tokyo and Kanagawa, and that the number of target users is 15,700, among which customer A, B, D. Namely, the number of target users of the proposed system changes according to the scale of the disaster, and performs access prediction and load balancing in accordance with the number of target users.

Figure 4 shows the load balancing flow using the access prediction model. An additional server is called a scale-out, reduction is a scale-in. The scale-out operation is not executed, if the service is acceptable with the normal configuration of servers for the number of target users at the time of disaster; if unacceptable, scale-out executes with the appropriate number of servers based on the access prediction model. Scale-out executes equal load balancing by the Web server deployed to each AZ under ELB. Moreover, an e-mail is sent to target users after the completion of scale-out to avoid access concentration before the construction of a load-balancing environment.

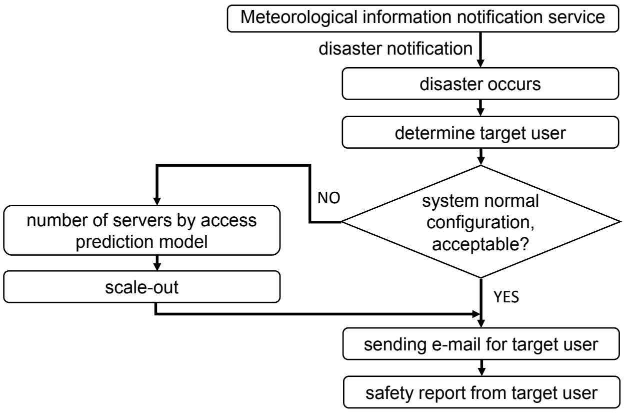

Meteorological information notification service
disaster notification
disaster occurs
determine target user
NO
system normal configuration, acceptable?
number of servers by access prediction model
scale-out
YES
sending e-mail for target user
safety report from target user

Figure 4: Flow of load balancing

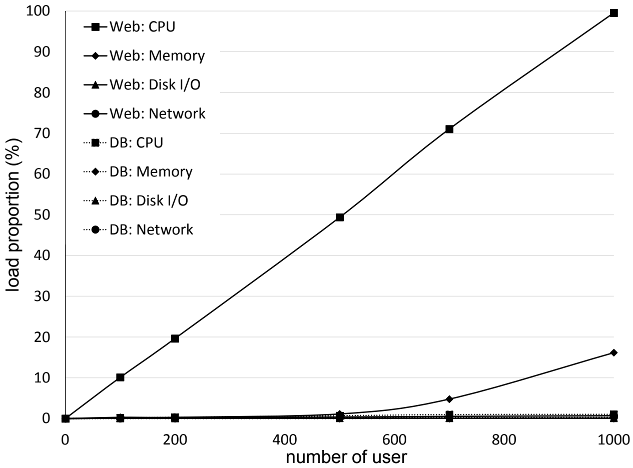


Figure 5: Web, DB resource usage proportion

In order to execute a scale-out, it is necessary to grasp the load point of the system resources. That is, it is possible to improve the processing power by adding a server when it accepts a certain load in terms of system resources. Therefore, it is necessary to grasp the load point of the system resources of the safety confirmation system at the time of disaster. Access to the system at the time of disaster accounts for more than 90% of safety report accesses, based on access logs. Thus, the load point of the system is the safety report access concentration at the time of disaster. We measured the resource consumption of the load point by using JMeter [14] to create a test scenario for safety report access. JMeter is a set of evaluation tools that enable a target system to be accessed via the web. Figure 5 shows the results of measuring each of the resources by changing access to the safety report every 10 minutes. The web server used EC2 t2.small and the DB server used EC2 c3.xlarge. Figure 5 shows that, the web server CPU usage increases greatly with increasing safety report access traffic and each of the resource loads is considerably smaller than web server CPU usage. This result indicates that the load point of the safety confirmation system increases web server CPU usage as caused by the safety report access concentration at the time of disaster. Therefore, the proposed system performs scale-out and scale-in by monitoring web server CPU usage. It should be noted, strictly speaking, the DB server was assigned a load, but this paper only targets the web server to simplify the content.

Table 1: EC2 instance types: UnixBench results

TYPE	vCPU	Memory (GiB)	System Benchmarks Index Score	Costs ($,hour)
t2.small	1	2	1702.30	0.034
t2.medium	2	4	2536.00	0.068
t2.large	2	8	2537.20	0.136
m3.medium	1	3.75	848.90	0.077
m3.large	2	7.5	1858.60	0.154
m3.xlarge	4	15	2945.00	0.308
m4.large	2	8	2025.00	0.14
m4.xlarge	4	16	3132.80	0.279

Instance types and number of servers to be used in scale-out and scale-in are decided based on single server processing power. Work that is related [15] to resource management of a web system attempted modeling the resource management of the whole system from the benchmark result of a single server. In reference to the related work, in this study, we calculate the suitable number of servers based on the processing power of a single server, to determine the access traffic obtained in the prediction. The access processing power of this study is determined by using the safety report access that can be processed per unit time. This was measured to clarify the relationship between the access processing power and each EC2 instance type. Table 1 shows the measurement result of the UnixBench [16] for a general purpose EC2 instance type. The overall index performance of the UnixBench is provided by the "System Benchmarks Index Score". Table 1 shows that the "System Benchmarks Index Score" increases by increasing the EC2 "vCPU". However, the "System Benchmarks Index Score" is not simply doubled if "vCPU" is doubled. Thus, the cost performance is higher of one vCPU type than two vCPU type. Therefore, the proposed system adopted t2.small from among the one "vCPU", considering the cost per hour and result of the "System Benchmarks Index Score". Incidentally, AWS EC2 defines the baseline of CPU usage for the t2 series including t2.small. If the CPU usage is above the baseline, the state becomes burst. Burst is a state to temporarily improve CPU performance, and it is able to continue by consuming the AWS CPU credits. If the CPU credits are exhausted, CPU performance cannot exceed the baseline. In this study, the processing power of one server is determined by the number of safety report accesses at a CPU usage of 20%, which is the baseline of t2.small, not considering the processing power of the burst. CPU usage at 20% of t2.small is able to process 200 safety report accesses in 10 minutes, as shown in Figure 5. Moreover, the processing power of the system with normal configuration is 400 safety report accesses in 10 minutes, because of using two t2.small for each web server.

3.2 Characteristic access distribution of the safety confirmation system

The prediction of access requires an understanding of the characteristic of access distribution to the safety confirmation system. Figure 6 shows the access distribution at the time of disaster.

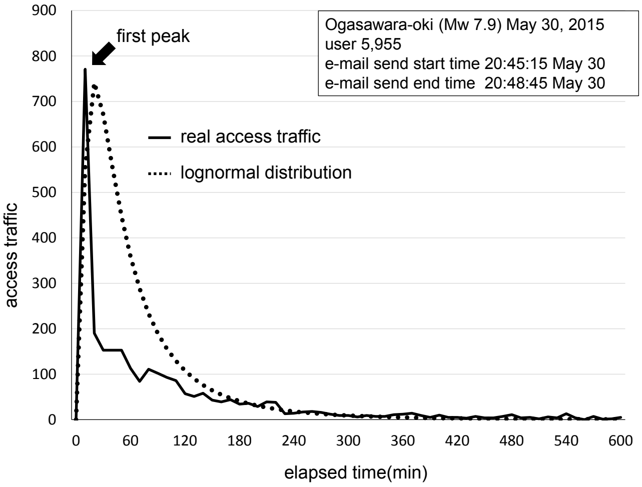


Figure 6: One-peak access traffic and lognormal distribution

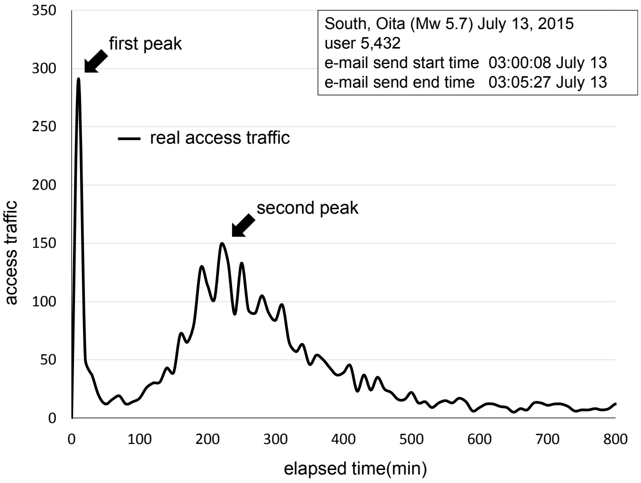


Figure 7: Two-peak access traffic

The access traffic at the time of disaster reaches a peak a short while after the initial e-mail is sent by the safety confirmation system, and then decreases with time. Moreover, Figure 6 shows that the lognormal distribution and access traffic at the time of disaster are similar. To model the counting data, usually, it is common to use a Poisson distribution. However, we proposed an access prediction method using a lognormal distribution for the purpose of predicting access to the safety confirmation system at the time of disaster, because we had confirmed a certain normality by a normality test to access distribution at the time of a disaster in the past research. When using this method is, access is in accordance with the lognormal distribution to decay period from peak, and got the effect that was expected in the number of the appropriate server calculated for access prediction; however, depending on the disaster situation, which occurs at the time of the disaster, a single lognormal distribution is problematic.

Figure 7 shows the access traffic of the disaster that occurred at 3:00 at night. Figure 7 shows two peaks, with the first peak being immediately after the occurrence, and the second peak a few hours after the occurrence. The second peak is reached in the morning and reflects human activity time.

This denotes that users who were sleeping at the time of the disaster only access the system after awakening. Therefore, our proposed access prediction method uses a plurality lognormal distribution for an access distribution consisting of two peaks.

3.3 Suitability of the mixed lognormal distribution for two peaks

Access prediction is carried out by modeling the access trend distribution analysis using data from past disasters. Our previous study entailed access prediction with a single lognormal distribution model for a one-peak disaster, as shown in Figure 6. Lognormal distribution is defined as in Eq. (1), mode *M* is Eq. (2), expected value *E* is Eq. (3), where μ is the expected value of the normal distribution, σ is the standard deviation of the normal distribution. *M* and *E* are determined from the statistics of the target user *TU* and access distribution of data from past disasters. Substituting, *M*, *E* in Eqs. (2), (3) determines μ and σ, and the parameter of Eq. (1). Equation (4) is the access prediction model, which is used to predict the access number *AN* at the time of *x* minutes. *A* is a coefficient of Eq. (1) for matching the peak access according to the number of target users *TU*.

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

$$M = \exp(\mu - \sigma^2) \tag{2}$$

$$E = \exp(\mu + \frac{\sigma^2}{2}) \tag{3}$$

$$AN = A * f(x) \tag{4}$$

Previous research [17] attempted the analysis using the mixed lognormal distribution model, for the concentration of multiple tweets from Twitter at the time of disaster. In this study to examine the suitability of applying the mixed log-normal distribution for access distribution with two peaks, in reference to previous research. The mixed lognormal distribution represented in Eq. (5) is based on Eq. (1). Moreover, it has two lognormal distributions considering adaptation to two peaks. *f(x)* is first peak lognormal distribution. *g(x)* is second peak lognormal distribution. α and β is weighting coefficients for the 1 of the cumulative probability density of *F(x)*. It shall have the relation α+β=1.

$$F(x) = \alpha f(x) - \beta g(x) \tag{5}$$

Determination of α and β is, the derivative of *F(x)* with two peaks utilize to take twice zero. It is because *F'(x1)=dF(x1)/dx*, the x-coordinate of the first peak is *x1*, and that of the second coordinate is *x2*, formulas represented by Eqs. (6), (7).

$$F'(x1) = \alpha f'(x1) + \beta g'(x1) = 0 \tag{6}$$

$$F'(x2) = \alpha f'(x2) + \beta g'(x2) = 0 \tag{7}$$

Equations (6), (7) are equal, and substituting the β=1−α relationship, can be represented as Eq. (8).

$$\alpha = \{g'(x2) - g'(x1)\}/\{f'(x1) - f'(x2) - g'(x2) + g'(x1)\} \tag{8}$$

f(x) is first peak, *g(x)* is second peak those each including, and considering *f'(x1)* and *g'(x2)* is zero, α and β can be represented as Eqs. (9), (10). Therefore, *F(x)* can be represented as Eq. (11).

$$\alpha = g'(x1)/(g'(x1) - f'(x2)) \tag{9}$$

$$\beta = -f'(x2)/(g'(x1) - f'(x2)) \tag{10}$$

$$\begin{aligned}F(x) &= (g'(x1)/(g'(x1) - f'(x2))) * f(x) + \\ &(f'(x2)/(g'(x1) - f'(x2))) * g(x)\end{aligned} \tag{11}$$

Equation (11) only strictly represents the probability distribution. Thus it is necessary to further determine the weighting coefficient to represent the access distribution. The access distribution is represented by Eq. (12).

$$H(x) = (A + B) * F(x) \tag{12}$$

$$H(x) = A\alpha f(x) - A\beta g(x) + B\alpha f(x) - B\beta g(x) \tag{13}$$

A the mainly first peak adjustment coefficient. *B* the mainly second peak adjustment coefficient. Equation (12) expands to become Eq. (13). Given that the effect of *A* in the second peak and *B* in the first peak is negligible, the second and third terms of Eq. (13) can be ignored. Therefore, Eq. (14) represents the access prediction model, which has two peaks.

$$H(x) = A * \alpha f(x) - B * \beta g(x) \tag{14}$$

Substituting *x1* and *x2* of the first and second peaks, respectively, in Eq. (14), and solving the simultaneous equations, enables *A* and *B* to be determined.

4 EVALUATION

4.1 Evaluation of suitability for access distribution

The suitability of the proposed method was evaluated for access distribution. The evaluation was carried out using 10 minutes unit access distribution during the Oita earthquake shown in Figure 7 (Figure 8), and the disaster drill in a company (Figure 9). A safety system have a similar load point in disaster and disaster drill. It is that access is concentrating from the start of the disaster drill. Therefore, disaster drill data is also used for the evaluation of the proposed method. Moreover, the reason for using disaster drill is similar to the access distribution midnight disaster. During a disaster drill the first peak occurs after the start of drill when the notification e-mail was sent in the morning, and there was a second peak during the lunch break. The number of target users of the Oita earthquake (Figure 8) is 5,432 people.

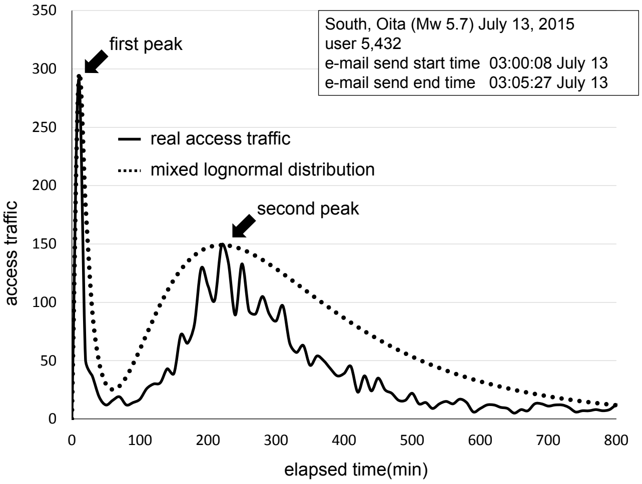


Figure 8: Oita earthquake

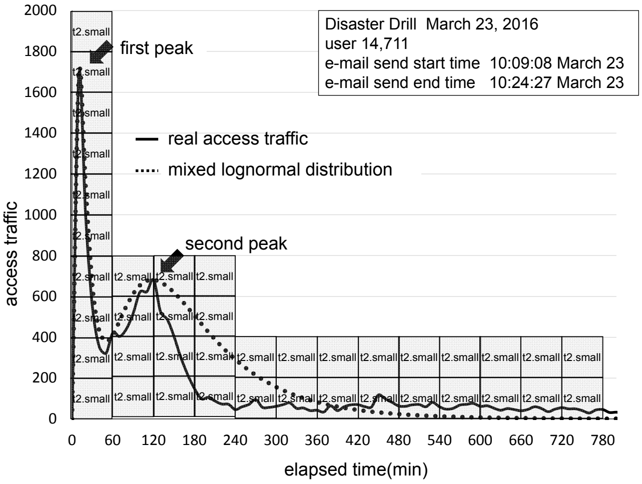


Figure 9: Disaster drill

At the first peak, 219 users accessed the system 10 minutes after the occurrence, and the second peak results from 149 users accessing the system 220 minutes after the occurrence. Substituting *x1*=10, *x2*=220 in Eq. (13), and solving the simultaneous equations produces the following results *A*=731,736.5, *B*=-55994.5. Then, σ and *μ* are determined from *M* and *E*, for the first and second distribution, respectively, which become a distribution curve of the proposed method as shown in Figure 8. The number of target users of the disaster drill (Figure 9) is 14,711 people, and the first peak occurs when 1,680 users access the system 10 minutes after the occurrence and the second peak when 681 users access of the system 120 minutes after the occurrence. Substituting *x1*=10, *x2*=120 in Eq. (13) and solving the simultaneous equations produces the following result *A*=3,535,204.1, *B*=-126719. The result shows that the peak of the proposed method is consistent with the access distribution because the known peak value of the past disaster is fitted to Eq. (13), however, the proposed method is also generally acceptable for subsequent distribution. In this paper was evaluated the suitability of Figure 7 based on the mixed lognormal distribution a case having two peaks. However, it is necessary to evaluate with many cases the future, since cases is little.

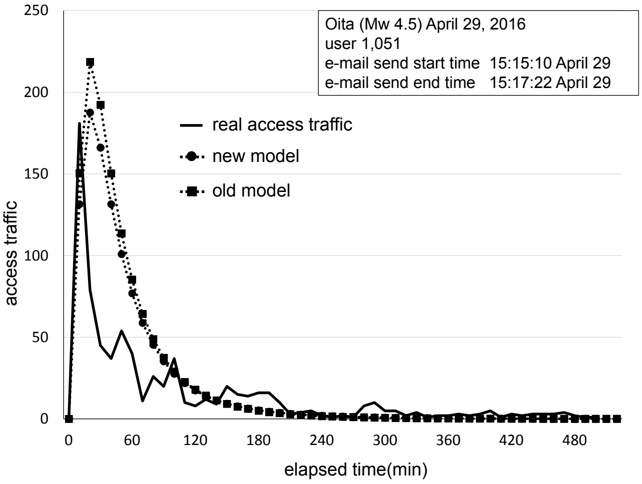


Figure 10: Accuracy of the model at Oita disaster

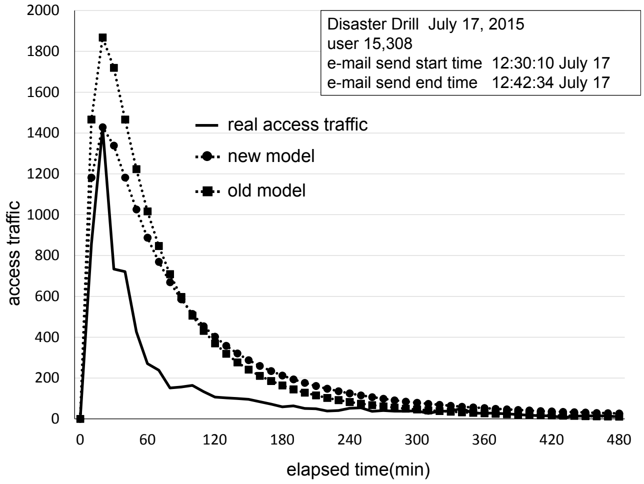


Figure 11: Accuracy of the model at disaster drill

It shows the calculation of the number of servers for the calculated access distribution using the proposed method. The number of servers was calculated based on the processing power required to access one server with an access distribution of 1-hour increments, because EC2 is to be charged on an hourly basis. Figure 9 shows that there is access of up to 1,680/10 minutes during 0-60 minutes. The processing power of t2.small is 200 safety report accesses in 10 minutes; therefore, 0-60 minutes is the 10 servers. Then we also calculated the number of servers in the same manner.

4.2 Accuracy of the model by the number of sample data

Each of the parameters that are used in the proposed model are determined by statistical analysis using data from past disasters and disaster drill data. As an example, *E* is determined by the approximation equation using the relation of the *TU* and *D*. *D* is difference between *E* and *M*. The coefficients to be granted to the lognormal distribution are calculated from the ratio of *TU* and the number of peaks. Therefore, a large amount of data for use in the analysis is expected to improve the accuracy of the proposed method. The access prediction model of a single lognormal distribution was a comparative evaluation of the number of data is less old model and the number of data is large new model. The old model was used until the summer of 2015, the number of data is 14, and the number of data of the new model to which data was subsequently added, is 29. Figure 10 shows a trace of the disaster data, whereas Figure 11 shows the trace of the disaster drill data. Compared to the old model, both new models are well suited for real access distribution, and it is seen that the accuracy of the new model is improved. The proposed method uses mixed lognormal distribution, which only fits the access distribution to disaster data and disaster drill data. However, as shown with a single lognormal distribution model in Figure 10 and 11, hereafter, data can grasp the tendency of each of the parameters to be given to the model by collecting, and can be expected to build access prediction model.

5 CONCLUSION

In this paper, we proposed an access prediction method using a lognormal distribution, to predict the concentration of access to a safety confirmation system at the time of a disaster. The proposed method uses a mixed lognormal distribution and suggested that it is possible to compute access prediction for an access distribution with two peaks resulting from the occurrence of a disaster situation.

Future challenges include the construction of the access prediction model using a mixed lognormal distribution and improving the accuracy. The mixed lognormal distribution model showed the suitability of the extent to which access distribution was allowed during the past disaster and disaster drill. However, this is a poor basis for relevance because the sample data for access prediction is small. A parameter of the model is determined based on past empirical rule with simple considered, however, using the least squares method or the maximizing likelihood method it can be expected to further improve the model accuracy. In addition, there is a need for verification and evaluation of the Poisson distribution as well as the normal distribution. In the future we plan to improve the accuracy and modifications of the proposed method, by collecting an additional amount of disaster data.

ACKNOWLEDGMENTS

We are grateful to Associate Professor Takahiro Hasegawa at the Center for Information Infrastructure, Shizuoka University who has developed safety confirmation system that became the basis for the proposed system. In addition, we are grateful to Kunihiko Murayama, president of AvanceSystem Corporation.

REFERENCES

[1] T. Hasegawa, H. Inoue, and N. Yamaki, “Development of a low running cost and user friendly safety information system”, Journal for Academic Computing and Networking, No.13, pp.91-98 (2009). (in Japanese)

[2] Google Person Finder, https://google.org/personfinder/global/home.html?lang=en, (2016).

[3] J-anpi, http://anpi.jp/top, (2016) (in Japanese)

[4] Song Fu, “Failure-Aware Construction and Reconfiguration of Distributed Virtual Machines for High Availability Computing”, CCGRID, International Symposium on. IEEE/ACM, pp.372-379 (2009).

[5] M. Nagata, Y. Abe, I. Kinpara, M. Fukui, and H. Mineno, “A Proposal and Evaluation of a Global Redundant Safety Information System Based on Access Prediction Model”, IPSJ Transactions on Consumer Devices & Systems, Vol.6, No.1, pp.94 - 105 (2016). (in Japanese)

[6] T. Ishida, A. Sakuraba, K. Sugita, N. Uchida, and Y. Shibata, “Construction of Safety Confirmation System in the Disaster Countermeasures Headquarters”, 3PGCIC. Eighth International Conference on. IEEE, pp.574 - 577 (2013).

[7] J. Wang, Z. Cheng, I. Nishiyama, and Y. Zhou, “Design of a Safety Confirmation System Integrating Wireless Sensor Network and Smart Phones for Disaster”, MCSoC, IEEE 6th International Symposium on, pp.139 - 143 (2012).

[8] Y. Hiroaki, and N. Suzuki, “Development of Cloud Based Safety Confirmation System for Great Disaster”, WAINA 26th, International Conference on. IEEE, pp.1069 – 1074 (2012).

[9] H. Echigo, H. Yuze, T. Hoshikawa; K. Takahata; N. Sawano, and Y. Shibata, “Robust and Large Scale Distributed Disaster Information System over Internet and Japan Gigabit Network”, AINA 21st, International Conference on. IEEE, pp.762 - 768 (2007).

[10] Amazon Web Services (AWS), https://aws.amazon.com/?nc1=h_ls, (2016)

[11] Amazon EC2, https://aws.amazon.com/ec2/?nc1=h_ls, (2016)

[12] Availability Zones, http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html, (2016)

[13] Elastic Load Balancing (ELB), https://aws.amazon.com/elasticloadbalancing/?nc1=h_ls, (2016)

[14] Apache JMeter, http://jmeter.apache.org/, (2016)

[15] Murta,C.D., Dutra,G.N. “Modeling HTTP service times”, In Global Telecommunications Conference, GLOBECOM'04, IEEE Vol.2, pp.972-976 (2004).

[16] UnixBench, https://github.com/kdlucas/byte-unixbench, (2016)

[17] F. Toriumi, K. Shinoda, T. Sakaki, K. Kazama, S. Kurihara, I. Noda, “Analysis of Retweet under the Great East Japan Earthquake”, IPSJ SIG Technical Report, Vol. 2012-ICS-168, No.3, pp.1-6 (2012). (in Japanese)

Are They Effective in Actual Workspace? – Case of Pressure Ulcer Prevention Features Derived From Under-controlled Experiment

Yu Enokibori and Kenji Mase

Graduate School of Information Science, Nagoya University, Japan
enokibori@is.nagoya-u.ac.jp, mase@nagoya-u.jp

Abstract - There are many features obtained from laboratory or under-controlled experiments, and proposed for many research areas. However, there are also big gaps lying between actual and laboratory experiments. In this paper, we present a result of effectiveness confirmation for pressure ulcer prevention features using weave-structure-based e-textile pressure-sensor. We do laboratory experiments with 10 elderly subjects and 12 young subjects, and yielded several effective features to estimate pressure ulcer risk, such as number and size of over-70-persentile high-pressure areas of body pressure. Although equipment used in the above experiment are the same ones used in actual workspace, the features cannot show the same effectiveness with the data collected from actual workspace experiment that is done for 18 elderly subjects who are over 80 olds in special elderly nursing home. The causes of such non-effectiveness that were yielded from discussion with staffs of the nursing home are difference between body shape and characteristic among healthy and unhealthy elders; effects of pressure dispensing cares, wrinkles generated during periodic body position changing care; and so on.

Keywords: e-textile, pressure ulcer prevention, pressure sensor, textile structure based sensor, actual experiment

1 INTRODUCTION

Many features for solving actual workspace issues are explored through laboratory experiments. Unfortunately, such features are sometimes unsuitable to solve target issues even if they are appropriate in laboratory experiments. In such cases, the knowledge about why they are unsuitable in actual workspaces is not published as public knowledge and silently ignored. Our case is such an example. However, fortunately, we can publish an overview and discuss it in this kindly workshop.

Our research is part of “The Knowledge Hub of AICHI, The Priority Research Project.” Our target is pressure ulcer prevention using a weave-structure-based e-textile pressure sensor that has the potential to be ubiquitously distributed in our daily lives. We created an interdisciplinary research group of nursing researchers, textile specialists, sensor companies, and information scientists and explored the elements that can be used for pressure ulcer prevention and risk assessment through both laboratory and actual workspace experiments.

Pressure ulcers (as known as bedsores) occur in people who have disabilities related to body control, such as senior citizens and infantile paralysis and spinal cord injury sufferers. One cause of pressure ulcers is continuous high pressure on the skin; e.g., in the prone posture, body pressure is added to the coccyx because it sticks out due to softening of body tissues in aging. In healthy patients, such high pressure is ameliorated by involuntarily body posture changes. However, people suffering from the above disabilities cannot do this adequately. Of course, many other risk factors should be considered, such as how soft the body tissues have become, remaining quantity of muscle mass, and the relationship with the body mass index (BMI). However, since most cannot be adequately confirmed through continuous measured data, we do not know exactly sure which factor can be used in actual healthcare environments with our e-textile sensor. Therefore, in our project, we first explored the features that can be used for pressure ulcer prevention and risk assessment by comparing elderly subjects (high-risk group) and young subjects (low-risk group) and then confirmed that the obtained features can be used in actual healthcare environments.

The feature exploration seemed successful from our laboratory experiment results, which was done with ten elderly subjects and twelve young subjects. We identified several features that represent the risks of pressure ulcers, such as the number and the size of the over 70-percentile high-pressure areas of body pressure. However, such features were unsuitable with actual healthcare experiments, even though the equipment used in the experiments, such as beds and pressure dispensing mattresses, was identical.

The remainder of this paper is organized as follows. In Section 2, we describe related works. Section 3 gives an overview of our e-textile-based sensor. In Sections 4 and 5, the laboratory experiment, the actual workspace experiment, and its result are explained. Discussions about the differences between the two experiment results and their causes are also explained. Finally, in Section 6, we conclude this paper.

2 RELATED WORKS

The basic guidelines of pressure ulcer prevention have been described [1]. In Japan, the guidelines of the Japanese society of pressure ulcers are the basic ones [2]. Typical care in hospitals and nursing homes includes body pressure dispersion and air-controlled mattresses. Changing the patient’s posture every two hours (even overnight) is another important basic care precept for people with body control disabilities including seniors and patients of infantile paralysis and spinal cord injuries. However, such posture changes are very heavy works even for professional caretakers. Therefore, systems are required that reduce such workload efforts. For example, if a system can detect specific body parts, such as the heels, they get high pressure, the

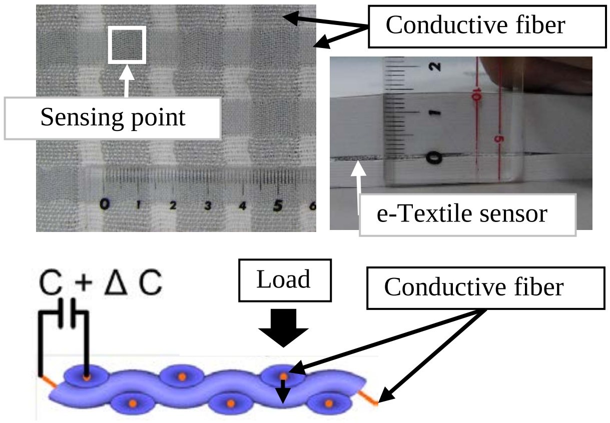


Figure 1: Overview of weave-structure-based e-textile pressure sensor

Table 1: Specification of e-textile sensor

Sensor size	180 × 90 cm (depending on circuit limitation)
Sensing point size	1 cm^2
Resolution	2 cm - $2\sqrt{2}$ cm (between center of sensing points)
Sampling rate	10 Hz (improved)
Connectability	Wi-Fi (improved)

necessary of caretakers is only to change positions of the body parts. Such a scenario might reduce the workload of caregivers.

Sheet style pressure sensors are a way to detect the amount of body pressure on the skin of patients [3,4]. Such examples include our proposed weave-structure-based pressure sensor, which can be woven with a typical weaving machine with little customization [5]. It is suitable for mass-production and disposable use. Such characteristics are also suitable for medical and healthcare uses.

Typically, sheet style pressure sensors are unsuitable for detecting absolute pressure. They are suitable for detecting relative pressure. On the other hand, most risk assessment methods for pressure ulcer prevention are based on absolute pressure values. In addition, most risk assessment and prevention methods for pressure ulcers are not based on continuous pressure measurements from sheet style pressure sensors. Therefore, we first explored the risk assessment features with zero-based thinking methodology.

3 WEAVE-STRUCTURE-BASED E-TEXTILE PRESSURE SENSOR

An overview of our weave-structure-based e-textile pressure sensor is illustrated in Figure 1. This device is an improvement of a previously described one [5]. Most structures are identical as those already described. The sensing points are the cross points of the conductive weft and warp whose capacitance changes are converted to pressure values with a linear regression function trained through an initialize phase.

The points that have been improved are the sensing specifications (Table 1). Sensor size, sensing point size, and the resolution haven't changed. However, the sampling rate was improved from 2 to 10 Hz, and a Wi-Fi connection feature was added. We also added several noise-reduction features.

Table 2: Characteristics of laboratory experiments

	Elderly	Young
Age	68.8 ± 4.0	21.3 ± 0.8
Males	6	6
Females	4	6
Body Mass Index (BMI) (n.s. with t-test**)	21.8 ± 3.7	21.3 ± 3.0

* All subjects were healthy.
** Welch's t-test

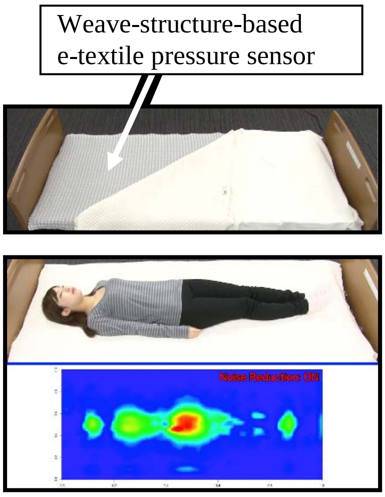


Figure 2: Sensor installation and an example of body pressure

4 LABORATORY EXPERIMENT TO EXPLORE PRESSURE ULCER RISK FEATURES

In this section, we describe our preliminary laboratory experiments that we did as part an actual workspace experiment from September through December, 2013. We compared the data from the elderly and young subjects to identify the features that can be used for pressure ulcer risk estimation. As almost researchers know, healthy elders do not have significant higher risks than young people. However, there are similar predispositions even if these are not enough to develop pressure ulcers

4.1 Experiment setting and procedure

The characteristics of our ten elderly and twelve healthy young subjects are summarized in Table 2 . They had no significant body shape differences because of the body mass indexes (BMIs) and their Welch's t-tests (p-value > 0.05).

Sensor installation is illustrated in Figure 2. An overview and an example of the experiment's collected data are illustrated in Figure 3. Weave-structure-based e-textile pressure sensors were installed between the mattress and the

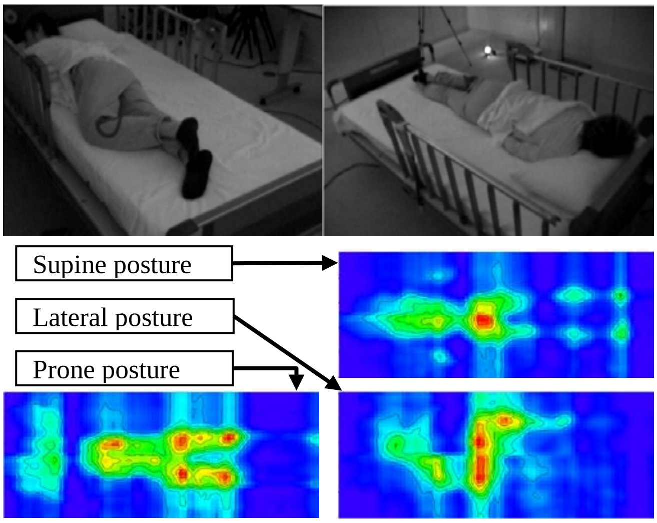


Figure 3: Overview of laboratory experiment and examples of collected body pressure data

Table 3: Number of posture change

	Elderly	Young
Average	14.4 ± 4.6	20.2 ± 6.6
Min – Max	7-20	9-30
p-value *	**0.03**	

* Welch's t-test

Table 4: Ratio of 0.3 and 0.7 area sizes

	0.7/0.3	p-value *
Elderly	0.118 ± 0.046	**0.02**
Young	0.133 ± 0.047	
Male	0.130 ± 0.050	0.27
Female	0.125 ± 0.043	
BMI: < 19 (thin)	0.133 ± 0.043	All pairs > 0.05 **
BMI: 19-25 (normal)	0.128 ± 0.050	
BMI: 25 < (plump)	0.115 ± 0.038	

* Welch's t-test
** Welch's t-test and Bonferroni revision

bed-pad to make a space between the sensor and the bodies to remove the capacitance changes caused by the close proximity of the bodies. In addition, a plastic sheet was placed between the sensor and bed-pad to remove the capacitance change caused by evaporated sweat. As illustrated on the top two pictures in Figure 3, there were no heavy blankets and cushions except a pillow. The collected data are very vivid and can distinguish among the laying postures (bottom, Figure 3).

Experiments were done individually for each subject around 1 p.m. in the same room. Subjects were instructed to sleep for over four hours. The rooms included a bed and a body pressure dispersion mattress, both of which are commonly used in nursing homes. The lights were turned off after subjects lied down on their beds. The pictures that collected the posture of the subjects as ground truth were taken by two infrared cameras.

Table 5: Number of 0.3 and 0.7 areas

Prone posture

	Elderly	Young
# of 0.3 area	4.8 ± 1.2	4.4 ± 1.6
p-value* for above	0.63	
# of 0.7 area	2.2 ± 1.0	2.6 ± 1.2
p-value* for above	**0.02**	

Lateral posture

	Elderly	Young
# of 0.3 area	2.3 ± 1.4	2.5 ± 1.1
p-value* for above	**0.03**	
# of 0.7 area	2.4 ± 1.1	2.3 ± 1.0
p-value* for above	2.52	

* Welch's t-test

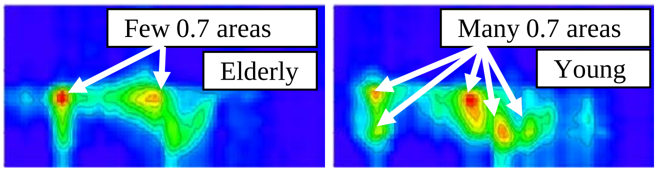


Figure 4: Example of different 0.7-area numbers

4.2 Sensor data preprocessing for discussions

The sensor data were preprocessed in the following manner because the sensors cannot output absolute pressure values instead of the relative pressure values. The relative values were normalized based on the max value of each subject: 1.0 for maximum and 0.0 for minimum. We converted the sensor output for the 2D color images (Figure 3). The blue is the low pressure area, and the red is the high pressure area. In addition, we added eleven contours, 0.0 (blue) to 1.0 (red) by 0.1, for the images. In addition, we agreed that the 0.3 lines roughly represent the body shape of the subjects by visual observations and discussions among healthcare researchers.

Note here that the converted values have different meanings among subjects because they are based on the max value of each subject. Therefore, even if there are many red or huge red areas, that does not denote risk. This suggests that the body pressure is well distributed. On the contrary, only one small red area denotes high risk because there is concentrated high pressure.

4.3 Findings from comparison of elderly and young subjects

The features gleaned from our discussions to distinguish elderly (high-risk group) and young subjects (low-risk group) are described in Tables 3 to 6. We focused on the number of posture changes, the size ratio of the 0.3 and 0.7 areas, the number of 0.3 and 0.7 areas, and the roundness of the 0.7 area because they showed suitability in brute force style feature exploration. Next we only explain the overviews of the yielded features below because this paper is limited to information science.

Table 6: Roundness of 0.7 areas

Prone posture

	Roundness	p-value *
Elderly	0.887 ± 0.036	**0.02**
Young	0.690 ± 0.175	
Male	0.754 ± 0.207	0.37
Female	0.782 ± 0.135	
BMI: < 19 (thin)	0.762 ± 0.226	All pairs > 0.05 **
BMI: 19-25 (normal)	0.747 ± 0.156	
BMI: 25 < (plump)	0.935 ± 0.013	

Lateral posture

	Roundness	p-value *
Elderly	0.828 ± 0.103	**0.02**
Young	0.614 ± 0.206	
Male	0.632 ± 0.245	1.00
Female	0.755 ± 0.144	
BMI: < 19 (thin)	0.716 ± 0.282	All pairs > 0.05 **
BMI: 19-25 (normal)	0.703 ± 0.173	
BMI: 25 < (plump)	0.645 ± 0.357	

* Welch’s t-test

** Welch’s t-test and Bonferroni revision

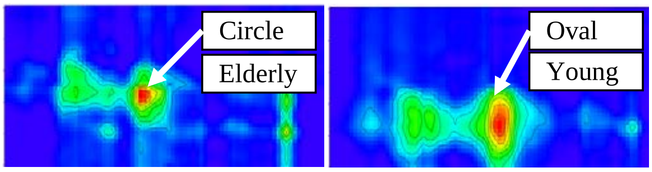


Figure 5: Example of difference of 0.7 roundness

1. Number of posture changes

Table 3 shows significant differences between the elderly and young subjects with Welch’s t-test (p-value < 0.05). It represents how well the subjects can control their bodies as well as their sensitivity to skin context changes. Typically, such abilities declined in the high-risk group.

2. Size ratio of 0.3 and 0.7 areas

An 0.3 area roughly represents the body shape. Therefore, the size ratio of the 0.3 and 0.7 areas represents the distribution of their body pressure, as described in the last part of section 4.2. If the body pressure is concentrated on a point, the 0.7-area size will be small.

We found a significant difference between the elderly and young subjects (Table 4). Considerable causes that affect this feature include remaining muscle mass, how soft their body tissues are, and so on. It was not confirmed.

On the other hand, we found no significant differences between the male and female subjects or among the BMI difference groups, suggesting that this feature can estimate the risk of pressure ulcers without regarding gender and body shape issues.

3. Number of 0.3 and 0.7 areas

As described in Table 5, there is a significant difference between the elderly and young subjects with the 0.7-area numbers on the prone postures. “Number of areas” denotes the low-risk group with more supporting points than the high-risk group and distributed body pressure (Figure 4). This feature could be used for the risk estimation of pressure ulcers. In this experiment, however, we were unable to confirm its causes, such as remaining muscle mass, and how soft their body tissues have become.

There was also a significant difference between the elderly and young subjects with the numbers of the 0.3 areas in lateral posture. This result denotes that the high-risk group has fewer contact points between their bodies and the bed surfaces. This is perhaps one more feature to represent how soft their body tissues have become. Imagine two linked sausages. If their material is solid, a space will be made under the link point. If the material is very soft, such as melted cheese, the sausage’s shape cannot be kept, and the space under the link point is filled.

4. Roundness of 0.7 area

Roundness is defined by the following equation:

$$R = \frac{s}{\pi\left(\frac{l}{2\pi}\right)^2}, \qquad (1)$$

where R is roundness, s is size of area, and l is length of circumference of area. This equation depends on the fact that area size is maximized with circle shape if length of circumference is the same.

As shown in Table 6, there are significant differences among elderly subjects and young subjects with roundness of 0.7 area on both of prone and lateral postures. It perhaps represents how their bones stick out from their body as described in the last part of last section.

On the other hand, there are no significant differences between male and female subjects, and among BMI difference groups. It means that this feature can be used for risk estimation of pressure ulcer without dependences of gender and body shape.

4.4 Conclusion of laboratory experiment

Our laboratory experiments found several features that can be used for pressure ulcer risk estimation. Based on discussions with nursing researchers, we believed that they can be used for groups at higher risk than the laboratory experiments’ subjects, such as seniors over 80, dementia sufferers, and so on.

5 ACTUAL WORKSPACE EXPERIMENT TO CONFIRM SUITABILITY OF FEATURES

Since we successfully found significant characteristics that are suitable for pressure ulcer risk assessment in laboratory experiments, we confirmed their suitability in an actual workspace experiment that was done in collaboration with a special elderly nursing home, categorized as “Tokuyou” in Japan, around Nagoya City. The details of this experiment are given below.

5.1 Experiment setting and procedure

We agreed to keep the details of the subject characteristics private based on the nursing home’s request. Our experiment was done with 18 subjects over 80-years-old in a special elderly nursing home from November 20 to 23, 2014. However, we only describe 16 because one subject withdrew after the experiment began, and the data of another subject was not successfully collected due to mechanical trouble. The gender ratio was basically even. Since most were unable to turn their bodies over well, this group had higher pressure ulcer risk than the laboratory experiment's elder group.

The sensor installation was basically identical as in the laboratory experiment. Please refer to section 4.1 for the details of the sensor placement. In contrast with our laboratory experiment, there were a few varieties of beds and mattresses including the ones used in laboratory experiment. Because such beds do not directly make contact with the subjects and all of the mattresses are pressure-dispensing types, we assumed that such bed and mattress differences will not affect the experiment results.

Our experiment, which was designed to avoid interrupting the typical healthcare procedures of the nursing home, involved no special attention except the sensor setup and removal that was done by the researchers. Both the sensor setup and removal were done during lunch time when the occupants had moved to the dining space. We collected the 24-hour data and removed the sensors from the subject beds during the next lunch period. No typical over-night care was disturbed, including posture changes every two hours. The cushions that were inserted between the subject's body and the bed are also typical nursing home care. In contrast with our laboratory experiments, ground truth pictures of the laying postures were not able to be collected because of privacy issues.

5.2 Sensor data preprocessing for discussions

We preprocessed the sensor data in almost the same manner as the laboratory experiments. The relative value was normalized based on the max value during each eight-hour non-overlapped time window of each subject: 1.0 for the maximum and 0.0 for the minimum. We converted the sensor output for the 2D color images. The blue is the low pressure area, and the red is the high pressure area. In addition, we added eleven contours, from 0.0 (blue) to 1.0 (red) by 0.1, for the picture.

5.3 Result

Before starting this discussion, we planned to confirm the suitability of the identified features during laboratory experiments by comparing the actual workspace and laboratory experiment data. However, we aborted such plans because this experiment's 2D color map showed significantly different shapes (Figure 6). These samples are very typical meaningful images obtained through this experiment. Of course, there are also many meaningless images.

As seen in Figure 6, the shapes are significantly different from the laboratory experiment's ones (Figure 3). We were unable to confirm that the 0.3-line areas roughly represent the shape of the subjects. Most 0.3-line areas are too small to be considered rough body shapes.

The top five pictures are the body pressure data that were estimated as lateral postures. The bottom four pictures are the body pressure data that were estimated as supine postures.

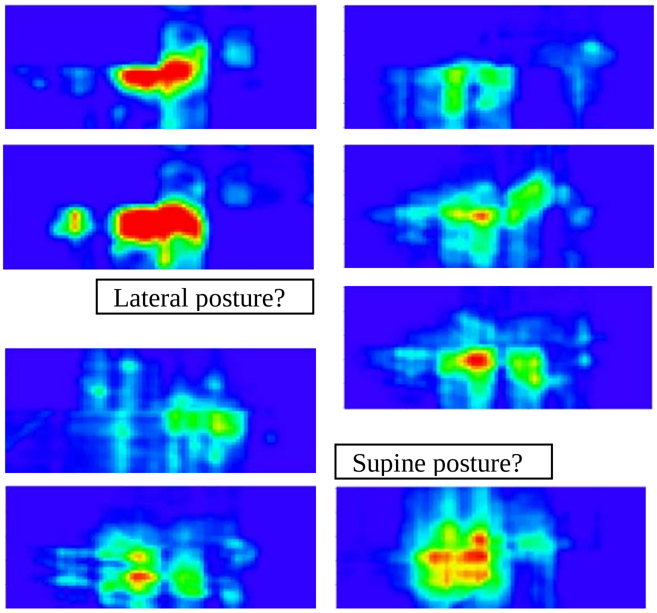


Figure 6: Examples of collected body pressure data from actual workspace experiments

The above two estimations were based on our subjective discussions because the lying posture classifiers that were trained with the laboratory experiment data are unsuitable for this data. Because of the lack of ground truth pictures, we confirmed the estimations through discussion among the nursing home staffs who are familiar with the typical sleeping postures of the subjects.

5.4 Discussions

Next we discuss the causes of the following odd collected data and focus on the following three main causes:

1. Contractured body shape

The first cause that we considered is contractured body shapes, such as kyphosis. Most high-risk subjects have such a body shape. Their backbones and sometimes their arms and legs were bent and stiff. Therefore, such stiff body parts barely touched the bed's surface. For example, the backbones directly touched the bed surface during the supine posture (right-bottom picture, Figure 6). In addition, the arm contracture might have a slightly strange lean during the lateral postures. Such characteristics did not appear in the healthy senior group.

2. Wrinkles made through daily care

The second cause that we considered is the sheets wrinkles made through daily care. When we removed the sensors, we found that many wrinkles remained on the sheets. They may have caused the image blurs and fake pressures shown in the left-top picture of supine posture in Figure 6.

We also identified some wrinkles of sheets in the laboratory experiment; however, in the actual experiment they seemed large and numerous. We suspect that they are caused by the substandard daily care provided by the nursing home. Caretakers sometime climb up on subjects' beds during overnight body posture changes and support such lying positions. In addition, subjects were not simply lifted up and put on the bed in such care. They were often pushed, pulled, or dragged because humans are too heavy for “lift up and put on” style care.

3. Effects of body pressure dispersion care

The biggest cause that we have to consider is the effects of body pressure dispersion care. Because the highest priority of nursing homes is prevention of disease, staff members provide the best possible care, such as body pressure dispersion cushions, multiple clothes to make buffers around high-risk parts, adjusting the tension of bed shirts, and so on. Therefore, the collectable data from the nursing home is well-distributed body pressure data, such as the right-top picture of the lateral postures that cannot be used to analyze our algorithm and our system.

5.5 What should/can we do in experiments under actual healthcare situations?

Based on the above discussion, what we should do to collect suitable data from the high-risk group is “failed to take care during experiments.” Because most cooperative nursing homes are perhaps advanced nursing homes, there are only well-cared subjects. They provide only unsuitable data for algorithm and system analysis. A solution for this issue is collaboration with substandard nursing homes for our experiments. However, there are another difficulty that they are less motivations to cooperation. Even if we successfully corroborated with such nursing homes, how can we avoid offering advice for such inadequate care that might cause more pressure ulcers? It is also difficult because of moral issues.

Home care situations offer possibilities. Most caretakers in such situations are not trained professionals. Therefore, we can collect “poor care” data and might even collect “failed care” information. Even if we provide superior advice, such caretakers are unlikely to consider all of it. Perhaps the moral issue is mitigated.

However, improvements in information science do not allow such experiments. For example, IoT enables us to collect data in real-time. Moreover, we can give advice to caretakers through ubiquitous network connectivity and such advanced hand-held devices as smartphones. In addition, we have been addressing whether such systems can be implemented with little labor. Therefore, we might be severely criticized if we avoid efforts to build such prevention systems.

6 CONCLUSION

Many research projects obtain features from laboratory or under controlled experiments to solve actual workspace problems. Our project, which is an example of such an experiment, addresses pressure ulcer prevention. We obtained several characteristics that seem suitable for pressure ulcer risk estimation through laboratory experiments with ten elderly subjects and twelve young subjects. However, these characteristics were not suitable for the data collected from an actual workspace, a special elderly nursing home with 18 elderly subjects. The causes of such non-effectiveness yielded from discussions with nursing researchers and the staff members of the nursing home are the differences between body shape and the characteristics among healthy and unhealthy elders, the effects of pressure dispensing cares, and wrinkles of sheets generated during periodic body position changes.

Based on our discussion, we now face a moral dilemma. We have to collect data from “poor or failed care.” However, this dilemma might be prevented by state-of-the-art information science. Future work will address possible solutions to this moral issue.

ACKNOWLEDGEMENTS

Some of this research was supported by “The Knowledge Hub of AICHI, The Priority Research Project” and JSPS KAKENHI Grant Number 15H02736. We are grateful to Dr. Yuko Harasawa, Associate professor of the Graduate School of Nursing, Nagoya City University, Dr. Yoshiko Takahashi, research assistant of Life and Health Sciences, Chubu University, and Dr. Satoko Yanagisawa and Dr. Yoshiko Komatsu, Professors of the Graduate School of Nursing and Health, Aichi Prefectural University, for their help and discussions.

REFERENCES

[1] J. K. Stechmiller L. Cowan, J. D. Whitney, L. Phillips, R. Aslam, A. Barbul, F. Gottrup, L. Gould, R. Lisa, M. C. Robson, G. Rodeheaver, and others, Guidelines for the prevention of pressure ulcers, Wound Repair and Regeneration, Vol. 16, No. 2, pp. 151-168 (2008).

[2] R. Tsuboi, M. Tanaka, T. Kadono, Y. Nagai, K. Furuta, Y. Noda, and others, JSPU guidelines for the prevention and management of pressure ulcers, Jpn J Pu, Vol. 16, No. 1, pp. 12-90 (2014)

[3] J. Meyer, P. Lukowicz, and G. Troster, Textile Pressure Sensor for Muscle Activity and Motion Detection, 10th IEEE International Symposium on Wearable Computers, pp. 69-72, (2006).

[4] M. Rothmaier, P. M. Luong, and F. Clemens, Textile pressure sensor made of flexible plastic optical fibers. Sensors, Vol. 8 No. 7, pp. 4318-4329 (2008).

[5] Y. Enokibori, A. Suzuki, H. Mizuno, Y. Shimakami, and K. Mase, "E-Textile Pressure Sensor Based on Conductive Fiber and Its Structure," In Adjunct Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), pp. 207-210 (2013)

Session 2:
Sensor Networks

(Chair: Yu Enokibori)

Reducing Software Update Data Sizes for Networked Devices with Fixed-Length Operation CPU Architecture

Yusuke Fukuda[*], Yutaka Onuma[**] , Yoshiaki Terashima[***], and Ryozo Kiyohara[*]

[*] Kanagawa Institute of Technology University, Japan
[**]Graduate School of Kanagawa Institute of Technology, Japan
[***]Soka University, Japan

Abstract – Sensor networks have many devices connected together through slow wireless networks. Currently, these devices have a limited number of applications; however, in the future, these devices will be widely used for varied purposes. Therefore, there is a need for software update functions on slow networks. Because many devices are connected to one network, updating time should be small, and therefore a small data size for updates is beneficial. In this paper, we propose a method to reduce the data size for software updates. Our method is based on bsdiff, which is a well-known code-difference technique. We propose an improved bsdiff method and evaluate it using simulation. We found that it slightly reduced some of the data sizes.

Keywords: Sensor network, Software Updating, Data Compression, Fixed Size Operation Code, Bsdiff

1 INTRODUCTION

In recent times, because of the reduction in size as well as cost of various electronic devices that can connect to the internet, including smartphones, utility devices, and robots, IoT (Internet of Things) has grown widely. This has led to a sharp increase in the types and scale of content generated by such devices.

Most of these devices are continuously running, such as various sensing services and machine controllers, which need to operate without halting. These devices might require their software to be updated regularly.

Furthermore, these devices might not be connected via a fast network, such as a cellular or broadband network. For example, an in-vehicle network is a slow network, which might consist of, say, more than 70 automotive ECUs (Electronic Control Units); this is depicted in Figure 1. An example of a typical in-vehicle network is a CAN (Controller Area Network) that can have a maximum operating speed of 1 Mbps depending on the length of the in-vehicle network. The devices in such in-vehicle networks might require their software to be updated to resolve any issues that were identified after the vehicle was shipped or to enhance their efficiency [1][2].

Currently, for these software updates, the vehicles need to be sent to the factories or brought to car sales dealers; this is not only time consuming, but also inconvenient for the users. Therefore, it is required that, in the near future, this new software be downloadable to devices (e.g., PCs, tablets, or smartphones) at the user’s home or office as is shown in Figure 2. In addition, the software for vehicles has to be

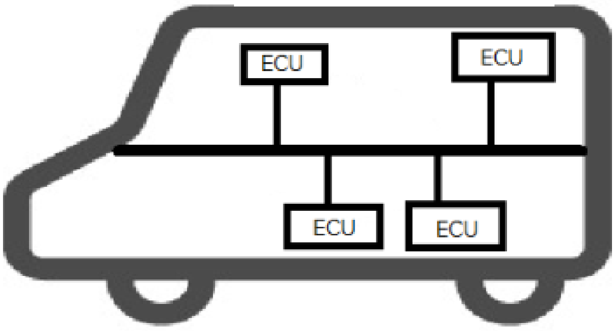


Figure 1: Example of vehicle network

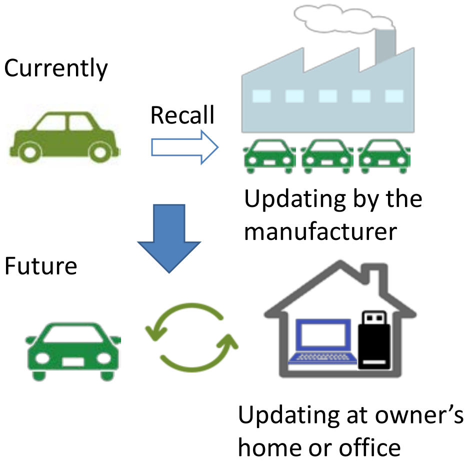

Currently
Recall
Updating by the manufacturer
Future
Updating at owner’s home or office

Figure 2: Example of sensor networks

updated automatically as well. However, during this software update, the users cannot drive their car; therefore, the software update time should be small.

Similar to in-vehicle networks, in most sensor networks, the size of device software as well as the network capacity is relatively small. Each ECU on the in-vehicle network has different functions and therefore different software. However, in the case of many sensor networks, most of its devices have the same function, which is performed by the same software and therefore all of the devices need to be updated simultaneously. In both cases, the time taken for the software update is crucial. In this study, we consider the feature of networks and focus on binary difference technologies. We consider bsdiff, which is a widely used and effective algorithm, and propose a new algorithm based on it.

This paper is organized as follows. In Section 2, we discuss the various features of sensor and in-vehicle networks that are important with regards to software updates. In Section 3, we introduce the related works, and in Section 4, the bsdiff algorithm is presented and its problems are discussed. Then, in Section 5, we propose a new algorithm based on bsdiff, and in Section 6, we evaluate our method. We conclude the paper with a brief summary in Section 7.

2 FEATURE OF DEVICES ON SENSOR NETWORKS

As shown in Figure 1, ECUs on the in-vehicle network are connected via buses. Each ECU is independent and in many cases sourced from different manufacturers. Therefore, the software update first needs to be downloaded to other connected devices such as a PC, tablet, smartphone, or car navigation device. Alternatively, these devices can connect directly or through the OBD (On-board Diagnostic) interface port.

As previously mentioned, many sensor networks have the same devices, all of which have the same functions and same software. However, the network topology may differ, such as bus type, ring type, mesh type, and tree type. In all these cases, all the devices must be updated simultaneously. Therefore, the size of data should be small.

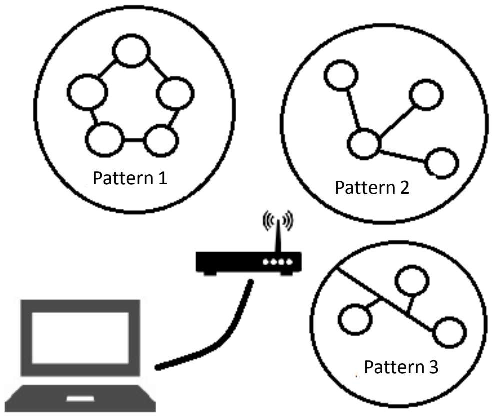


Figure 3: Typical network topology of sensor network

The time required for software updates in the case of both, networks and individual devices, depends on the downloading time, as shown in Figure 4.

We focus on the compression of data for software updating. Differential updating technologies should be applied. Though there are various differential software updating technologies, in this study, we focus on the bsdiff, which is a well-known algorithm, and propose an improved method of differential update based on bsdiff.

3 RELATED WORKS

During an ECU software update, the vehicle must be stationary and should not be driven; however, the engine must remain on. Therefore, the update time should be considerably small.

Furthermore, reducing the software update time requires the following three considerations: structure of the software, download time, and update time for flash memories.

3.1 Software Structure

The software structure is related to the differences in binary code. Figure 5 shows a typical binary update. The left –hand side of Figure 5 shows the source code, which has

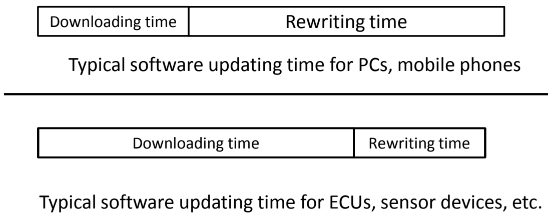

Downloading time
Rewriting time
Typical software updating time for PCs, mobile phones
Downloading time
Rewriting time
Typical software updating time for ECUs, sensor devices, etc.

Figure 4. Software updating time

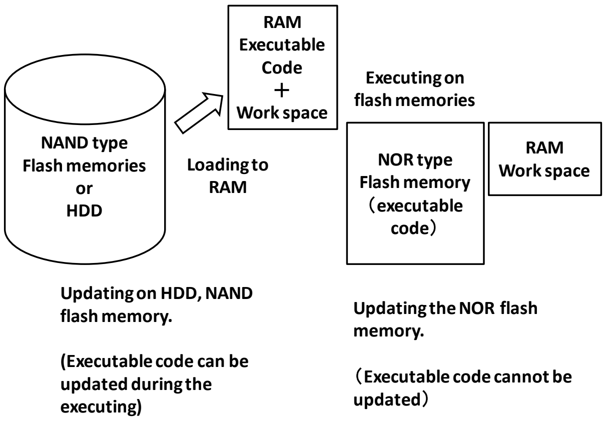


Figure 5: software updating time for each process

one line of code modified. The right–hand side shows the binary code, which has many differences between the new and old versions because of immediate data that refers to modified addresses (i.e., new instructions were inserted, therefore, the old instructions moved to higher addresses). These references cause a considerable difference between the original and modified binary codes.

Reference [1] presents a solution to this problem. However, this solution requires information from the development environment, which is difficult to acquire. For example, it requires the symbol information that is used in the linker, and the redundant additional spaces, which reduce the reference information that is changed by memory sliding.

3.2 Downloading Time

Let *T(D,S)* be the time that cannot serve many functions. This time depends on *D*, which is the size of the binary difference. S is the size of the blocks to be rewritten. Let *R(S)* be the rewrite time that depends on *S*. *T(D,S)* can be calculated with the following equation.

$$\mathrm{T(D,S)} = \frac{\mathrm{D}}{\mathrm{V}} + R(S)$$

where *V* is the network bandwidth.

If a high-capacity network can be used, e.g., for PCs or smartphones, then *V* is considerably large, and the ratio *D/V* is negligible; therefore, *T* depends on *R(S)*, i.e., the software updating time. However, if a high-capacity network cannot be used, the bandwidth *V* is small, and therefore *D/V* is considerably large. Then, *T* depends on *D/V*. As *V* cannot be changed, *T* depends on *D*. As previously mentioned, the in-vehicle network has a considerably low capacity; thus, the size of *D* has to be reduced.

Many studies have been conducted on binary differences [2][3][4][5][6]. These studies proposed the representation formats and methods of searching for binary differences, which are based on [5].

3.3 Rewriting Time

The rewrite time depends on the type of flash memory. Because NAND flash memory can be updated during execution, as shown in Figure 5, we need not consider the rewrite time. However, we do need to consider it for NOR flash memory.

Moreover, the rewrite time depends on the number of blocks to be rewritten, which depends on the size of the flash memory’s erase blocks. Reference [1] shows how to gather the differences to the same erase blocks. However, because the software size of automotive ECU program and sensor devices is not that large, the erase blocks are not an issue.

3.4 Downloading Protocol

Not only are there many network topologies, but the software updates also need to be conducted simultaneously in the case of sensor networks. Therefore, many protocols are proposed. Reference [7] introduced one such protocol.

3.5 Goal

When updating software for an automotive ECU or sensor devices, the download time is crucial, because of the poor network capacity. Therefore, in this study, we focus on reducing the binary difference. This can be achieved by bsdiff [2], which is a well-known, efficient algorithm for PC software. However, this algorithm is optimized for a non-fixed-length operation set (e.g., the Intel architecture). In this study, we modified it for a fixed-length operation set (e.g., the ARM architecture)..

4 BSDIFF

In this section, we introduce bsdiff. The bsdiff algorithm consists of the following four parts:

(1) Searching for a common sequence,
(2) Extension of the common sequence
(3) Deciding the differential area, and
(4) Compression

These parts are processed as shown in Figure 6. Common sequences are searched repeatedly based on an old algorithm [7] and suffix-array sorting. Then, each matching common sequence is extended based on the similarity of the new code with the old code.

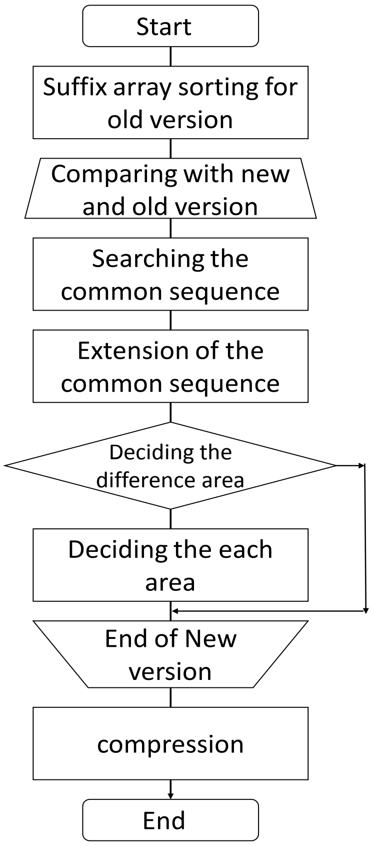

Start
Suffix array sorting for old version
Comparing with new and old version
Searching the common sequence
Extension of the common sequence
Deciding the difference area
Deciding the each area
End of New version
compression
End

Figure 6 Algorithm of Bsdiff

The directions of the extension are the upper and lower sides of the same code areas. The similarity is evaluated using the percentage of same code and length of the different code. These parameters are 50% and 8 bytes, respectively, in the original bsdiff, i.e., if the code differs by more than 50% or has 8 continuous bytes different, then it is considered as different code.

If the area in the new code has no similarity with the previous code, then it is adjudged to be a different code. Figure 7 shows an example of such a difference. The representation format consists of same areas, similar areas, and different areas (referred to as diff). Table 1 shows the commands of this format.

After representing the difference, the differential data of this format is compressed using a general-purpose compression algorithm (e.g., bzip2). The causes of the differences in similar areas are references that have slid to new addresses. In this case, the difference can be efficiently compressed.

Bsdiff is suitable for Intel i386 CPU architecture, which has non-fixed operation code. Therefore, we proposed an improved method for similar subsequences in programs [2].

The proposed method is based on the following two terms:

(1) There are many similar codes such as function calls.
(2) There are many similar codes such as macro codes.

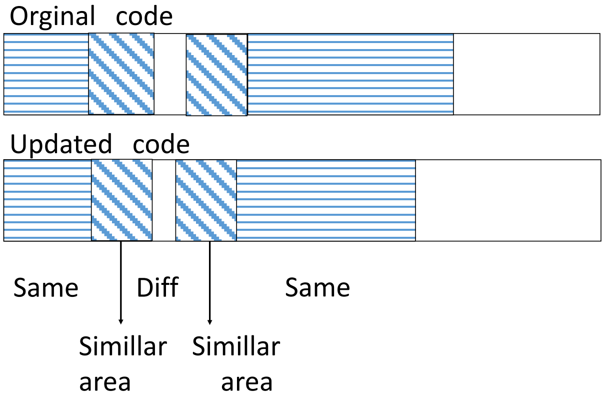


Figure 7 Example of diff

Table 1 Diff format

item		size
header	"BSDIFF40"	8
	Length of similar area	8
	Length of diff	8
	Size of new file	8
Operation block	ADD	8
	COPY	8
	SEEK	8
	…	
Similar area		
Diff area		

In this paper, we focus on other aspects which is bsdiff is for non-fixed operation code.

5 PROPOSED METHOD

In order to optimize the bsdiff algorithm for fixed-operation CPUs, we change the rules of determining the diff area from the additional area, which are shown in Figure 7.

The following are the rules established for determining diff in the original version of bsdiff algorithm:

(1) If the more than 50% of code is changed, the code is additional code. Otherwise, the code is changed code.
(2) If 8 continuous bytes of code are different, the code is additional code.

The features of CPU with fixed operation code are as follows:

(1) Almost all CPUs have a 32-bit architecture. Therefore, in many cases, address data is changed only by 1 byte.
(2) Almost all operation codes require 4 bytes.

Therefore, we propose the parameter should be changed.

6 EVALUATION

Tables 2 and 3 show the sample file specification for evaluation. Table 2 shows the detailed specification for the original software, whereas Table 3 shows the detailed specification for the software after it is updated by codes a, b, and c.

Though these are small programs, the diff algorithm can be evaluated, because, the software size of the sensor device is also small and there are no long jump operations.

The format of executable code is ELF. It is clear from Table 2 and 3 that even if the source code is small, the executable code is relatively large. The reason for this is that the executable code includes standard libraries, which are not changed.

Table2 sample file specification before updating

Source program	Size of file (bytes)	Number of lines	Size of executable file(bytes)
a	946	39	6622
b	464	22	5831
c	792	32	6350

Table3 sample file specification after updating

Source program (updated)	Size of file (bytes)	Number of lines	Size of executable file(bytes)
A	1087	43	7478
B	631	26	5831
C	829	33	6362

Table 4 the size of bspatch

% old → new	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
a→ A	150	150	150	150	150	150	150	150	150	150	150	150	150	150	2860	2860	2860	2860	2860
b→ B	788	788	788	788	788	788	788	788	788	794	794	780	781	789	3152	3152	3152	3152	3152
c→ C	711	711	711	711	711	711	711	712	703	717	729	696	696	3884	3884	3884	3884	3884	3884

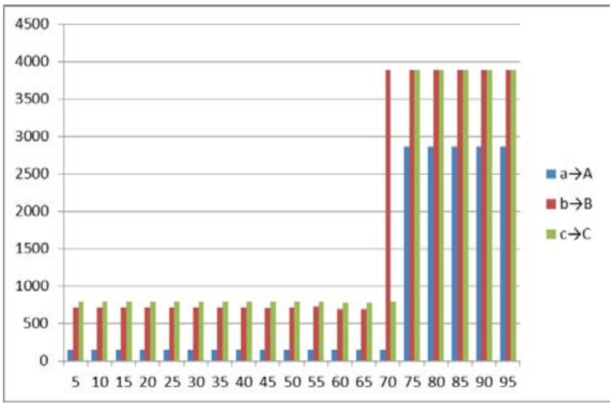


Figure 8 the size of bsmatch

Table 4 shows the size of bspatch for applying our method to each file. The rules to determine similarity using percentage (i.e., threshold) is changed from 5% to 95%. Figure 8 shows the graph for same data. This result shows the size of bspatch is the almost same when the threshold 5% up to 70%. In the case of the threshold being greater than 70%, the size of bspatch is large.

Therefore, the value of threshold is between 60% and 70%. However, it depends on the software. Therefore, if the CPU is powerful enough, in the development environment, the bspatch generation should be repeatedly executed, and the best threshold should be found.

7 CONCLUSION

In this paper, we propose an improved version of bsdiff for CPUs of fixed-operation code architecture. Our proposed method uses a modified threshold value. We evaluate our method using three sample codes. Our evaluation determines that the best threshold value is 70%. However, it is not enough data and depends on the software. Therefore, it is repeatedly applied and evaluated before the bspatch file is generated.

REFERENCES

[1] R. Kiyohara, M. Kurihara, S. Mii, and S. Kino, “A delta representation scheme for updating between versions of mobile phone software,” Electronics and Communications in Japan, Vol.90, No.7, pp.26-37, 2007.

[2] Colin Percival, “Matching with Mismatches and Assorted Applications ,” doctoral thesis, Wadham College University of Oxford http://www.daemonology.net/papers/thesis.pdf <accessed 2015/11/08>

[3] The VCDIFF Generic Differencing and Compression Data Format, http://www.ietf.org/rfc/rfc3284.txt.

[4] xdelta, http://xdelta.org/

[5] Ryozo Kiyohara, Satoshi Mii, Mitsuhiro MatsumotoMasayuki Numao, and Satoshi Kurihara、A new method of fast compression of program code for OTA updates in consumer devices 、 IEEE Transactions on Consumer Electronics, Vol.55, Issue 2,pp.812-817, 2009

[6] Ryozo Kiyohara, Satoshi Mii, Koichi Tanaka, Yoshiaki Terashima, and Hidetoshi Kambe、Study on Binary Code Synchronization in Consumer Devices、IEEE Transactions on Consumer Electronics, Vol.56, Issue 1,pp.254-260, 2010.

[7] J．W． Hunt and T．G． Szymanski， ” A fast algorithm for computing longuest common subsequences， ” Commun． ACM， vol． 20， no． 5， pp． 351-353， 1977

Visualizing Wireless Sensor Networks for Practical Network Management

Yuki Urata[†], Takuya Yoshihiro[‡], Yutaka Kawahashi*

[†]Graduate School of Systems Engineering, Wakayama University, Japan
[‡]Faculty of Systems Engineering, Wakayama University, Japan
* Center for Information Science, Wakayama University, Japan
{s171009, tac, yutaka}@center.wakayama-u.ac.jp

Abstract - To realize practical wireless sensor networks, many studies have been trying to develop low-energy technologies to realize long-life sensor devises. Thus, several studies have tried a passive management methodology that gets to know the network state from several measured information attached to data packets and collected to the sink nodes. Though these existing studies basically aim at detecting root causes after failure occurrence, finding signs in advance of failure is also important to keep the performance of networks as an infrastructure from the experience of the Internet management. In this paper, we design and develop a new system to visualize network states to manage wireless sensor networks. Our design is based on the guideline of the infrastructure (i.e., the Internet) network management. Through evaluation, we confirmed that the proposed system design enables operators to watch networks and aware of every change and symptom that should be noticed by operators.

Keywords: Wireless Sensor Networks, Operation, Management, Visualization

1 INTRODUCTION

Wireless Sensor Networks (WSNs) are expected as a practically important technology to support the up-coming IoT (Internet of Things) world. However, to realize practical WSNs, reliable communications have to be done within very low power consumption. For this challenge, various communication protocols as well as sensor-node hardwares that consume extremely low power have been developed so far.

As a well-known communication standard for sensor networks, IEEE802.15.4[1] has been promoted, and many commodity devices have been developed. However, IEEE802.15.4 in fact consumes considerable power because it is designed to cover wide variety of practical scenes, and also to include many functions to enhance flexibility of communications. Thus, to achieve a minimum power consumption level to collect sensed values to sink nodes, lots of low-power MAC protocols for WSNs have been proposed [2][3][4][5].

On the other hand, similar to the networks, we have to manage and operate WSNs in practice to keep WSNs work correctly to gather sensed data values. To this end, we need a system with which we can watch the state of WSNs and find the trouble as soon as it occurs. There are several tools and methods developed for this purpose. In early days, query-based methods to collect value of WSNs has been tried, which is a similar approach to SNMP [6] in the Internet management. However, in low-power duty-cycle MAC protocols, such queries require quite long delay due to long sleeping time of each node, and thus it is impossible approach in WSNs.

Alternatively, one of the basic approaches is to collect required values to sink nodes by piggy-backing them on data packets to find anomalies of WSNs. There are several proposals from this approach: Ramanathan et al. proposed a system Sympathy that identifies the root causes by constructing decision trees [7]. Liu et al. proposed a method PAD that finds the root causes by constructing causal diagrams [8].

These methods carefully model the relationship among possible events to specify the essential causes of troubles. However, since the events that are considered in them are limited to the ones that are supposed in advance, new phenomena cannot be treated. As the Internet management process poses, the troubles in network management have too large variety, so that those proposals cannot cover all possible cases. Also, finding a small sign of trouble before it occurs is an important issue for prevention of troubles. Thus, in addition to the predefined diagnose systems, we need a system that visualizes the state of WSNs to help finding troubles or their prior signs in managing and operating WSNs.

In the literature, there are several studies that visualize WSNs such as Octopus [9], SRNET[10], etc. Also, several visualization tools have been provided as bundled software in WSNs devices or implementations; they are introduced in a survey paper [11]. However, they are not designed from the viewpoint of managing networks to utilize the experiment of Internet management, nor evaluated from the viewpoint.

In this paper, we propose a new design of systems that visualize the state of WSNs and help managing them. Our system design is based on the principle of Internet management, i.e., we designed to perform (a) structure management, (b) failure management, and (c) performance management. We implemented and evaluated our system using simulation traces of a WSN to confirm that the system enables users to find important signs from the viewpoint of principles (a)-(c).

This paper is organized as follows. In Sec. 2, we describe the design policy of WSN management systems. Especially, we discuss on the principal of Internet management and requirements for WSN management systems. In Sec. 3, we present the proposed system, and we evaluate the system in Sec.4. Finally in Sec.5, we conclude the work.

2 REQUIREMENTS AND THE DESIGN

2.1 Requirements for WSN Management Systems

In this study, we suppose wireless sensor networks (WSNs) that deploy low-energy MAC protocols such as B-MAC and RI-MAC, so that any query-based management protocol such as SNMP cannot be used to manage them.

In managing WSNs, we are required to carefully watch the networks to keep them correctly working for collecting sensed data values. For this purpose, not only detecting the problems occurring, but also finding implicit signs of future problems is important. One of our primary goals is to make network administrators possible to find these signs surely and speedily by visualizing the values collected at sinks. Note that it is difficult to manage multi-hop WSNs in real time as long as it is managed in the passive way, i.e., managed through the values collected at sink nodes. In other words, we have to allow a certain level of delay on finding failures or troubles in the network management tasks. If we have a requirement on the delay performance, they should be covered by the deployed MAC and routing protocols. Accordingly, we in this paper focus on just inquiring the causes of the failures without caring the time delay that takes to find them.

On the other side, in the Internet, network management has been a critically important issue to provide stable and secure services to end users since the Internet now has a role of social Infrastructure. In this context, the area of network management has naturally been grown in the past decade and formed a rough but firm consensus on how to manage networks. Such consensus is issued as some documents, e.g., reference [12] describes the required knowledge and administrative operations that should be performed in managing networks. The document says that the network management consists of 5 specific aspects of management domainsshown in the following.

Structure Management
: Structure management is the task that maintains physical and logical elements in networks. Network structure is the basis of all management task, so it is siginificantly important to grasp the latest state of the network structure. In contrast, the structure of current WSNs is quite simple in which sensor devices, their configurations, neighbor relationship, and the paths from each node to sinks are included.

Failure Management
: Failure management is the task (1) that defines the event regarded as failure, (2) considers the detecting strategy, countermeasures and preventive measures for each failure, (3) and executes them. We have two types of the failure detection approaches: active detection based on queries sent to each device, and passive detection based on the reports coming from each device.

Performance Management
: Performance management is the task that maintains WSNs to keep a constant level of communication performance. Generally, the performance of networks include such as throughput, packet loss ratio, latency, jitter, congestion frequency, etc. In WSNs in this study, we expect each packet to reach a sink reliably within a certain latency. Thus, especially packet loss ratio and latency are the important measure of the performance.

Resource Management
: Resource management is the task that maintains the system resources required in the operation of systems. Note that resource management includes CPU, memory, and network capacity management that prevents shortage of those dynamically used resources.

Security Management
: Security management is the task that protects the system and the contents from the threat outside of the network. In WSNs in contrast to the Internet, the main concern is the information leaking, which can be prevented by deploying some encryption facility.

In this study, we only consider three management aspects, i.e., structure, failure, and performance management, and omit considering resource and security management. Resource management includes CPU and memory management, However, WSNs use small amount of CPU and memory resources so that they are not important to manage in most cases. As above, in WSNs, three management issues, i.e., structure, failure, and performance management, are essentially important.

2.2 The System Design

For structure, failure, and performance management, the system must collect the required administrative information from WSNs and visualize them appropriately to help operators manage WSNs properly. Collecting information is done such that each node measures several administrative values and include them in the packet that the node generates. By piggy-backing the values required for management on packets, passive management using the values gathered on the sink nodes is enabled. Note that the administrative values are not added at each node in the collection paths, but added only at the originated node of the packet. Therefore, the overhead incurred from the administrative values does not change depending on the size of WSNs.

User interface is designed in order for operators to easily find events related with the three management aspects. For structure management, we prepare the *delivery tree view* to see the network state intuitively. In addition, we can overlap two delivery trees at different time points, which enables us to see the transition of the delivery tree.

In failure management of WSNs, when the packets from a node do not arrive at sinks, we can regard that the node or link failure. Thus, we prepare the *alert table* in which possible node failure events as well as other alerts are listed up to check whether the failure really occurs or not. In addition, to operate any failures, we prepare the *administrative values table* in which all the data values collected in sinks are seen per source node, and also prepare the *line graph view* that visualize the data values per node and per item to see the values intuitively.

Finally, for performance management, we mainly watch residual power, packet loss ratio, and delivery delay at each node. By listing the events in the *alert table* when these values

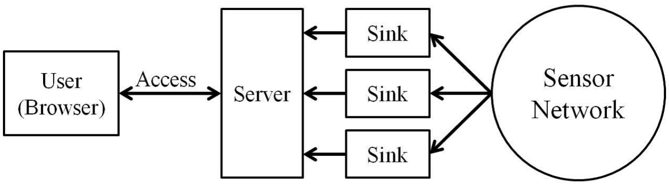


Figure 1: System Structure

exceed a threshold, the operators easily find the performance degradation. After the operator is notified of the performance anomalies, they usually explore for the level of the degradation and specify what is the root cause of this trouble. We can use the *administrative values table* and the *line graph view* again for this purpose.

With the basic design policy described above, our system has the following characteristics that differentiate our system from the others.

(1) Visualizing Delivery Tree Transition
The function of overlaid display of multiple delivery trees obtained from different time points is unique to our system, which enables operators to grasp the transition of delivery paths easily and intuitively.

(2) Alert Table to be Aware of Administrative Events
We prepare the *alert table* that makes administrators keep aware of important administrative events occurred in WSNs. This is done by simply applying thresholds to several carefully-selected administrative data items

(3) Intuitive Operation
From the base *delivery-tree view*, we can intuitively transit to other views by clicking entities nodes and buttons. In our user interface design, the administrators are possible to access the related data values to explore for the state of WSNs.

3 THE PROPOSED SYSTEM

3.1 System Structure

The proposed system visualize the network state from the administrative values collected to sinks. We show the system structure in Fig. 1. A sink node collects the sensed values generated periodically at every node. Note that there may be multiple sink nodes in a WSN, but the server collects values from all sink nodes. The server provides the function of web servers so that the administrators access to the server via Web browsers to visualize the state of WSNs.

3.2 Attaching Administrative Values to Packets

In our framework, we piggy-back the administrative values on packets to collect them to the sinks in WSNs. As the administrative values, we used the following 18 items that are typically used in administrating networks. Note that every item is measured at each node by itself.

Items: (a) Reception Time at Sink, (b) Sink Node ID, (c) Source Node ID, (d) Sequence ID, (e) Parent Node ID(Next-hop of Source Node), (f) Number of Transmitted Frames per Unit Time, (g) Number of Received Frames per Unit Time(Incl. Overhead Frames), (h) Number of Transmitted ACKs per Unit Time, (i) Number of Received ACKs per Unit Time(Incl. Overhead Frames), (j) Number of Transmitted Control Messages per Unit Time, (k) Number of Received Control Messages per Unit Time, (l) Number of Received Frames per Unit Time(Excl. Overhead Frames), (m) Number of Received ACKs per Unit Time(Excl. Overhead Frames), (n) generated Time of Sensed Values, (o) Accumulated Awaking Time per Unit Time, (p) Accumulated Sleep Time per Unit Time, (q) Sensor Coordinate(If device is with GPS), (r) Residual Power.

Most items above are well-defined but we would add an explanation for several items. Note that these administrative values are added to data packets only when a sensor value is measured and the corresponding data packets are generated, and not added at the relay nodes.

Item (d) is a value uniquely assigned to each packet by a node, which is used to check the loss of data packets. Items (f)-(l) are the values measured per unit time, where typically counting is done after previous sensed-value generation since we assume sensing is done periodically at each node. Note that the number of transmissions (f)(h)(j) include not only the frame generated at the node but also the frame that the node relays to sinks. Also note that several items have two sort of values that include and exclude overheard frames, where overhearing is the event in which a node receives a frame that does not destine the node itself. The number of overheard events can be useful in a specific case, e.g., to measure the congestion ratio around the node. Item (r) represents residual power at each node when the packet is generated.

Sensor location (q) may be collected at sink, but we have to consider that many sensor-node devices do not have GPS due to large power consumption, and also due to relatively large errors in computing positions.

From the items (a)-(r) above, we compute several values useful for managing WSNs shown as follows. (s) Packet Loss Ratio of the Next-hop Link per Unit Time, (t) Delivery Delay, (u) Elapsed Time after Last Frame Reception from Each Node at Sink.

Item (s) represents a quality of next-hop link that is computed from the transmission count (f) and the ACK reception count (m) using the formula (s)= $\{((f)-(m))/(f)\}\times 100[\%]$. Item (t) is the average time of packets taken to travel to sinks from the packet is generated. This value is computed as an average of (t)=(a)−(n) for each packet. Item (u) implies a potential anomaly if it is far larger than the sensing time interval. This value is computed from item (a): if we let a_n be the arrival time of a packet with sequence number n, (u)= $a_n - a_{n-1}$.

Note that those three values (r)-(u) are especially important in managing WSNs since worse values immediately imply performance degradation of WSNs. Thus, in our system, we set a threshold value for each of (r), (s), (t), and (u) so that the system can notify administrators of the abnormal state through the *alert table*.

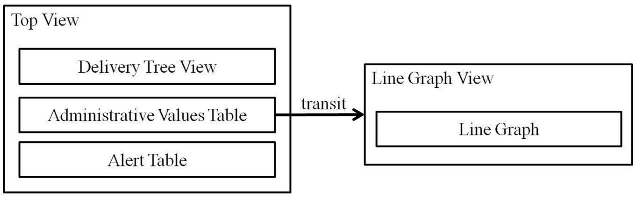


Figure 2: User Interface

3.3 User Interface

3.3.1 Transition of Views

The transition of user's view in our system is shown in Fig. 2. In the top view, the delivery tree is displayed to show the overview of the current state of the WSN. The administrative values table and the alert table are aside of delivery tree to show the specific data values and important alerts to notice.

At the top of each item in administrative values table, we placed a button that pop-up a new window and display the line graph of the specified data item. In combination of those four components, administrators can watch the state of WSNs and explore the detailed behavior to find what is going on under troubles.

3.3.2 Delivery Tree View

Delivery tree view displays the delivery tree at the specific time point. Normally it would display the latest tree, but user can specify arbitrary time to see the tree at that time point. Delivery tree basically consists of a set of nodes which are placed at the right position and the set of next-hops to which each node forwards packets destined to sinks. In managing WSNs, we heve to know the place of each node, so we assume that the coordinate of each node is known by some mean, for example, GPS or static map that include the coordinate of each node. To show the transition of the delivery tree in time, our view has a function to overlay the past or the future delivery tree over the tree of the current time point. See Fig. 3 for this overlaid view. The current delivery tree is shown with solid blue lines whereas the past tree with dotted pink lines. The past tree consists of the previous next-hop links that are different from the current one and are reported not before the time $c - t$ where c is the current time and t is a predetermined threshold. Overlaying the future delivery tree is done in the similar way. By labeling links with the time reported (i.e., the generation time of reported packets), administrators can understand the transition of trees with exact reported time.

3.3.3 Administrative Values Table

The administrative values table shows all the administrative values collected at sinks per node. We show an example of the administrative values table in Fig.4. By clicking a node in the delivery tree, the administrative values table is updated to show the values sent from the clicked node around the time specified in the delivery tree view. With the administrative values table, administrators can refer the required administrative values intuitively and efficiently from large amount of data. Also, the button on top of each column invokes the line-graph window to see the data values intuitively.

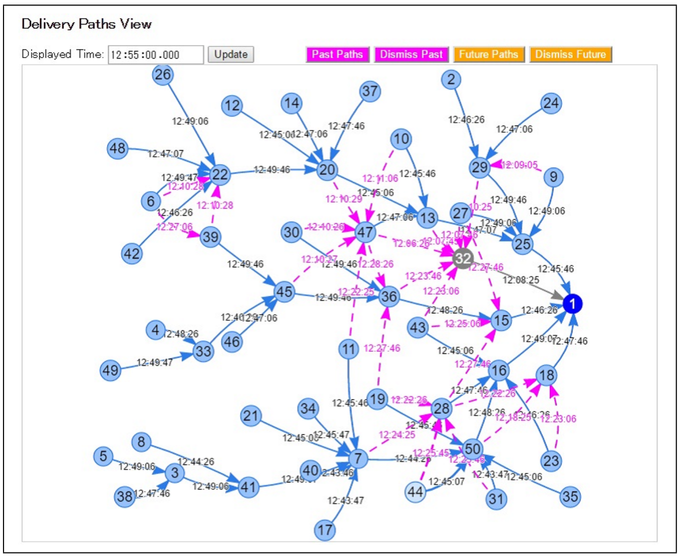


Figure 3: Delivery Tree View (When Two Time-points are Overlayed)

3.3.4 Alert Table

The alert table has a role to keep administrators being aware of the important signs of WSN state changes. Especially, performance changes are essentially important to notify since they are invisible in the delivery tree so that not easy to notice for administrators. To this end, as mentioned in Sec.3.3.2, we display alert messages on communication performance by applying a threshold for each data items (r), (s), (t), and (u). An example of the alert table is shown in Fig. 5.

3.3.5 Line Graph View

The line graph shows the values listed in the administrative values table in the line graph fashion to enable administrators to grasp the trend of values intuitively. An example of the line graph view is shown in Fig. 6. As shown in this example, administrative values are plotted in time series where the horizontal axis is the reported time of the value.

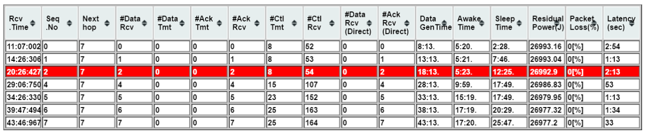
4 EVALUATION

4.1 Implementation

The proposed system is implemented as a web application implemented using javaScript and AJAX. We use a library vis.js[13] for graph visualization, and highcharts[14] for drawing line graphs. In our system implementation, our application runs on Apache ver.2.4.6[15].

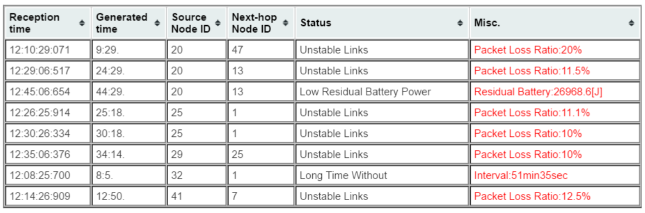
4.2 The Evaluation Method

We run simulation of WSNs to obtain a trace from a self-designed scenario, and visualize it using the proposed system. For the simulation, we implement a receiver-initiated MAC



Rcv .Time	Seq .No	Next hop	#Data Rcv	#Data Tmt	#Ack Tmt	#Ack Rcv	#Ctl Tmt	#Ctl Rcv	#Data Rcv (Direct)	#Ack Rcv (Direct)	Data GenTime	Awake Time	Sleep Time	Residual Power(J)	Packet Loss(%)	Latency (sec)
11:07:002	0	7	0	0	0	0	8	52	0	0	8:13.	5:20.	2:28.	26993.16	0[%]	2:54
14:26:306	1	7	1	0	0	1	8	53	0	1	13:13.	5:21.	7:46.	26993.04	0[%]	1:13
20:26:427	2	7	2	0	0	2	8	54	0	2	18:13.	5:23.	12:25.	26992.9	0[%]	2:13
29:06:750	4	7	4	0	0	4	15	107	0	4	28:13.	9:59.	17:49.	26986.83	0[%]	53
34:26:330	5	7	5	0	0	5	23	152	0	5	33:13.	15:19.	17:49.	26979.95	0[%]	1:13
39:47:494	6	7	6	0	0	6	25	163	0	6	38:13.	17:19.	20:29.	26977.32	0[%]	1:34
43:46:967	7	7	7	0	0	7	25	164	0	7	43:13.	17:20.	25:47.	26977.2	0[%]	33

Figure 4: Administrative Values Table



Reception time	Generated time	Source Node ID	Next-hop Node ID	Status	Misc.
12:10:29:071	9:29.	20	47	Unstable Links	Packet Loss Ratio:20%
12:29:06:517	24:29.	20	13	Unstable Links	Packet Loss Ratio:11.5%
12:45:06:654	44:29.	20	13	Low Residual Battery Power	Residual Battery:26968.6[J]
12:26:25:914	25:18.	25	1	Unstable Links	Packet Loss Ratio:11.1%
12:30:26:334	30:18.	25	1	Unstable Links	Packet Loss Ratio:10%
12:35:06:376	34:14.	29	25	Unstable Links	Packet Loss Ratio:10%
12:08:25:700	8:5.	32	1	Long Time Without	Interval:51min35sec
12:14:26:909	12:50.	41	7	Unstable Links	Packet Loss Ratio:12.5%

Figure 5: Alert Table

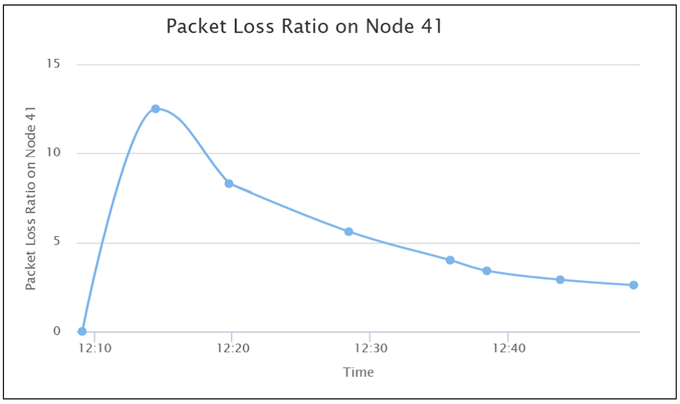


Figure 6: Line Graph View

protocol combined with a low-power routing protocol proposed in [4]. We implement these protocols on Contiki[16], which is an OS for sensor network devices, and simulated it over simulator Cooja included in Contiki OS package. As a simulation scenario, we placed 50 nodes at a random coordinate in a 200[m]×200[m] rectangular field, and we placed a sink node on the central position of the left side of the field. Each node generates a sensed value in 5[min] interval, and sends periodical beacons in 40[sec] interval. We run the simulation in 60[min], but we intentionally invoke a node failure about the middle of the simulation time. Other parameters related to the simulation is shown in Table 1. As a result of simulation, we obtained a set of trace data set, i.e., the set of administrative values collected at the sink.

As for the parameters on the proposed system, we used the threshold values to generate alert messages in Table 2. These values are determined through previously performed test runs.

Evaluation process is in the following. We prepare the proposed system that loaded the above trace data set. We asked one subject, who is a student who have studied both the Internet management and sonsor network protocols, to use the system for 20 minutes and also asked to list up the events that is important from the viewpoint of three management aspects, i.e., structure, failure, and performance management. After the above operations, we carefully examined the trace data set to find all the events that administrators should be aware of. We confirmed whether the subject can find all the events or not.

Table 1: WSN Environment

Simulation Start Time	12:00
Simulation Time	60[min]
Field Size	200[m]×200[m]
Number of Nodes	50
Communication Range	Circle with Radius 50[m]
Beacon Interval	40[sec]
Sensing Interval	5[min]
Battery Capacity	27000[J]

Table 2: Threshold for Alert Table

Packet Loss Ratio	10%
Delivery Delay	60[sec]
Elapsed Time after Last Frame Reception	30[min]

4.3 Results

By examining the events that the subjects found, we evaluate the practical efficacy of the proposed system. First of all, we say that the subject found all the events that we judged administrators should be aware of. This result shows that the proposed system works well and useful in managing WSNs. In the following, we see the detailed description seen from the three aspects of network management, i.e., structure, failure, and performance management.

4.3.1 On Node Failure (Failure Management)

The subject indicated the failure of node 32, which we intentionally did in the middle of the simulation time. From hearing from the subject, this event was found by watching the delivery tree view. Since packets from node 32 has not been reached the sink for a long time, next-hop of the node 32 disappears and the color of node 32 turns gray. Simultaneously, this event is displayed in the alert table. In this way, the system provides several mechanisms to help administrators find node failure, and the subject actually found failure of node 32 in our evaluation test. We confirmed that the failure management was well performed using our system.

4.3.2 On Path Transition (Structure Management)

The subject explained how and why delivery tree changes as time passes. This is also possible using delivery tree view. When the subject saw the delivery tree just after failure of node 32, which is shown in Fig. 7, he found that all nodes whose next-hop was node 32 changed their next-hop one after one. Also, the subject found that, node 20 observes far larger number of control frames than usual just after node 32 fails. Note that the deployed MAC and routing protocol [4] transmits far larger number of control messages in face of topology changes to speed up the tree reconstruction. The subject observed this behavior of nodes, which also supports that the root cause that changes next-hops is failure of node 32. As above, we confirmed that the structure management was well performed.

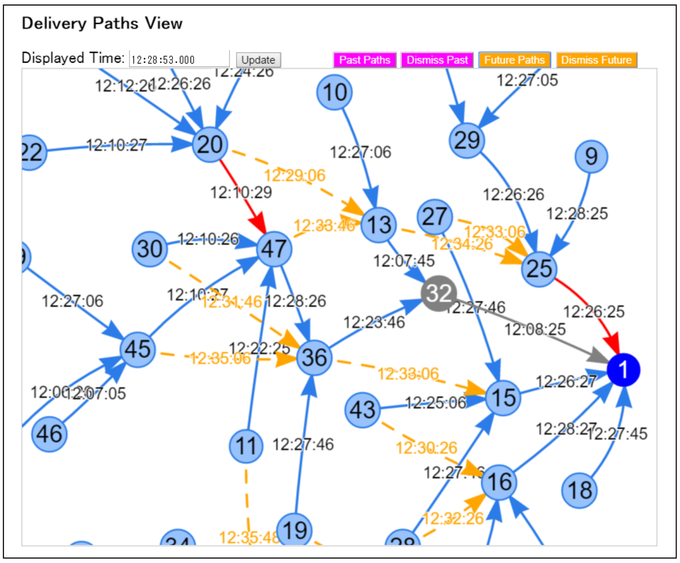


Figure 7: Delivery Tree View in Tree Transition

4.3.3 On Performance Degradation (Performance Management)

As for the performance of the network, the subject indicated that the nodes 20 and 41 had high frame loss ratio, which is found by the alert table and the delivery tree view. When the subject examined the network state around node 20, several nodes adjacent to node 20 transmit control frames more frequently. So, the subject naturally judges that those control messages are the cause of the congestion and the frame loss. As for node 41, he found that the next-hop of node 41 does not receive frames for a certain time period for some reason. Although node 41 repeatedly transmits a frame for the time period, the next-hop does not receive it, nor the delivery tree is not changed. This is an inconsistent behavior. Thus, it is inferred that the root cause would be a bug in program codes, or some kind of design errors of protocols. As above, we confirmed that performance management was also well performed.

In summary, through the evaluation test, we found that the subject performed all the tasks of failure, structure, and performance management. Consequently, we confirmed that the proposed system works well in a practical scenario of WSN management.

Additionally, note that our result implies that four basic components (i.e., delivery tree, administrative values table, line graphs, and alert table) are sufficient for a management tool of WSNs. However, from the viewpoint of functional design, the propose system can offer a new principle of WSN management tools.

5 CONCLUSION

In this paper, we proposed a system to manage WSNs that deploy low-power MAC protocols. We designed the system from the viewpoint of the network management in the Internet, i.e., we aim at executing structure, failure, and performance management. We evaluated the proposed system using a simulation trace data set. As a result, all the events that should be noticed are listed up by the subject, while all three management aspects are well managed, showing that the proposed system possibly works effectively in the real WSN management operations.

To apply the system to the real environment instead of simulation is one of the important task for the future.

ACKNOWLEDGMENT

This work is supported by JSPS KAKENHI (15H02691).

REFERENCES

[1] IEEE802.15.4, http://www.ieee802.org/15/pub/TG4.html

[2] J. Polastre, J. Hill, and D. Culler, ”Versatile Low Power Media Access for Wireless Sensor Networks,” In Proc. of SenSys’04, pp.95–107, 2004.

[3] Y. Sum, O. Gurewits, and D. B. Johnson, ”RI-MAC: A Receiver-initiated Asynchronous Duty Cycle MAC Protocol for Dynamic Traffic Loads in Wireless Sensor Networks,” In Proc. of SenSys’08, pp.1-14, 2008.

[4] S. Kojima and T. Yoshihiro, “A Low Management Cost Wireless Sensor Network Based on Receiver Initiated MAC Protocols,” Journal of Information Processing, Vol.57, No.2, pp.480?-493, 2016 (In Japanese).

[5] X. Fafoutis, A.D. Mauro, M.D. Vithanage, N. Dragoni, Receiver-initiated medium access control protocols for wireless sensor networks, Computer Networks, Volume 76, Pages 55-74, 2015.

[6] SNMP Standard, http://www.snmp.com/

[7] N.Ramanathan, K.Chang, R.Kapur, L.Girod, E.Kohler, D.Estrin ”Sympathy for the Sensor Network Debugger,” In Proc. Sensys’05, 2005.

[8] K. Liu, M. Li, and Y. Liu, ”Passive Diagnosis for Wireless Sensor Networks,” IEEE/ACM Transactions on Networking, Vol.18, Issue 4, pp.1132–1144, 2010.

[9] R.Jurdak, A.G. Ruzzelli, A. Barbirato, and S. Boivineau, Octopus: Monitoring, Visualization and Control of Sensor Networks, Wireless Communications and Mobile Computing, Vol. 11, Issue 8, pp. 1073–1091, 2011.

[10] E. Karapistoli, P. Sarigiannidis, and A.A. Economides, SRNET: A Real-time Cross-based Anomaly Detection and Visualization System for Wireless Sensor Networks, In Proc VisSec’13, 2013.

[11] B.Parbat, A.K.Dwivedi, O.P.Vyas, ”Data Visualization Tools for WSNs: A Glimpse,” International Journal of Computer Applications, Vol.2, No.1, pp.14–20, 2010.

[12] Information-technology Promotion Agency, Japan (IPA), “Knowledge for Network Management I,” Open Source Software Model Curriculum Version 1, https://jinzaiipedia.ipa.go.jp/wp-content/uploads/oss/basic_Guidance_12.pdf

[13] ”vis.js,” http://visjs.org/

[14] ”Highcharts,” http://www.highcharts.com/

[15] ”The Apache Software Foundation,” http://www.apache.org/

[16] Contiki, http://www.contiki-os.org/ (accessed at 1st Feb 2016).

A Study of Collaboration with “Aware Wheelchair” with Sensor Networks for Safe Actions in Home

Taizo Miyachi*, Gulbanu Buribayeva*, Saiko Iga** , and Takashi Furuhata***

*School of Information Science and Technology, Tokai University, Japan
**ICT Center, Tokai University, Japan
***University of Utah, USA
{198905 }@tokai-u.jp

Abstract - A wheelchair is very important technical aid for an independent and healthy life in super-ageing society since a few tens presentation older people need wheelchairs. However the number of accidents by a wheelchair was the worst among all technical aids (walking frame, bed, shower chair, etc.) in 2012 in Japan since both a wheelchair user and caregiver have blind spots and sometimes cannot find dangerous obstacle in a wide 3D space. We investigate safe collaboration methodologies by “Aware Wheelchair” with sensor networks that detect dangers and distribute awareness and suggest suitable actions to both user and caregivers. Aware Wheelchairs also send sensor data to a care server and accumulate them in the server. We also discuss how to give them effective awareness from psychological view point and reduce stress in order to utilize the awareness for safe actions in the collaboration both in active psychological ways and by automatic mechanical assists corresponding to personal type among different generation and physical ability. Real-time sharing of serious accidents would let the users and caregivers avoid similar accidents in similar situations.

Keywords: aware wheelchair; sensor network, collaboration for safe action, acoustic guide, declaration of intention

1 INTRODUCTION

A wheelchair is very important technical aid for an independent and healthy life in super-ageing society. Body exercise is indispensable for keeping health and physical ability. Therefore most wheelchair users use manual wheelchairs in Japan. However, a research in Japan 2012 showed that a wheelchair caused many more accidents and injuries than all the other technical aids (e.g. a walking frame, bed, shower chair, etc.) [Nihon Wheel Chairs Co. Ltd, 2012]. Human is very excellent but not almighty. Human cannot notice variety of dangers in blind spots. Both a wheelchair user and a caregiver usually cannot find danger in many kinds of blind spots. Human is also not good at finding a small jut in a wide 3D space by visual information. S(he) sometimes hit a foot or a hand, etc. by invisible obstacles and such a jut.

We propose “Aware Wheelchair (AW)” with a sensor network [9] that detects “dangers within two meters” and sends voice awareness to both a user and a caregiver in order to take safer actions in 3D space. Both a wheelchair user and a caregiver should carefully check invisible dangers and risks in a route since accidents by wheelchair was the worst in all the technical aids. It is not easy for them to safely control moved wheelchair and suddenly stop without troubles. We investigate on intension of the wheelchair user and sharing intentions between the wheelchair user and the caregiver.

2 ACCIDENTS, PHYSICAL ABILITY, INTENTION, AND STRESS

Most wheelchair users use manual wheelchairs in Japan because body exercise with technology aids is necessary for making elderly people and impaired persons healthier. The exercise also influences brain clearer and a user get great sense of achievement. However, human cannot aware all variety of dangers. The worst rates of accidents happened by technology aids were wheelchair (48.6%), and bed (13.3%) [6] (See Fig. 1). A wheelchair user (WU) should safely take actions several times in a day although s(he) does not move on a wheelchair or a bed in a whole day.

Physical ability, blind spot, physical phenomena, misunderstanding of intention, and stress, etc. would cause accidents. The major causes of wheelchair accidents can be classified into ten categories.

Causes.

(C1) Slow recognition and slow reaction for fast moving wheelchair. Human needs a longer time period than several 0.1 seconds to find danger and take an action to avoid, for example, a foot blow. A wheelchair usually moves too fast for an aged person to quickly stop it in a few seconds.

(C2) Blind spot. Human usually cannot find dangers in a blind spot (Fig. 1). S(he) cannot think of the serious damage between invisible foot and invisible obstacle.

(C3) Small jut and a rough surface of flat board. Human is not good at finding a small jut that scratches a leg in a wide 3D view although s(he) would be seeing around with the wide view. Human cannot also find dangers in a rough surface of flat board

(C4) Difficult simulation of moving obstacle. Human cannot recognize an exact path image of moving object in 3D space. S(he) cannot also avoid a collision between it and a part of his/her body.

(C5) Tired caregiver’s mistake. A

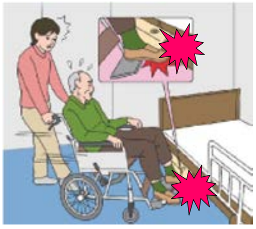


Fig. 1 An accident with wheelchair use [6]

tired caregiver sometimes could not find dangers and caused serious accident. Some operations of a technical aid need heavy physical load and are stressful. Users and caregivers usually get tired by such operations with embedded dangers. They sometimes suffer from several kinds of arthritis in several years. Many caregivers got tired everyday and left their jobs. The job of separation rate has been increased.

(C6) Misunderstanding of unfamiliar user's ability by a sub-caregiver. A sub-caregiver could not safely care an unfamiliar wheelchair user while two caregivers were collaborating. Misunderstanding of the unfamiliar user by the sub-caregiver caused an accident.

(C7) No record of sensor data and environments in many cases of both close calls and accidents. A user and caregivers could not record dangerous contexts and situations related to both human and environments.

(C8) A caregiver does not know correct care in a moment when s(he) suddenly happen to meet a new danger.

(C9) Falling down by physical phenomena, such as gravity and inertial force.

(C10) Stress and distruction. A caregiver has accumulated stress for not only his/her jobs but also unpredictable actions by a wheelchair user. We call this "accumulation type of stress." It is also uneasy for them to quickly control the speed of wheelchair and suddenly stop when they suddenly meet some danger.

A WU expects to safely take easy actions since s(he) would be impaired person and lack some physical ability. On the other hands, a caregiver expects to take effective actions in a short time. Therefor a gap of intentions "Intention Mismatch (IM)" between WU and caregiver sometimes happens. Caregivers take care of such IM.

3 AWARE WHEELCHAIR AND INTENTION SHARING FOR SAFE ACTION

WU and caregivers usually take safe actions in daily life. They sometimes caused accidents when they confronted difficult situations or caused negligence by stress in hard works. They could avoid or cope with most dangers in close cooperation if dangers could be discovered and WU and caregivers could know them ahead of time.

We propose "Aware Wheelchair (AW) with a sensor network" that could be supplements of the weak points of human such as blind spots, and slow recognition etc. in order to avoid accidents with wheelchair and bed (See Fig. 2). AW also gives both WU and caregivers a chance of start collaboration to take cooperative safe actions.

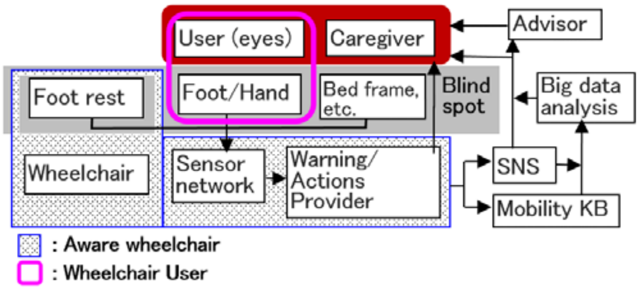


Fig 2. A feature of assist system for aware wheelchair

Aims of AW are mainly four categories.

Aims.

A1. Avoid a danger around and take safe actions

A2. Reduce both stress and load in each operation of a technical aid and make control of wheelchair, etc. easy.

A3. Let a user take self-actualizations with his/her intension in some short time period

A4. Give a chance of starting safe collaboration between a WU and caregivers with a shown same goal and share intentions for avoiding dangers or risks.

Functions.

AW finds a danger in blind spots and provides both a wheelchair user and a caregiver with both awareness and candidate solutions for the dangers. AW also gives them a chance of start collaboration in order to share an intention and take cooperative safe actions by providing the awareness. On the other hand, AW sends sensor data to SNS and mobility knowledge base server [8]. Expert advisers can know the message in the SNS and could send effective advices to both WU and caregivers. AW allows users, caregivers, and the advisers to analyze Big-data of both close calls and accidents, and to find dangerous spots and collaborative solutions for dealing with the dangers.

Safe collaborations between WU and caregivers are very important since actions by WU and caregivers sometimes caused accidents. The collaboration mainly consists of six sub-collaborations.

(SC1) Caregivers should share WU's intention for a goal. WU should show caregivers his/her intention (See Fig. 3).

(SC2) Caregivers and WU should choose safe and effective actions for the goal corresponding to both WU's physical ability and state of health.

(SC3) Caregivers should carefully assist WU's actions based on the shared intention. They should say a shared intention for a goal and remind check points and embedded dangers. Then they should carefully take safe actions.

(SC4) WU should declare himself/herself his/her intention for a goal and get feeling of attainment in case of success.

(SC5) Remote experts should advice a weak WU with effective solutions monitoring SNS.

(SC6) Caregivers should talk the shared intentions and

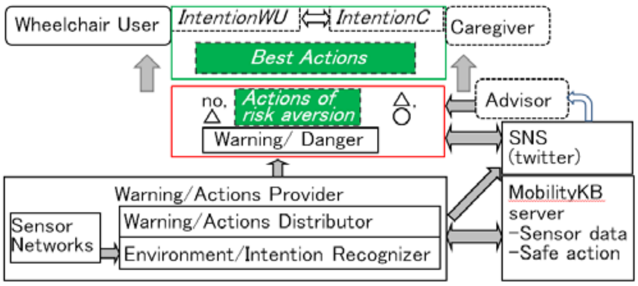


Fig 3. Sharing intentions and collaboration by a warning

A typical example of AW tasks for intensions of the wheelchair user and sharing intentions between the wheelchair user and the caregiver consists of six steps.

(S1) detect and predict dangers and risks
(S2) send warning to a wheelchair user, caregivers, and remote advisor in order to start avoiding both dangers and risks early to give himself/herself plenty of time
(S3) show recommended solution to WU and caregivers
(S4) let caregiver assist the lack of physical ability of WU
(S5) let both WU and caregivers cry with intention presentation to them for careful check of risks and finding the best solution.
(S6) Decrease stress of both caregivers and WU in order to avoid distraction for detecting dangers and to get feeling of attainment

4 EXPERIMENT

Wheelchair users and caregivers could avoid accidents or decrease serious injuries if sensor network could find a danger and distribute warnings to them ahead of time. However, an aged person cannot quickly take safer actions for a sudden situation. A caregiver should share an intention of WU and collaborate with WU so as to timely take safe actions although they cannot always take correct actions for unpredictable situations with a panicked user. We experimented a scene of reaching nearby a bed and moving from a wheelchair to a bed by a warning from a sensor network system with real-time sensor data and safe action knowledge in order to investigate how people collaborate to take safe actions or assist in a blind spot, and reduce stress in hard situations.

Conditions. Suppose, (A) wheelchair user, and (B) caregiver of the wheelchair user reach nearby the bed. The user is going to move to the bed from the wheelchair.
Devices. A wheelchair with a sensor network by ultrasonic distance sensor (HC-SRO4) and Arduino UNO/MEGA. Voice synthesizer: LSI ATP3011F4-PU
Scene: A caregiver pushed a wheelchair on which a user sat and was bringing the user to a bed in a high speed. When the sensor attached to the wheelchair detected the bed at the point of 90 cm away from the bed the sensor network system sent both the user and the caregiver a warning in voice messages, such as “Stop! Stop! Pull the body up to pull legs.”
Subjects: Group I. 30 persons between 18 and 22 years old.
Group II. 6 persons between 23 and 65 years old.

Case (a). A subject was a wheelchair user. S(he) tried to decrease the wheelchair speed and prepare not to hit his/her foot with the bed (See Fig. 1, Fig. 5).
Operations (a).
(Oa1) A subject heard the warning and explanations of actions in voice guides in Figure 6.
(Oa2) A subject tried to decrease the speed of wheelchair and pull his/her feet onto the foot rests.

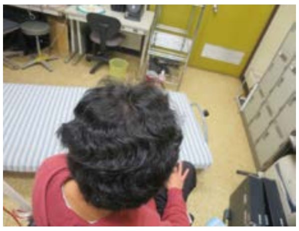
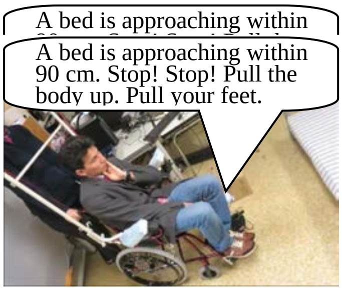


Fig. 5 A blow of foot by a bed frame in a blind spot


A bed is approaching within
A bed is approaching within 90 cm. Stop! Stop! Pull the body up. Pull your feet.
Fig. 6 Voice warning of bed approaching

Case (b). A subject was a caregiver. S(he) tried to stop the wheelchair and hold the user’s body up and pull his/her feet onto foot rests in order not to blow his/her feet with the bed.
Case (c). Remote experts were observing WU and caregivers when they would find an unpredictable hard situation.

Test.
Case (a). A wheelchair user and a caregiver acquired a warning and an instruction of avoiding damage
Question1. “Could you notice the danger by the warning ahead of time and avoid a blow of foot?”
Answer1. (See Fig. 7)
Group I: yes/no/no answer: 26/4/0
Group II: yes/no/no answer: 3/3/0

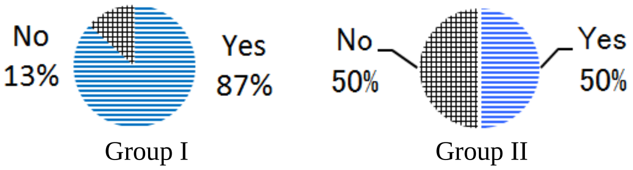


Fig. 7 Could you avoid a blow of foot with warning?

83 % young people and 50 % adults could agree with effectiveness of the acoustic warning. Many subjects worried that a WU with low physical ability could not cope with some hard situations if s(he) could get a warning of dangers.

Question2. “Could you get feeling of attainment in case of success if you would declare your intention for a goal to yourself before the start of actions?”
Answer2. (See Fig. 8)
Group I: yes/no/no answer: 26/1/3
Group II: yes/no/no answer: 5/0/1

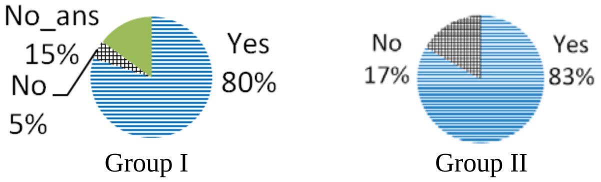


Fig. 8 Could you get feeling of attainment in case of your success with intention declaration?

80 % young people and 83 % adults could agree with effectiveness of the declaration of WU’s intention to WU in

order to avoid dangers and get feeling attainment. We could ensure that WU could carefully check dangers and take safe actions by the declaration when WU got a warning of danger from Aware Wheelchair. WU could get feeling of attainment and reduce stress.

Case (b). A caregiver acquired a warning and an instruction of avoiding damage
Question3. "Could you avoid accidents if you would declare the shared intention with WU to both WU and yourself?"
Answer3. (See Fig. 9)
Group I: yes/no/no answer: 26/4/0
Group II: yes/no/no answer: 6/0/0

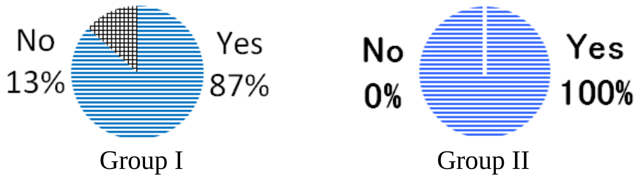


Fig. 9 Could you avoid accidents by declaration of the shared intention to both WU and yourself?

Interview3. Why could you avoid the accidents?
AnswerI3.
Group I and Group II: (a1) Both WU and caregiver could check dangers and carefully take safe actions with shared same intention.

Question4. "Could you reduce stress and easily control the wheelchair if AW could automatically reduce the speed of moving wheelchair in case of danger?"
Answer4. (See Fig. 10)
Group I: yes/no/no answer: 4/18/8
Group II: yes/no/no answer: 2/3/1

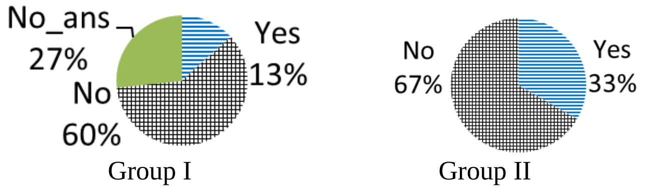


Fig. 10 Could you reduce stress and easily control the wheelchair by automatic speed reducer?

Interview4. Why could not you easily control the wheelchair?
AnswerI4.
Group I: (a2) I could not think that accidents could not be avoided only by slow seed of the wheelchair.
Group II: (a3) Automatic slow down function was useful. Machine would become danger when the machine would be broken.
(a4) Sudden stop caused stress of caregiver since caregiver autonomously controlled the wheelchair with some image of motion of the whelchair.

Case (c). Remote experts were observing WU and caregivers when they would find an unpredictable hard situation
Question5. "Could remote experts help WU and caregivers by Big-data analysis of accidents and close calls when they would encounter unpredictable hard situations?"
Answer5. (See Fig. 9)
Group I: yes/no/no answer: 27/3/0
Group II: yes/no/no answer: 3/3/0
Interview5. Why could remote experts help such unpredictable situation?
AnswerI5.

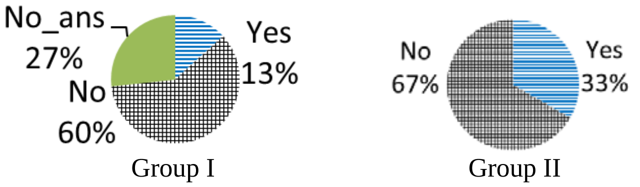


Fig. 11 Could remote experts help WU and caregivers by Big-data analysis of accidents and close calls?

Group I: (a5) Big-data analysis of accidents and close calls could be useful for unpredictable situations.
Group II: (a6) Only real-time information supply would be effective. Remote experts would be useful for training of trainees.

■Discussion
D1. AW could detect danger in a blind spot by sensor networks and distribute a warning to WU and caregiver. Warnings for blind spots were useful for avoiding accidents if WU or a caregiver would have enough physical abilities.

D2. AW's warning naturally caused the declaration by a WU that enabled to ensure intention and start reminding important check points. S(he) could autonomously proceed checking dangers and safe actions with his/her intention. S(he) could get feeling of attainment when s(he) could complete safe actions. The WU could trust himself/herself a nd reduce stress.

D3. Sharing intention enabled both WU and caregivers to easily collaborate with images of actions of surrounding people. Sharing intention would also avoid serious accidents.

D4. AW's warning naturally caused the declaration by a caregiver that enabled both WU and caregivers to share intention. They could start reminding important check points and checking dangers at the same time. They could also timely collaborate with safe actions in smaller load. Both WU and caregivers could trust each other and reduce stress.

D5. Remote experts would be useful to avoid dangers and cope with them in a new type of dangerous situation. Real-time distribution of useful information for a weak WU in the situation would be very important.

5 CONCLUSION

We investigate safe collaboration methodologies by "Aware Wheelchair (AW)" with sensor networks that detect dangers and distribute awareness and suggest suitable actions to both user and caregivers. We also discuss sharing intention for easier and stress-less collaboration between a wheelchair user (WU) and caregivers. Declaration of intentions that were triggered by awareness would be effective to avoid dangers and get feeling of attainment. Intelligent function should enhance the collaboration between WU, caregivers, and remote experts.

REFERENCES

[1] S. Timoshenko, and S. Woinowsky-Krieger, Theory of Plates and Shells, McGraw-Hill, pp.415-428 (1959).

[1] WILL, Inc. Features, "https://whill.jp/;" (2015).

[2] Yamaha Motor Co., Ltd., JW Active Plus+, "http://www.yamaha-motor.co.jp/wheelchair/lineup/#u-joy;" (2015).

[3] YDS design systems, WheelChair Vehicle (WCV), "http://www.yds-wcv.jp/index.html;" (2016).

[4] N. Wattanavarangkul, and T. Wakahara, Indoor and outdoor navigation system for wheelchair users, IPSJ SIG Technical Report, pp.1-6 (2012).

[5] H. Jogasaki, S. Mori, Y. Nakamura, O. Takahashi, Evaluation for the methodology of designing surface-code from the point of comfort on the road to create in-house navigation system to safely driving of wheelchairs, IPSJ SIG Technical Report 2015, CDS-12, No.12, pp.1-8 (2015).

[6] Nihon Wheel Chairs Co.Ltd., Case Study Welfare Equipment, "http://nwc-kurumaisu.com/fukushi-hiyari-hato/;" (2012).

[7] Miyachi, T., Maezawa, T., Nishihara, T., Suzuki, T., Affordance in Dynamic Objects Based on Face Recognition, Springer-Verlag, LNCS6277, pp.645-652, (2010).

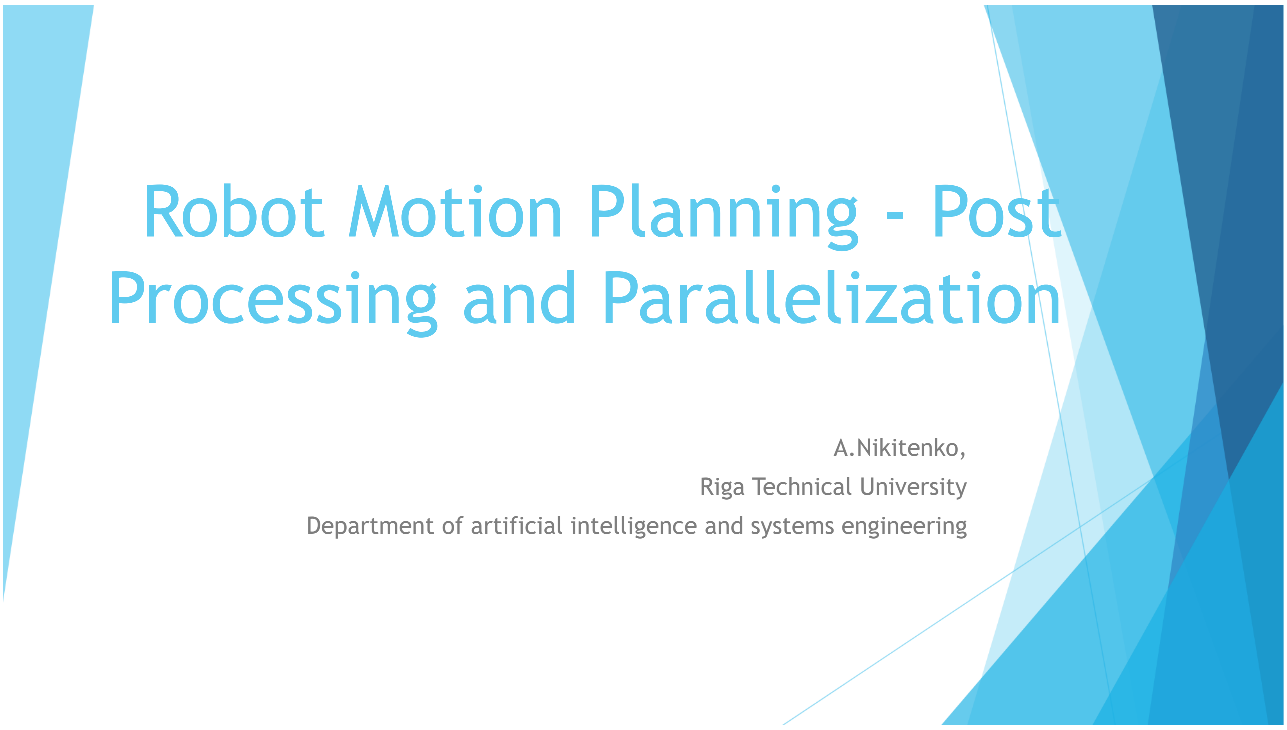
[8] Hemachandra, S., Walter, M.R., Tellex, S., and Teller, S. Learning Spatial-Semantic Representations from Natural Language Descriptions and Scene Classifications, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, (PDF), May (2014).

[9] Buribayeva, B., Miyachi, T., Yeshmukhametov, A., Mikami, Y., An Autonomous Emergency Warning System Based on Cloud Servers and SNS, ELSEVIER, Procedia Computer Science 2015, vol.60, pp.722 – 729 (2015).

Keynote Speech 2:

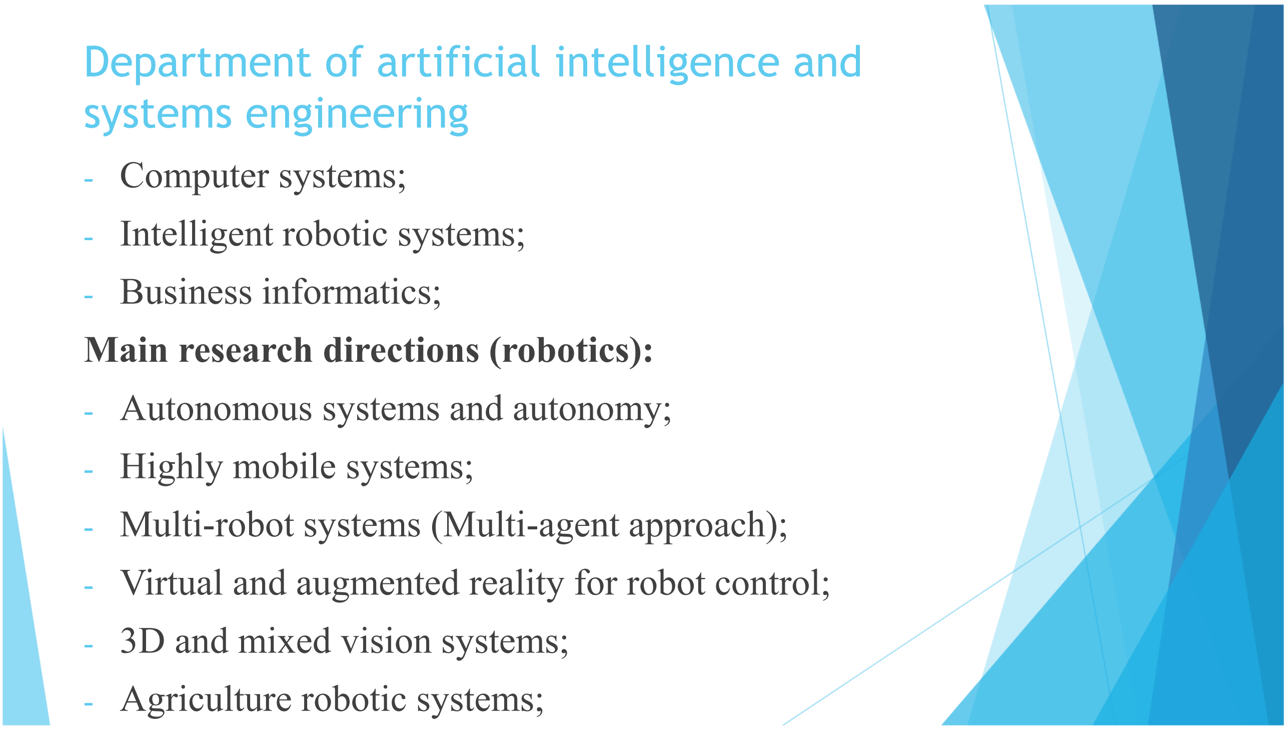
Prof. Agris Nikitenko

(Riga Technical University)



Robot Motion Planning - Post Processing and Parallelization

A.Nikitenko,
Riga Technical University
Department of artificial intelligence and systems engineering

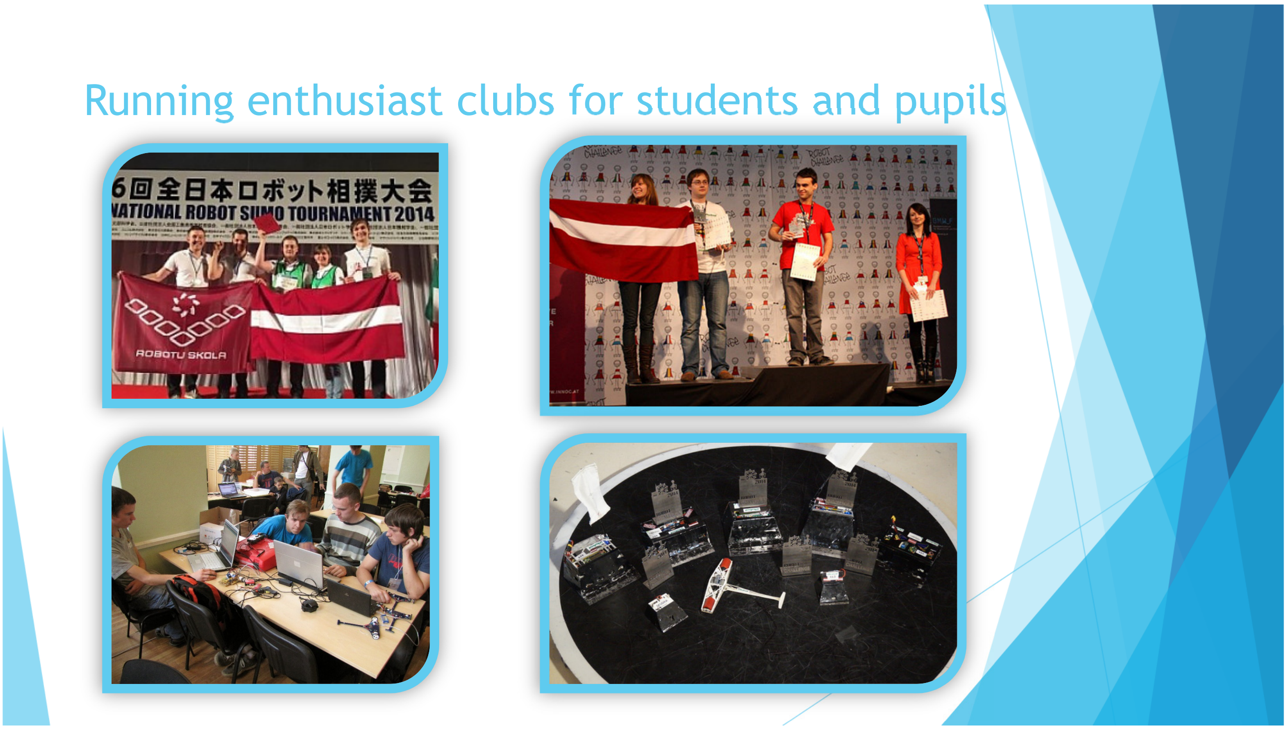


Department of artificial intelligence and systems engineering

- Computer systems;
- Intelligent robotic systems;
- Business informatics;

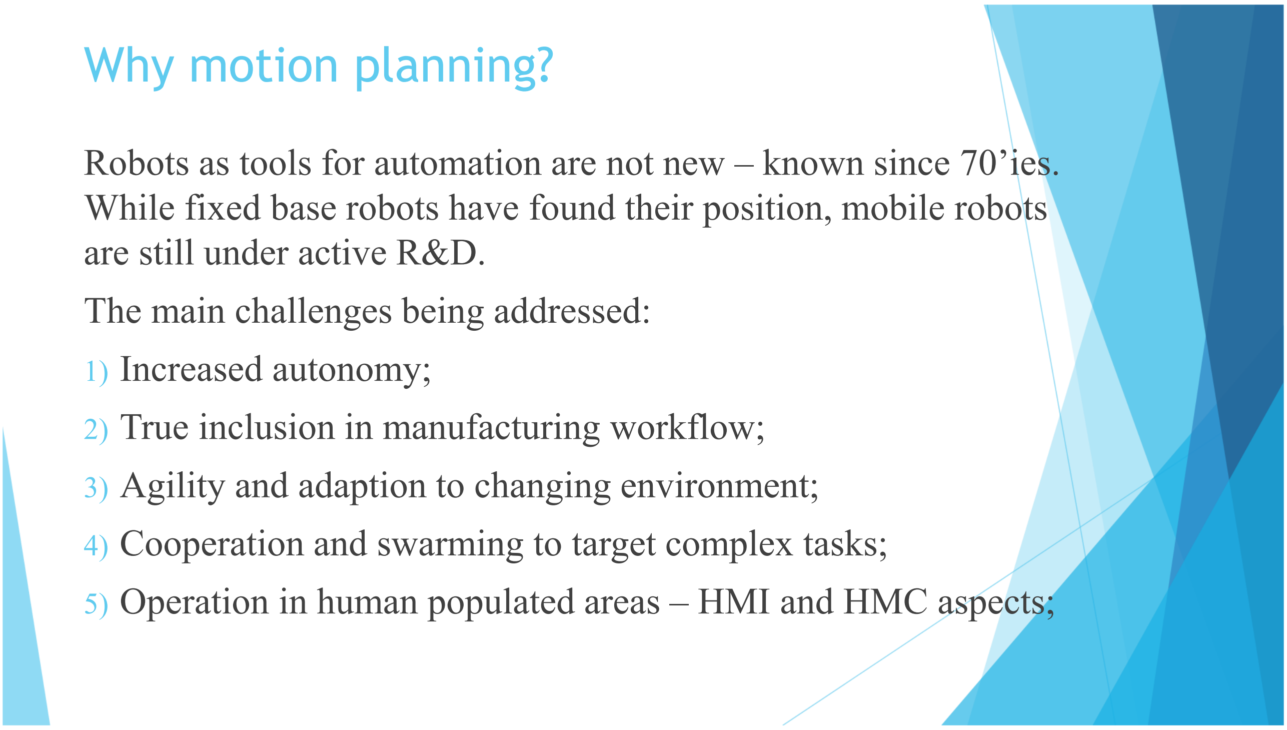
Main research directions (robotics):

- Autonomous systems and autonomy;
- Highly mobile systems;
- Multi-robot systems (Multi-agent approach);
- Virtual and augmented reality for robot control;
- 3D and mixed vision systems;
- Agriculture robotic systems;



Running enthusiast clubs for students and pupils

6回全日本ロボット相撲大会
NATIONAL ROBOT SUMO TOURNAMENT 2014
ROBOTU SKOLA



Why motion planning?

Robots as tools for automation are not new – known since 70'ies. While fixed base robots have found their position, mobile robots are still under active R&D.

The main challenges being addressed:

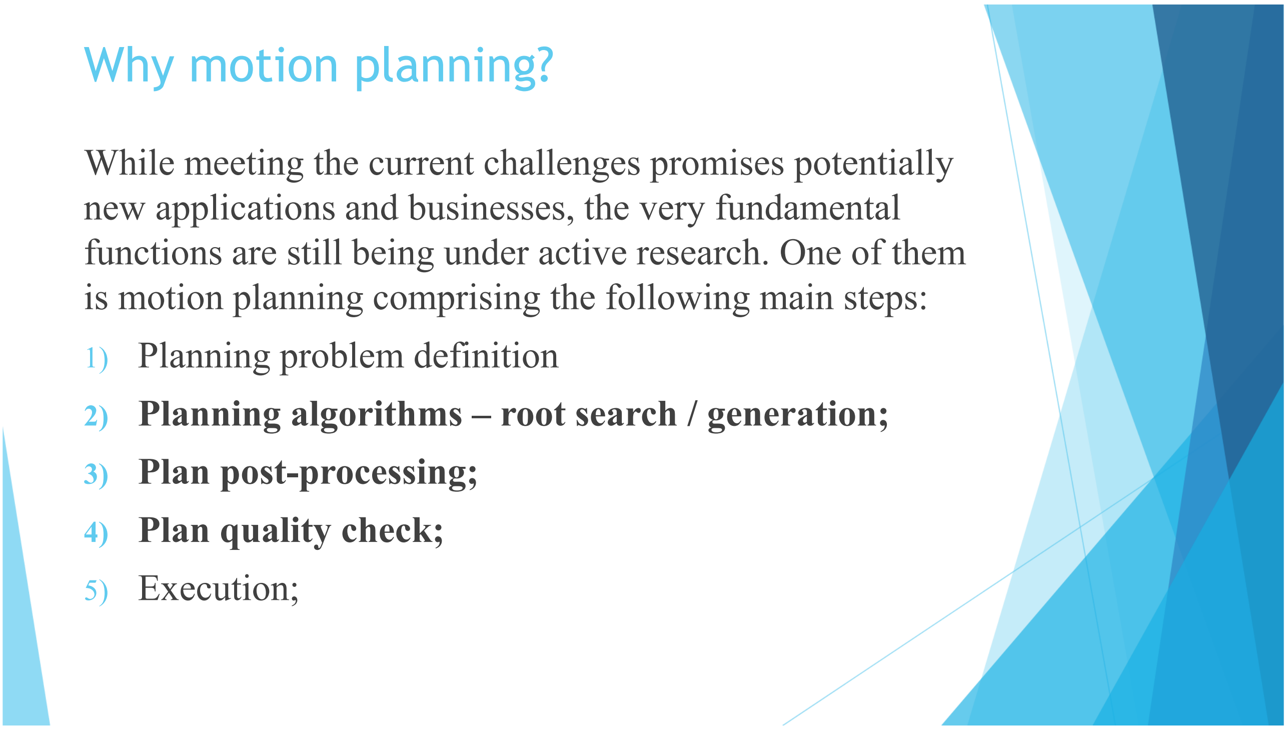
1) Increased autonomy;

2) True inclusion in manufacturing workflow;

3) Agility and adaption to changing environment;

4) Cooperation and swarming to target complex tasks;

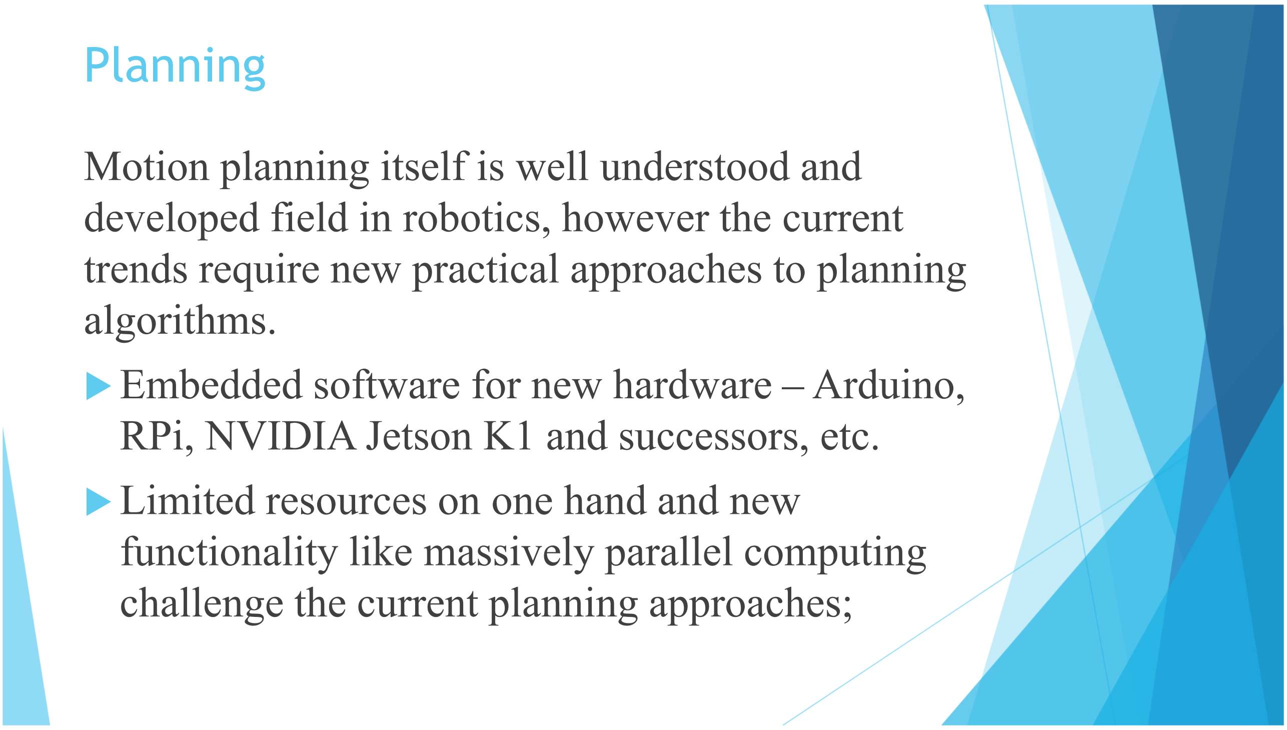
5) Operation in human populated areas – HMI and HMC aspects;



Why motion planning?

While meeting the current challenges promises potentially new applications and businesses, the very fundamental functions are still being under active research. One of them is motion planning comprising the following main steps:

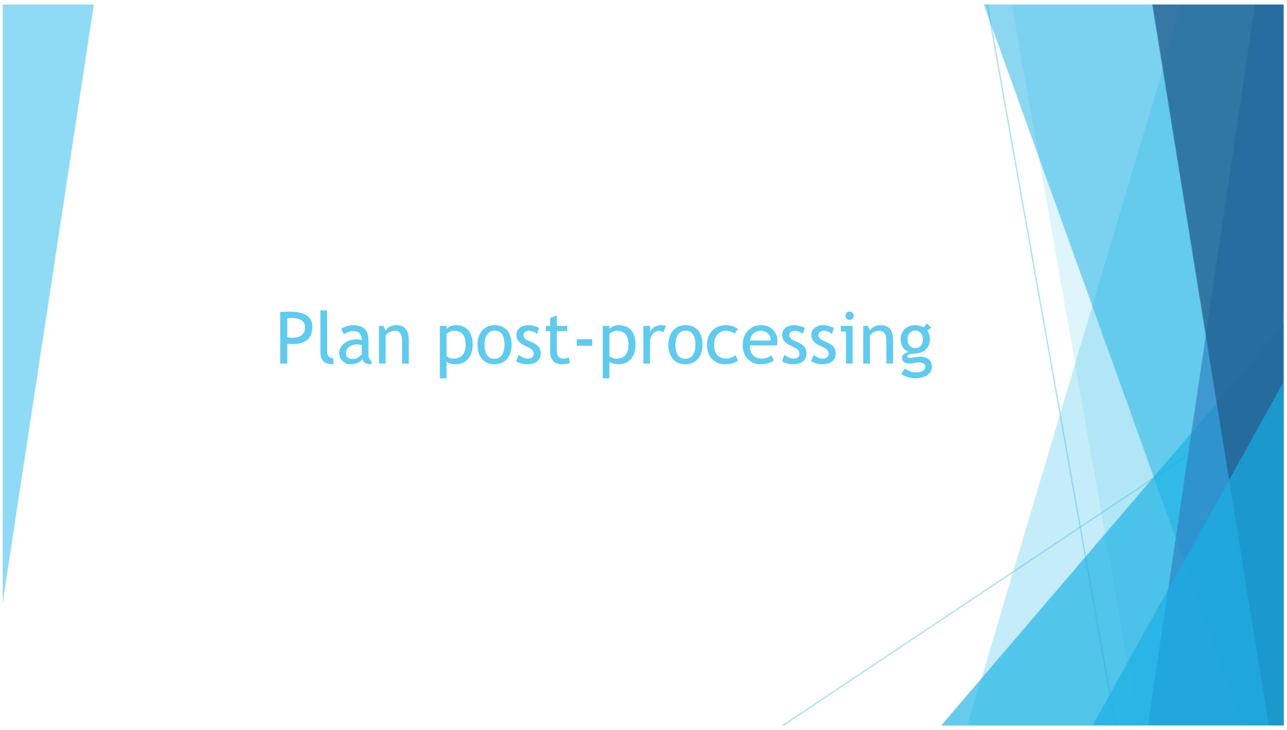
1) Planning problem definition
2) **Planning algorithms – root search / generation;**
3) **Plan post-processing;**
4) **Plan quality check;**
5) Execution;



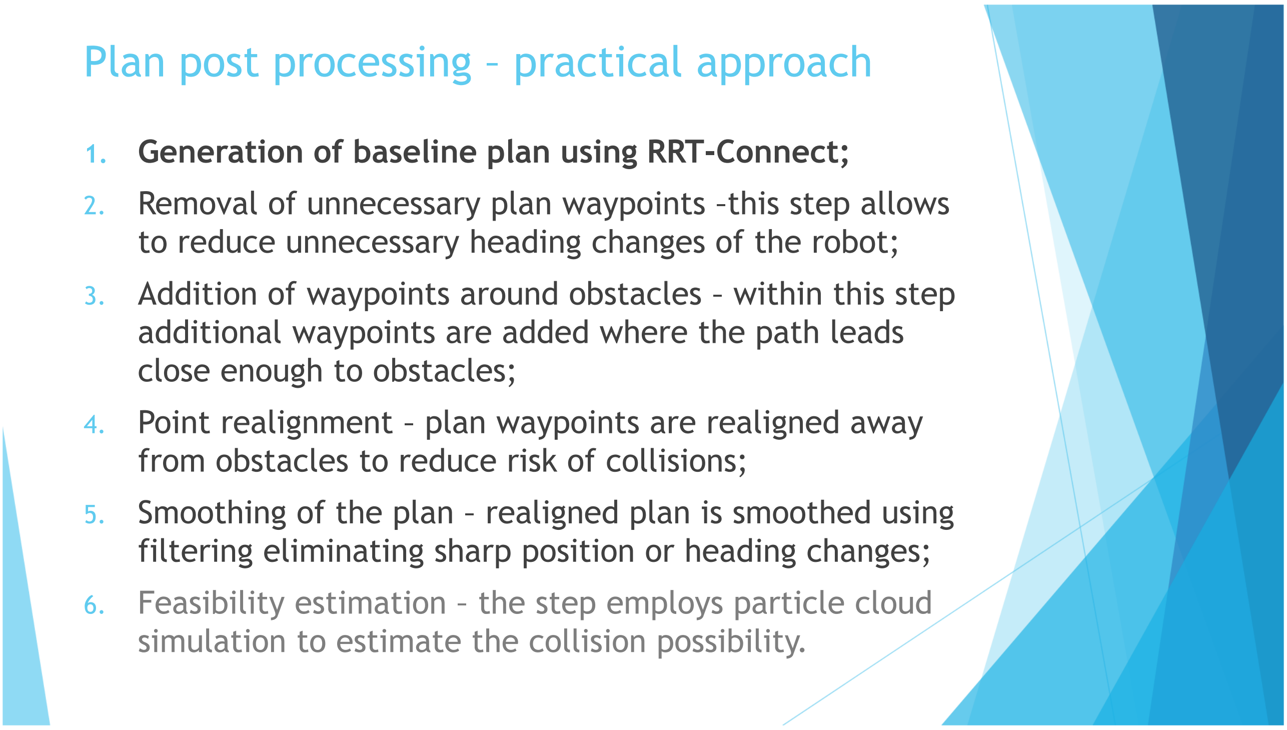
Planning

Motion planning itself is well understood and developed field in robotics, however the current trends require new practical approaches to planning algorithms.

- Embedded software for new hardware – Arduino, RPi, NVIDIA Jetson K1 and successors, etc.
- Limited resources on one hand and new functionality like massively parallel computing challenge the current planning approaches;



Plan post-processing



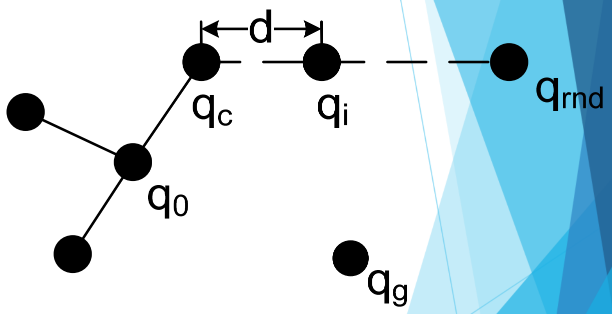
Plan post processing – practical approach

1. **Generation of baseline plan using RRT-Connect;**
2. Removal of unnecessary plan waypoints –this step allows to reduce unnecessary heading changes of the robot;
3. Addition of waypoints around obstacles – within this step additional waypoints are added where the path leads close enough to obstacles;
4. Point realignment – plan waypoints are realigned away from obstacles to reduce risk of collisions;
5. Smoothing of the plan – realigned plan is smoothed using filtering eliminating sharp position or heading changes;
6. Feasibility estimation – the step employs particle cloud simulation to estimate the collision possibility.

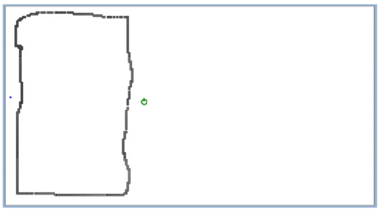
RRT-Connect

Algorithm: Base RRT

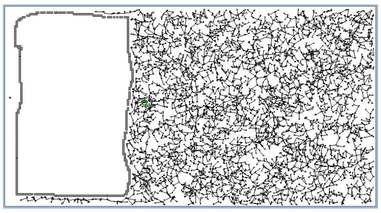
```
RRTmain()
Tree = q.0
q.rnd = q.0
while Distance (q.rnd , q.g) < ErrTolerance do
    q.target = SampleTarget()
    q.nearest = NearestVertex (Tree , q.target)
    q.rnd = ExtendTowards (q.nearest,q.target)
    Tree.add(q.rnd)
end while
return Trajectory (Tree,q.rnd)

SampleTarget ()
if Rand() < GoalSamplingProb then
    return q.g
else
    return RandomConfiguration()
end if
```



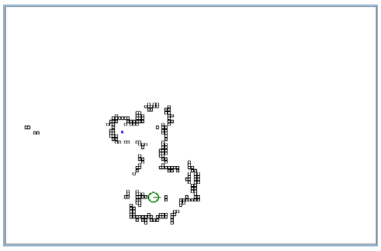
Problems in practical situations



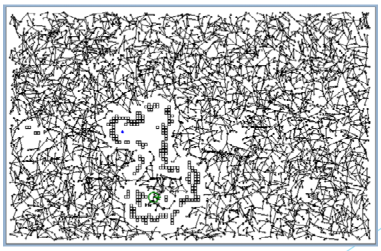
A - planning area 24m x 13m with a single large obstacle



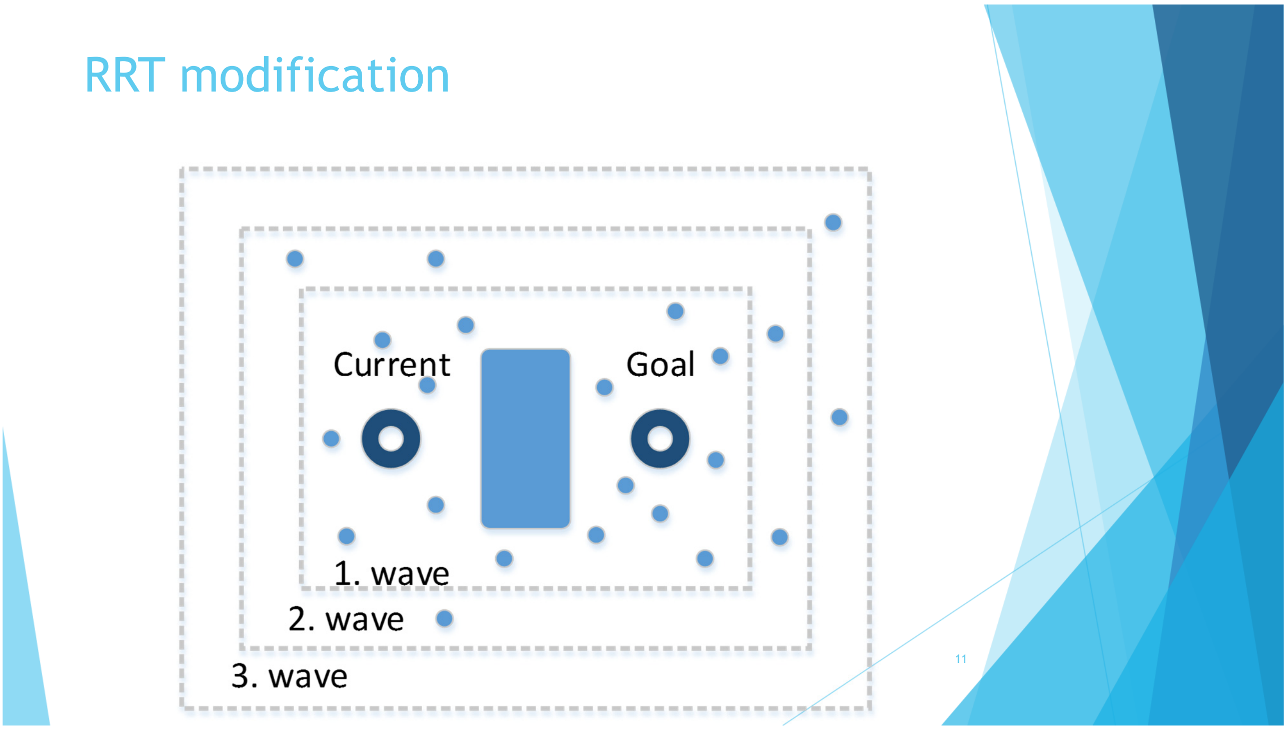
A - planning area 24m x 13m with a single large obstacle planning result - failure.



B - Planning area in small warehouse



B - Planning area in small warehouse with planning result - failure.



RRT modification

Current

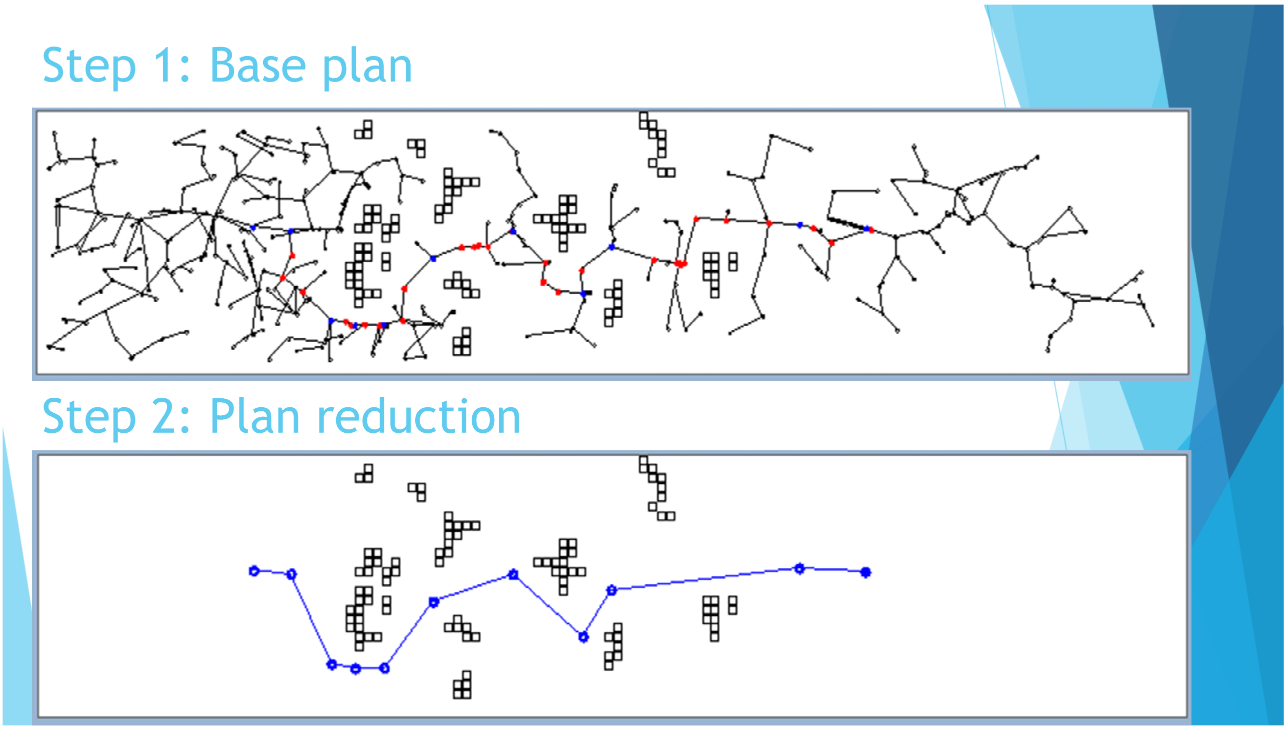
Goal

1. wave

2. wave

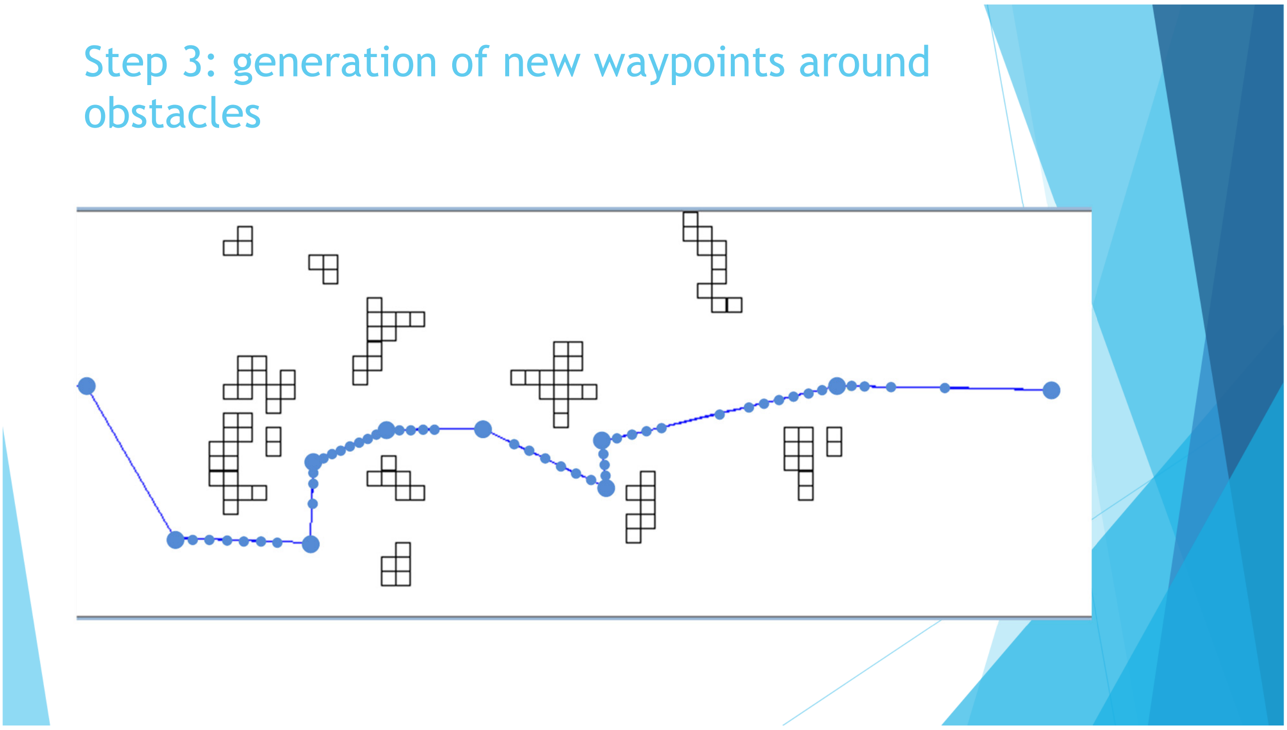
3. wave

11

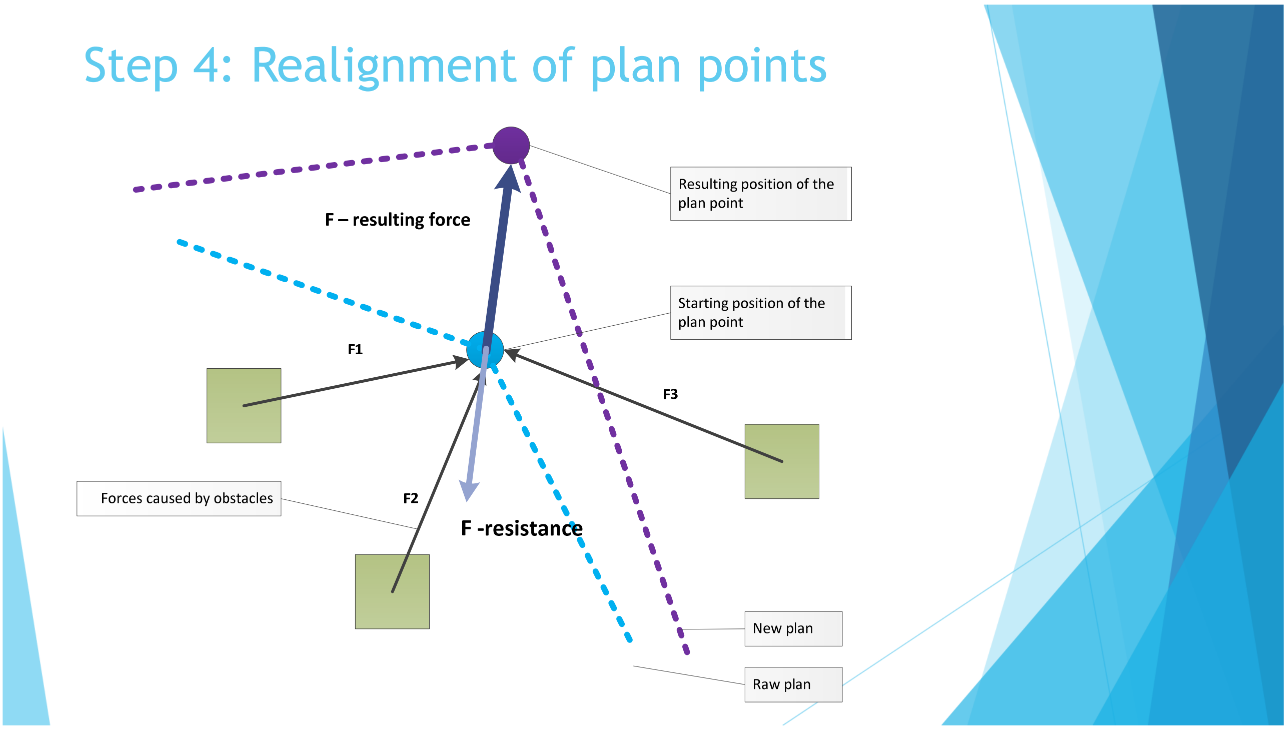


Step 1: Base plan

Step 2: Plan reduction



Step 3: generation of new waypoints around obstacles



Step 4: Realignment of plan points

F – resulting force

Resulting position of the plan point

Starting position of the plan point

F1

F2

F3

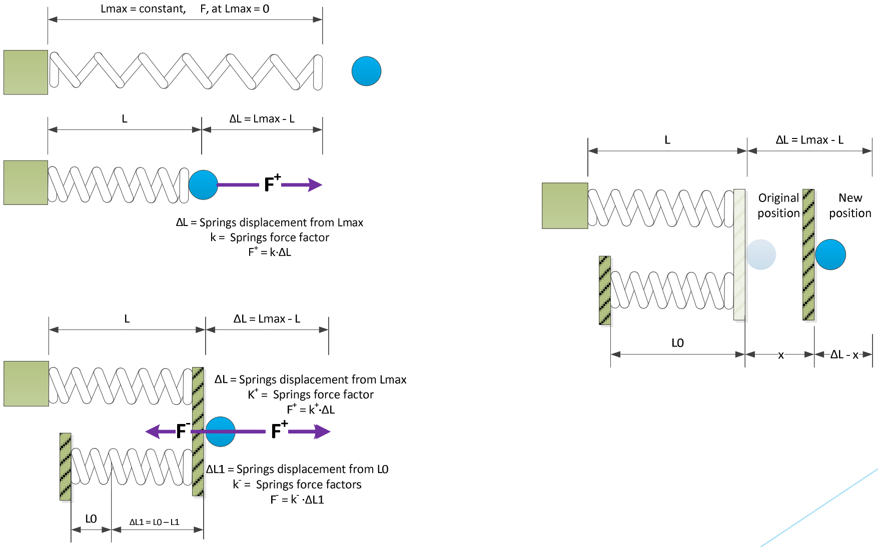
Forces caused by obstacles

F -resistance

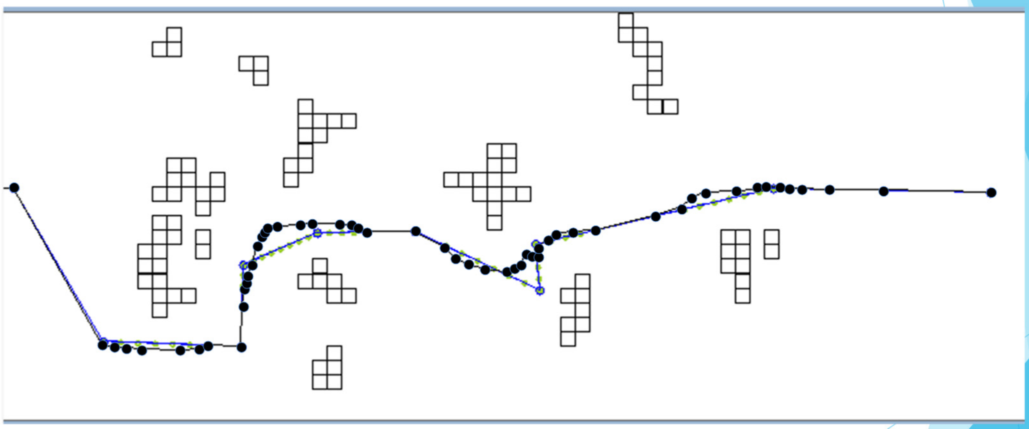
New plan

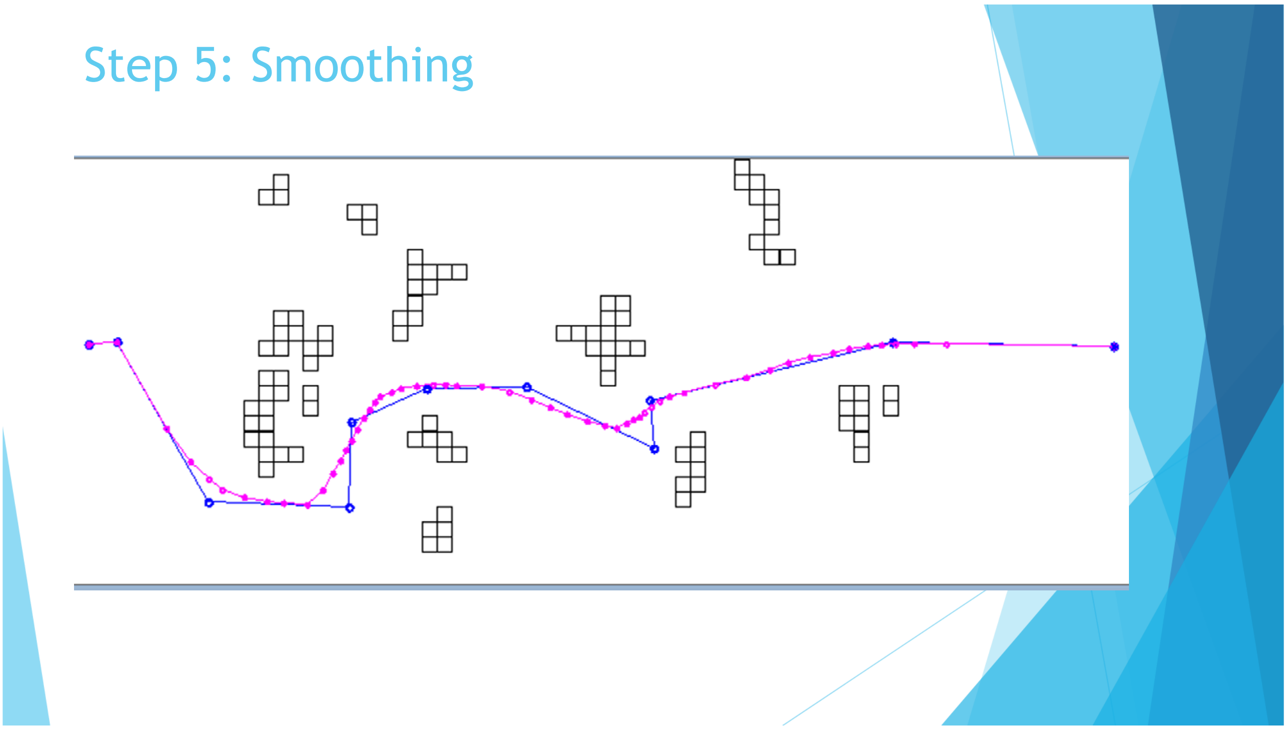
Raw plan

Step 4: Realignment of plan points

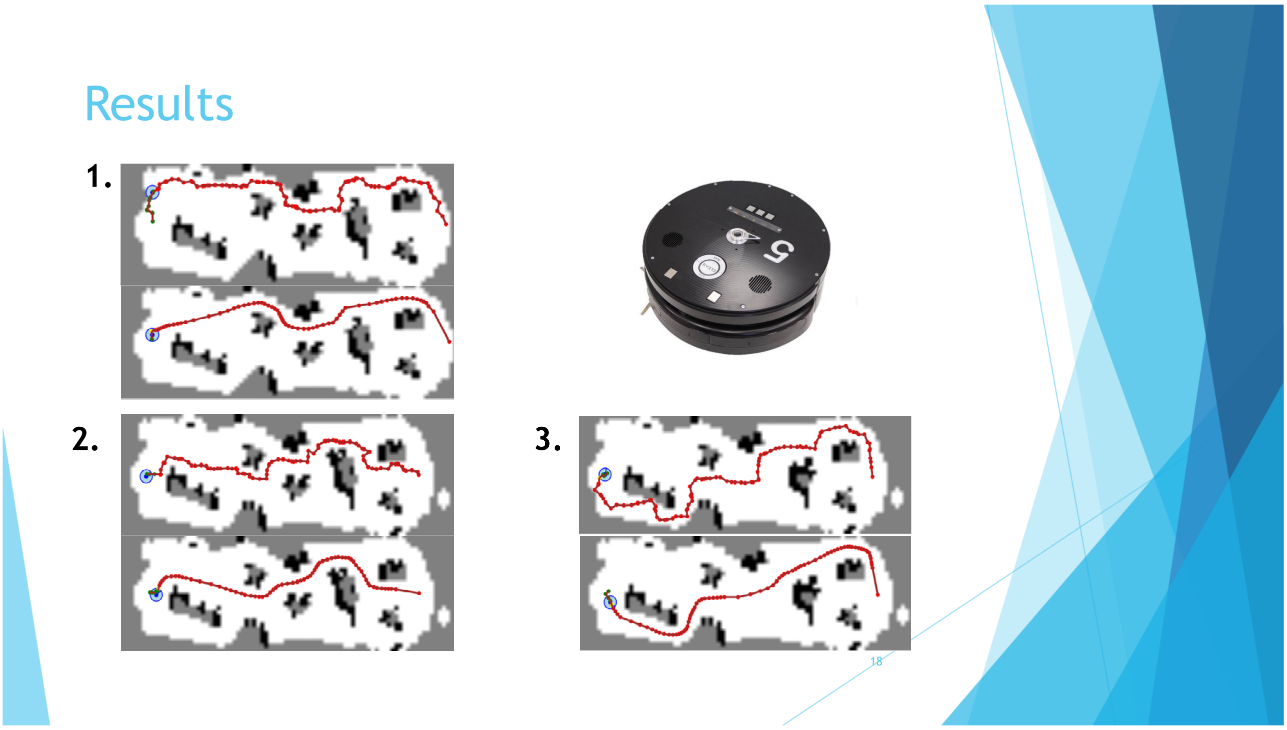

Lmax = constant, F, at Lmax = 0
L
ΔL = Lmax - L
F+
ΔL = Springs displacement from Lmax
k = Springs force factor
F+ = k·ΔL
L
ΔL = Lmax - L
ΔL = Springs displacement from Lmax
K+ = Springs force factor
F+ = k+·ΔL
F-
F+
ΔL1 = Springs displacement from L0
k- = Springs force factors
F- = k-·ΔL1
L0
ΔL1 = L0 – L1
L
ΔL = Lmax - L
Original position
New position
L0
x
ΔL - x

Step 4: Realignment of plan points

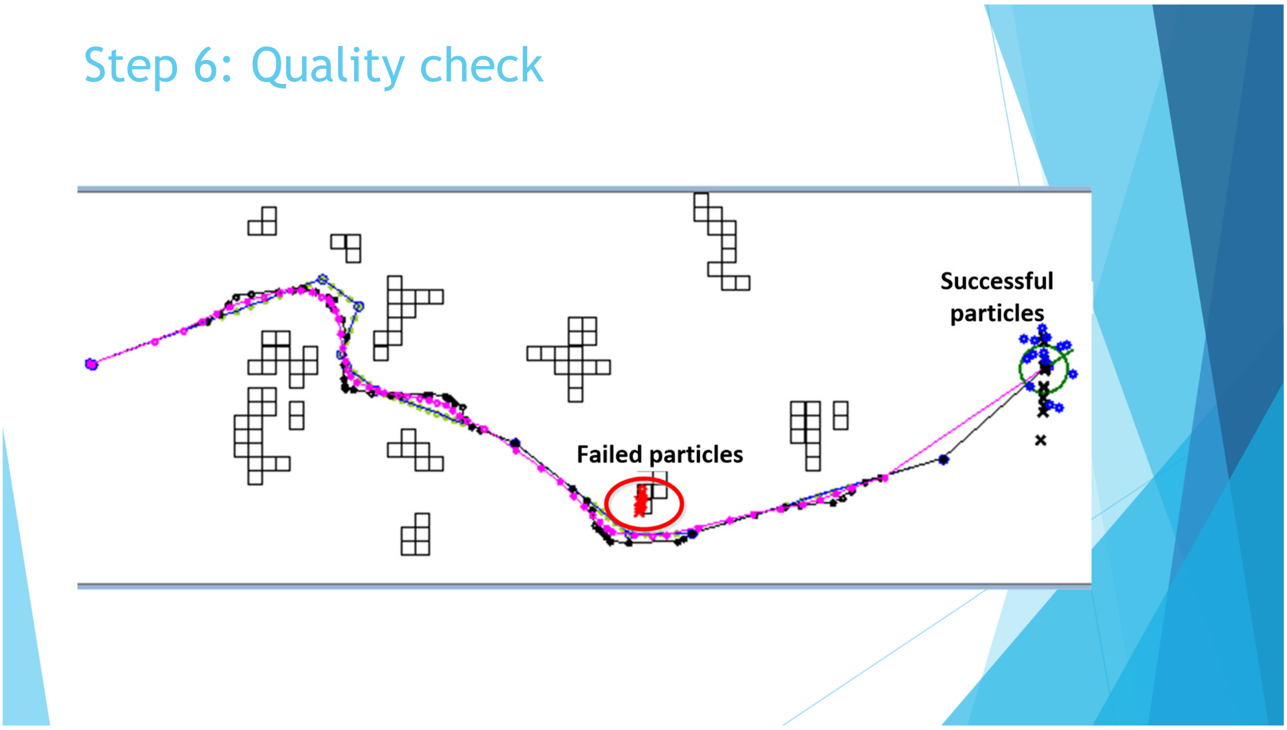




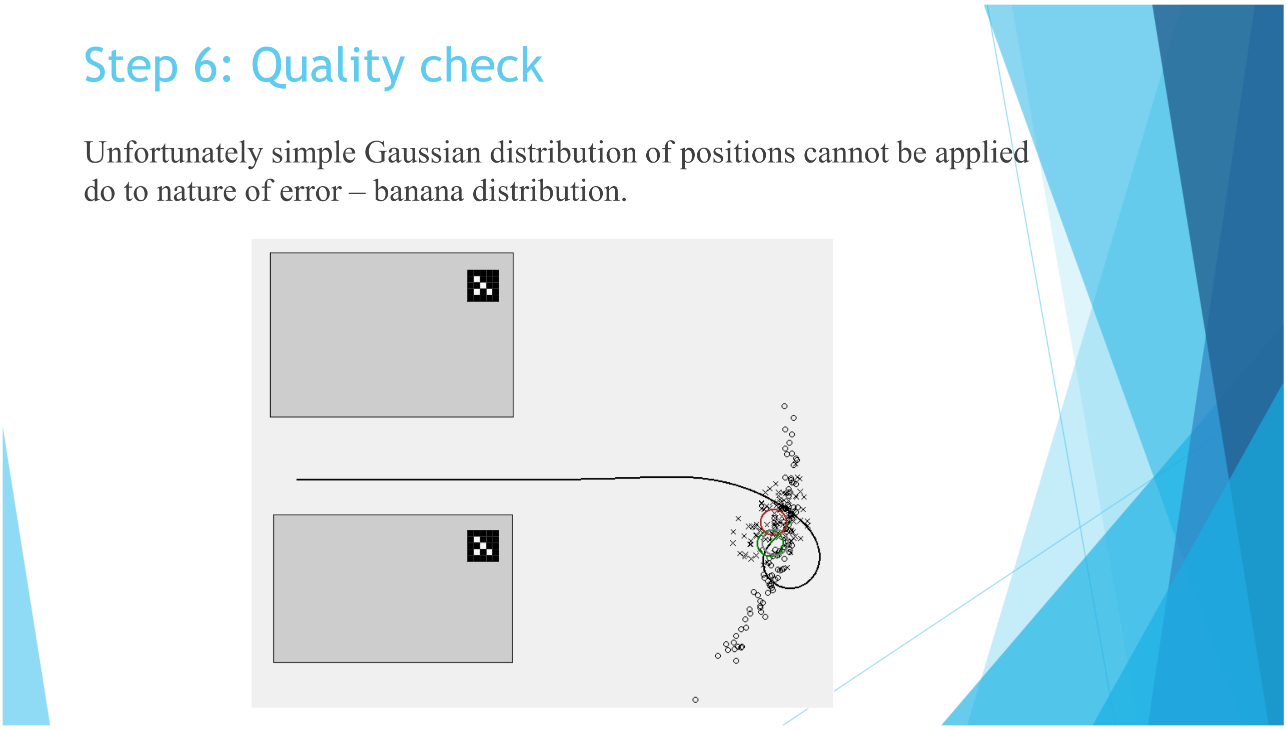
Step 5: Smoothing



Results

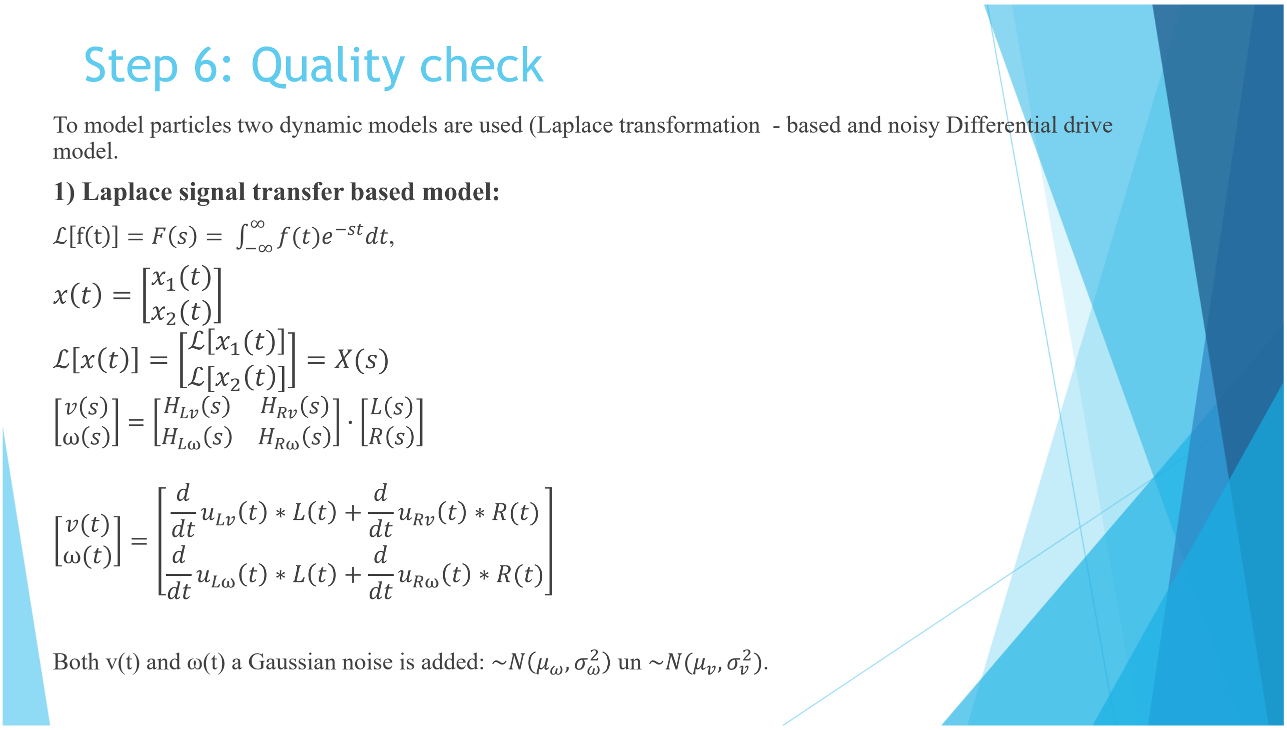


Step 6: Quality check



Step 6: Quality check

Unfortunately simple Gaussian distribution of positions cannot be applied do to nature of error – banana distribution.



Step 6: Quality check

To model particles two dynamic models are used (Laplace transformation - based and noisy Differential drive model.

1) Laplace signal transfer based model:

$$\mathcal{L}[\mathrm{f(t)}] = F(s) = \int_{-\infty}^{\infty} f(t)e^{-st}dt,$$

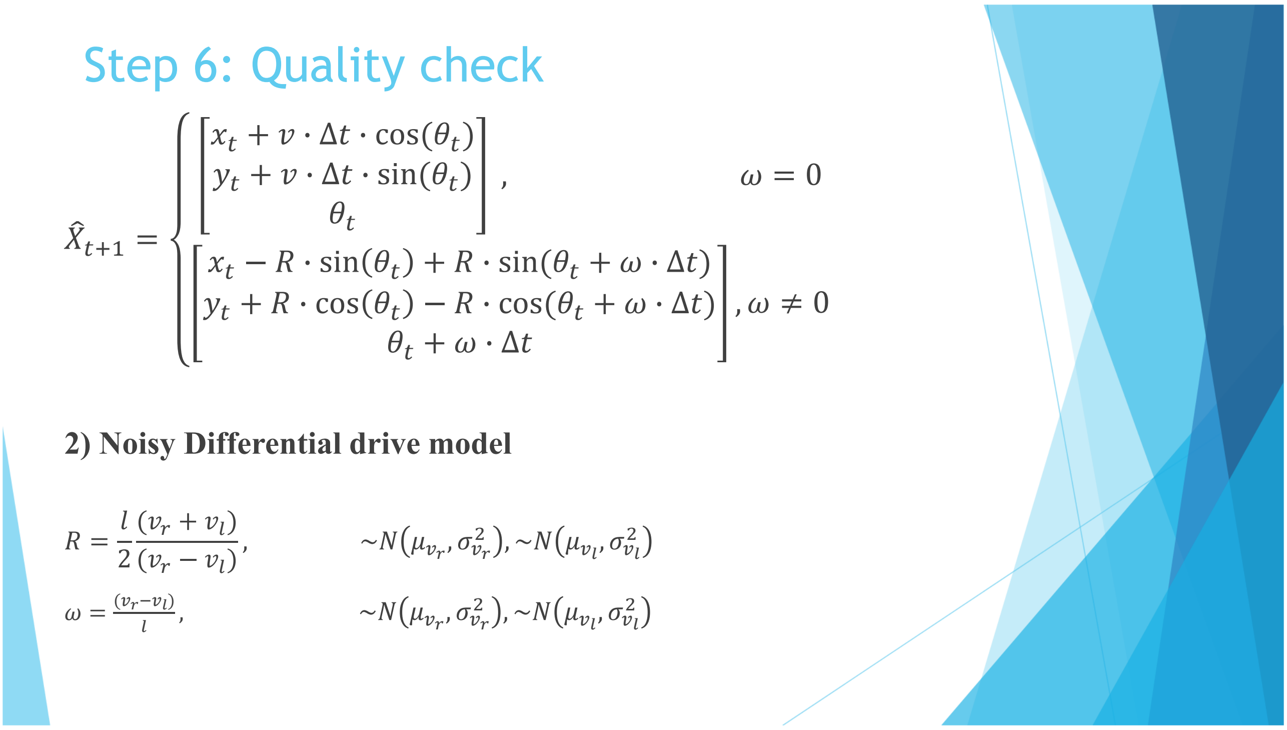
$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

$$\mathcal{L}[x(t)] = \begin{bmatrix} \mathcal{L}[x_1(t)] \\ \mathcal{L}[x_2(t)] \end{bmatrix} = X(s)$$

$$\begin{bmatrix} v(s) \\ \omega(s) \end{bmatrix} = \begin{bmatrix} H_{Lv}(s) & H_{Rv}(s) \\ H_{L\omega}(s) & H_{R\omega}(s) \end{bmatrix} \cdot \begin{bmatrix} L(s) \\ R(s) \end{bmatrix}$$

$$\begin{bmatrix} v(t) \\ \omega(t) \end{bmatrix} = \begin{bmatrix} \frac{d}{dt} u_{Lv}(t) * L(t) + \frac{d}{dt} u_{Rv}(t) * R(t) \\ \frac{d}{dt} u_{L\omega}(t) * L(t) + \frac{d}{dt} u_{R\omega}(t) * R(t) \end{bmatrix}$$

Both v(t) and ω(t) a Gaussian noise is added: $\sim N(\mu_\omega, \sigma_\omega^2)$ un $\sim N(\mu_v, \sigma_v^2)$.



Step 6: Quality check

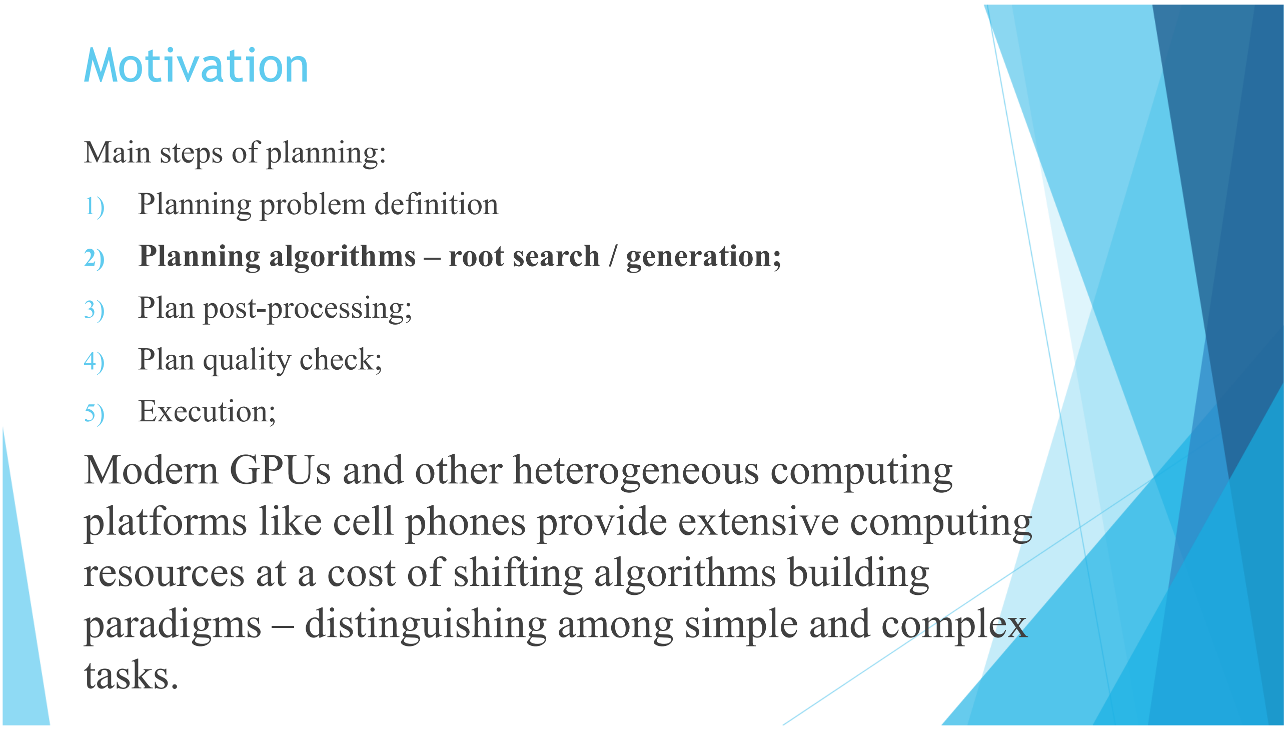
$$\hat{X}_{t+1} = \begin{cases} \begin{bmatrix} x_t + v \cdot \Delta t \cdot \cos(\theta_t) \\ y_t + v \cdot \Delta t \cdot \sin(\theta_t) \\ \theta_t \end{bmatrix}, & \omega = 0 \\ \begin{bmatrix} x_t - R \cdot \sin(\theta_t) + R \cdot \sin(\theta_t + \omega \cdot \Delta t) \\ y_t + R \cdot \cos(\theta_t) - R \cdot \cos(\theta_t + \omega \cdot \Delta t) \\ \theta_t + \omega \cdot \Delta t \end{bmatrix}, & \omega \neq 0 \end{cases}$$

2) Noisy Differential drive model

$$R = \frac{l}{2} \frac{(v_r + v_l)}{(v_r - v_l)}, \qquad \sim N\left(\mu_{v_r}, \sigma_{v_r}^2\right), \sim N\left(\mu_{v_l}, \sigma_{v_l}^2\right)$$

$$\omega = \frac{(v_r - v_l)}{l}, \qquad \sim N\left(\mu_{v_r}, \sigma_{v_r}^2\right), \sim N\left(\mu_{v_l}, \sigma_{v_l}^2\right)$$

Parallelization



Motivation

Main steps of planning:

1) Planning problem definition
2) **Planning algorithms – root search / generation;**
3) Plan post-processing;
4) Plan quality check;
5) Execution;

Modern GPUs and other heterogeneous computing platforms like cell phones provide extensive computing resources at a cost of shifting algorithms building paradigms – distinguishing among simple and complex tasks.

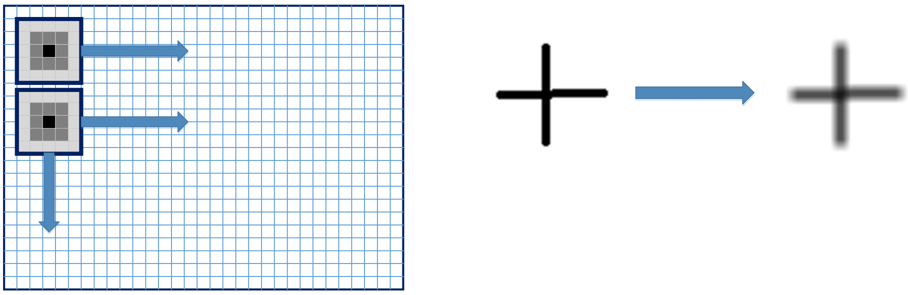
Main steps

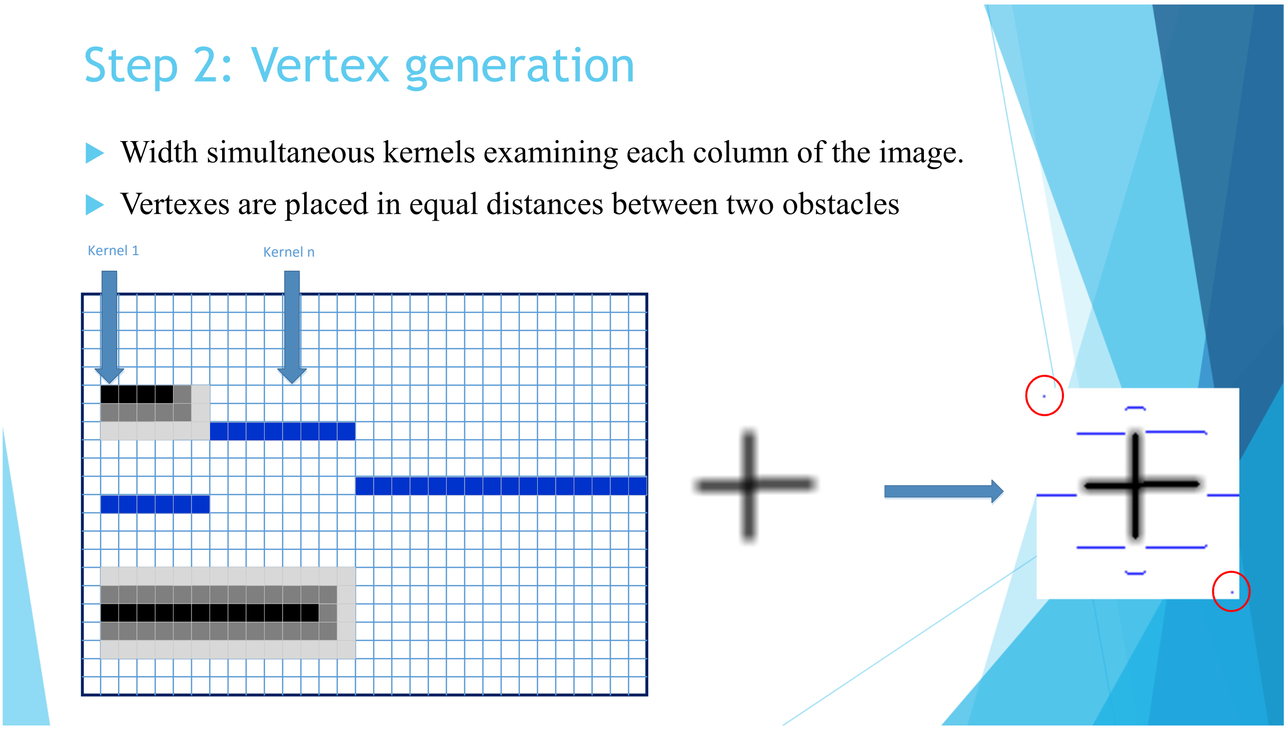
Transforming the planning problem to image processing problem allows to apply massive parallelization.

1) **Bluring the obstacle image ensuring safe distances to obstacles;**
2) **Generation of planning graph vertexes;**
3) **Building and planning graph;**
4) Searching the graph;
5) Post-processing;

Step 1: Bluring

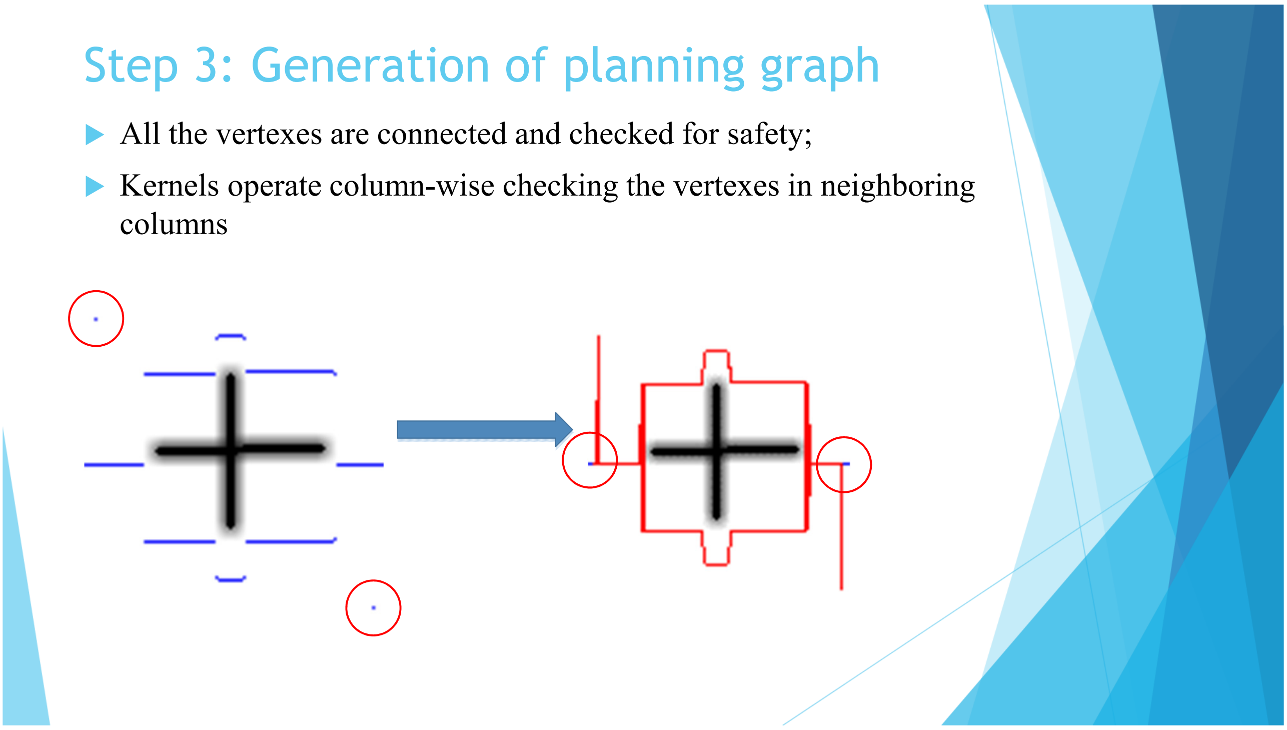
Width x Height simultaneous kernels





Step 2: Vertex generation

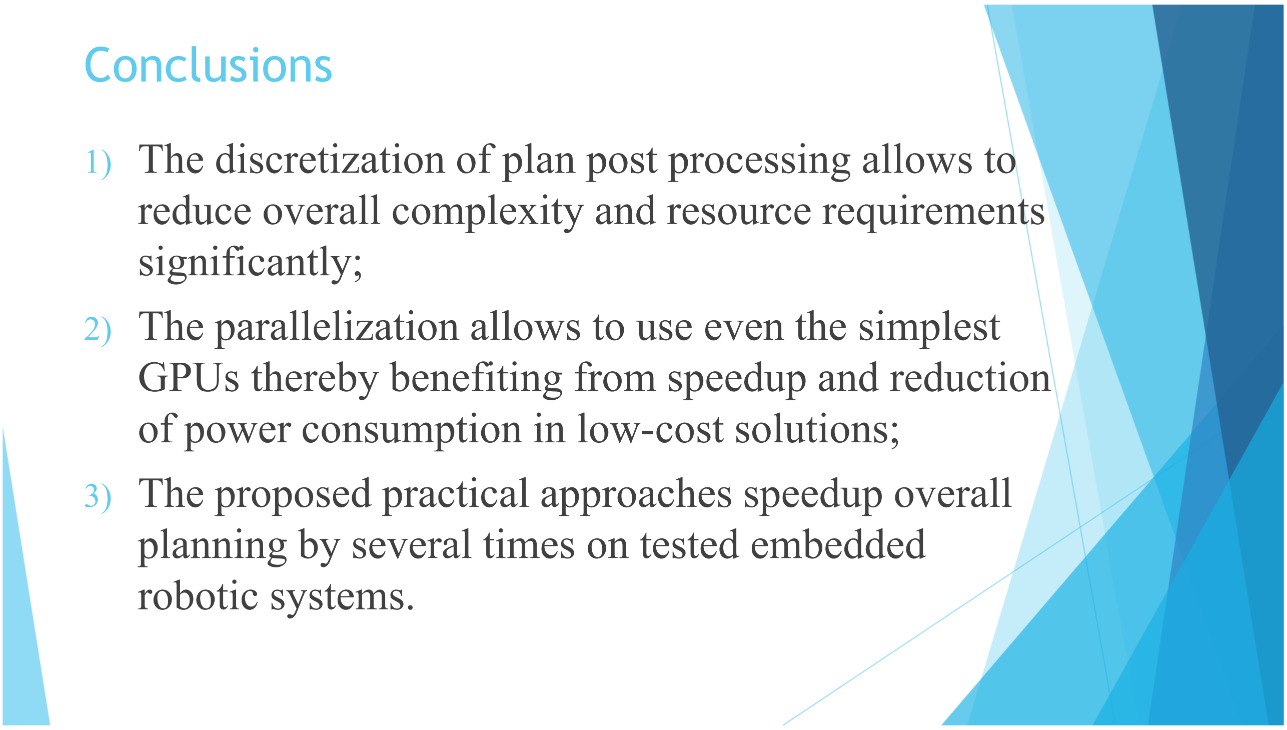
- Width simultaneous kernels examining each column of the image.
- Vertexes are placed in equal distances between two obstacles



Step 3: Generation of planning graph

- All the vertexes are connected and checked for safety;
- Kernels operate column-wise checking the vertexes in neighboring columns

Applications



Conclusions

1) The discretization of plan post processing allows to reduce overall complexity and resource requirements significantly;
2) The parallelization allows to use even the simplest GPUs thereby benefiting from speedup and reduction of power consumption in low-cost solutions;
3) The proposed practical approaches speedup overall planning by several times on tested embedded robotic systems.

Thank you!
Questions?

Session 3:
Intelligent Transportattion Systems

(Chair: Yoshitaka Nakamura)

Development of a Near-Miss Map System utilizing Driver's Emotions

Yoshia Saito[*]

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

Abstract – To realize a safe car society, it is necessary to support drivers not only in mechanical aspects but also human aspects. We propose a system for automatically generating a near-miss map utilizing driver's emotion. Existing near-miss map systems only use conditions of the car such as heavy braking to detect unsafe locations. However, the accuracy of the near-miss map is not so high because there are many false negative and false positive errors. To decrease the false positive errors and false negative errors for generating a high-accuracy near-miss map, we introduce human emotions into the detection of unsafe locations and realize a high-accuracy near-miss map system. In this paper, we report the design and implementation of our prototype system and evaluation results of the prototype system in terms of false negative and false positive errors.

Keywords: Emotion, Near-Miss Map

1. INTRODUCTION

There are 77 million cars in Japan [1] and 1.1 billion cars in the world [2]. The cars are useful and essential for people to live a daily life. On the other hand, there were 540 thousand traffic accidents in Japan. 4100 people died as the victim of the traffic accidents and 670 people were injured [3]. The cars also pose a significant danger to people and we must decrease the risk. As mechanical approaches, recent cars equip driving assistance technologies such as ESC (Electronic Stability Control) and ABS (Antilock Brake System). In addition, more advance technologies which include pre-crash safety are coming into practical use in recent years such as ADAS (Advanced Driving Assistant System) and DSSS (Driving Safety Support Systems). With assistance of these technologies, the traffic accidents decreased from 950 thousand in 2003, when there were the highest number of the accidents, to 540 thousand in 2015. The decrease ratio, however, draws a shallow slope in recent years and it needs other solutions than just the mechanical ones.

It is said that a reason caused majority of the traffic accidents is human error. To decrease the traffic accidents, we need to focus on human and support drivers introducing more human approaches. As the human approaches, it is recognized that the driver's emotion affect the driving performance. The typical example is “Road Rage” [4]. The road rage is a term used to denote aggressive behaviors by drivers when they are cut into their line, overtaken by the others and got angry by other reasons. It causes fatal and injury accidents. In this way, the human emotions are important for the driving safety. Many researchers study influence of the driver's emotion and apply the driver's emotion to the driving support system [11-16].

In this research, we develop a system which creates a high-accurate near-miss map utilizing the driver's emotions. The near-miss map is a map which shows unsafe locations gathering information. The information gathering of the unsafe location in the existed system is realized by 2 different ways. The first is a manually operated way which gathers from human (e.g. questionnaire survey) the second is an automatically operated way which gathers from sensors (e.g. heavy braking detection). The manually operated way requires a great deal of time and the automatically operated way has an accuracy problem of unsafe location detection. We introduce driver's emotion to the automatically operated way in order to improve the accuracy of unsafe location detection. In our proposed system, a smartphone is used to recognize driver's surprise and fear emotions by facial expression and detect an unsafe location using the driver's emotions in conjunction with sensed car conditions. The information of unsafe location is sent to the server via the Internet. The other drivers which come by near the location can receive warning to be aware of the risk of traffic accidents.

The paper is organized as follows. In the next section, we describe human emotion researched in psychology and applied studies of the human emotions in the automobile field. Section 3 shows existed near-miss map system and their problems as a preliminary study. We propose a near-miss map system utilizing driver's emotion in section 4 and implement a prototype system in section 5. Section 6 evaluates the prototype system and reports the experiment results. Section 7 gives some conclusions and our future work.

2. RELATED WORK

As related work, we explain human emotions at first. After that, we describe existed researches which utilize human emotions in automobile field and the effectiveness to apply human emotions to driving assistance systems.

2.1. Human Emotions

Human emotions have been researched in the field of psychology. Ekman defined six basic emotions by researching facial expressions of human [5]. The basic emotions are “happiness”, “surprise”, “fear”, “sadness”, “anger”, and “disgust”. Seven emotions added “neutral” to the six basic emotions are frequently used in the field of facial expression recognition with image processing technologies [6-8].

In some researches applied human emotions to information systems, the emotion model defined by Plutchik is used. Plutchik defines dimensions which include eight basic emotions “joy”, “trust”, “fear”, “surprise”, “sadness”, “disgust”, “anger”, and “anticipation” [9]. These emotions presents a circumplex model. In this model, similar emotions are placed on the neighborhood and opposite emotions are placed on the opposite side. Plutchik also defines eight combinations of two basic emotions “love”, “submission”, “awe”, “disapproval”, “remorse”, “contempt”, “aggressiveness”, and “optimism”.

Parrot classify human emotions and define them as hierarchical tree structure [10]. The primary emotions are “love”, “joy”, “surprise”, “anger”, “sadness”, and “fear”. Other a lot of emotions more than one hundred are also defined as secondary emotions and tertiary emotions.

In the above emotion models, there are several common basic emotions such as “surprise”, “anger”, “sadness”, and “fear”. From these common basic emotions, we focus on the “surprise” and “fear” emotions in our research because these emotions can be appeared on driver’s facial expression in near-miss situation.

2.2. Emotions in Automobile Field

There are a lot of researches to detect emotions of drivers using sensor devices and grasp conditions of the drivers. Jones researched a method to detect driver’s emotion using his/her speech and developed a system for emotion recognition in a car [11]. Riener proposed a method to detect driver’s emotions using his/her heart rate variability [12]. Haak tried to detect driver’s emotion by analyzing his/her brain signals [13]. Moreover, Anzengruber developed a system which can detect driver’s emotions using his/her face surface temperature and evaluated the system on a driving simulator [14]. In this manner, we can detect driver’s emotions utilizing information collected by various sensor devices.

There are also applied researches utilizing the detected driver’s emotions. Jeon studied effects of driver’s angry and fear emotion on his/her driving performance [15]. He found angry drivers made more mistakes than fear drivers and fear drivers had heavier workload than angry drivers. Cilfford described traffic accidents cloud be decreased if a navigation system changes emotional expression of voice guidance [16]. The energetic voice of the navigation system was effective when driver was happy and subdued voice was effective for upset drivers in the experiments. In this manner, these researches indicate the driver’s emotions affect his/her driving performance significantly. To grasp not only car conditions but also driver’s conditions using his/her emotions is necessary and utilization of the driver’s emotions will be effective to develop a system which creates a high-accurate near-miss map.

3. PRELIMINARY STUDY

In this section, we describe several existed near-miss maps and their map generating methods. Then, we clarify a problem of the generating methods.

3.1. Existed Near-miss Maps

Many local governments and organizations create and offer near-miss maps for citizens and communities. To create and offer the near-miss maps, it is necessary to gather information of unsafe locations and warn the information to people. In this research, we focus on information gathering of unsafe locations.

There are two methods for the information gathering to create near-miss maps. The first method is manual information gathering. It gathers information manually by means of questionnaire surveys and interview researches (called "manual information gathering" hereafter). Most near-miss maps [17-20] are created by the manual information gathering. This method can gather high-accurate information. However, a great deal of time and effort is required and it is difficult to gather information in real time. The second method is automatic information gathering. It gathers information automatically by means of analyzing sensor data (called "automatic information gathering" hereafter). Typical example is the SAFETY MAP of Honda Internavi [21]. The automatic information gathering detects a heavy braking and bad road using acceleration sensors and so on. This method does not need time and effort. However, accuracy of the map is not so high because it uses only car conditions although it does not always show the driver feels a sense of danger at that time.

3.2. Manual vs. Automatic Information Gathering

To determine whether there is difference between manual information gathering and automatic information gathering, we compare the existed near-miss maps. In this preliminary study, we compared a near-miss map at the station of Muikamachi, Minamiuonuma city of Niigata in Japan [20] with a map of the SAFETY MAP [21] at the same location. The former map is created by manual information gathering (called "manual near-miss map" hereafter). The latter map is created by Honda Internavi which detects -0.2G acceleration and estimates a heavy breaking (called "automatic near-miss map" hereafter).

By comparing these two near-miss maps, we found the manual near-miss map said unsafe locations but not unsafe locations in automatic near-miss map. This problem is called “false negative error”. The term, false negative error is a result that indicates a given condition is not fulfilled but the actual condition is fulfilled. There are many false negative errors in the automatic near-miss map. Thus, these false negative errors mean the driver felt a sense of danger but did not brake hard. Especially, these errors were found around locations where a city areas prohibited speeding.

On the other hands, we found the automatic near-miss map said unsafe locations but not unsafe location in manual

		Condition (Correct Answer)	
		Positive	Negative
Test Outcome	Positive	True Positive	**False Positive**
	Negative	**False Negative**	True Negative

Table 1: False negative and false positive error

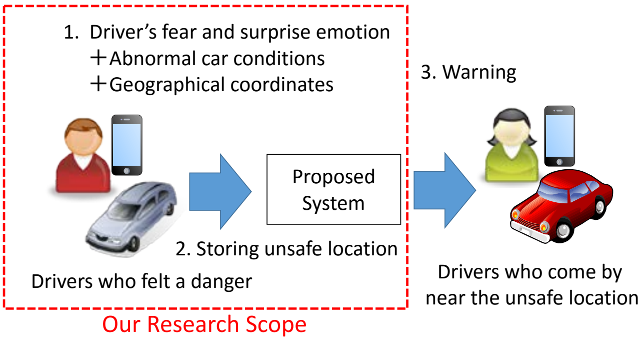

1. Driver's fear and surprise emotion
＋Abnormal car conditions
＋Geographical coordinates
Proposed System
2. Storing unsafe location
Drivers who felt a danger
3. Warning
Drivers who come by near the unsafe location
Our Research Scope

Figure 1: The model of our proposed system

near-miss map. This problem is called “false positive error”. The term, false positive error is a result that indicates a given condition is fulfilled but the actual condition is not fulfilled. There are many false positive errors in the automatic near-miss map. Thus, these false positive errors mean the driver did not feel a sense of danger but braked hard. Especially, these errors were found around locations where long straight road with a good view allowed speeding.

Table 1 shows supplementary explanation for the false negative and false positive errors. The false negative and false positive errors must be solved in order to increase accurate of the automatic near-miss map. As mentioned above, car conditions such as heavy brakes are not enough to detect actual unsafe locations where the driver feels a sense of danger. If we apply the driver's emotions to the automatic information gathering, more high-accurate near-miss map can be created without time and effort.

4. PROPOSED SYSTEM

We propose a system which creates a high-accurate near-miss map utilizing the driver's emotion.in order to realize a safe car society. Figure 1 shows the model of our proposed system. The proposed system utilize driver's fear and surprise emotions and abnormal car conditions to gather information of unsafe locations. By using not only car conditions but also driver's conditions, it improves accuracy of the automatic information gathering for unsafe locations. The proposed system uses the information of unsafe locations and warns of a danger to drivers who come by near the locations in order to encourage them to drive carefully. In this research, our research scope is the mechanism of information gathering for unsafe locations in a high accuracy.

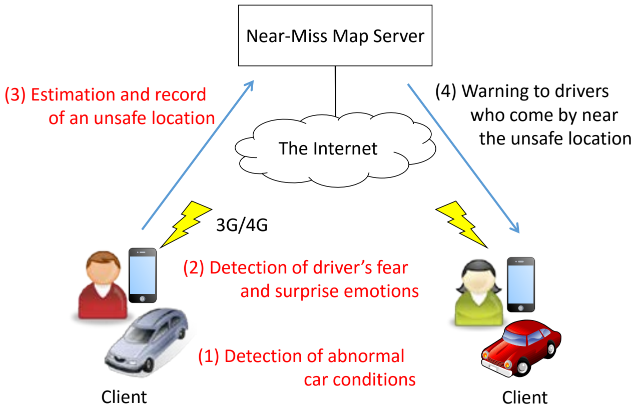

Near-Miss Map Server
(3) Estimation and record of an unsafe location
(4) Warning to drivers who come by near the unsafe location
The Internet
3G/4G
(2) Detection of driver's fear and surprise emotions
(1) Detection of abnormal car conditions
Client
Client

Figure 2: The architecture of the proposed system

The proposed system use general-purpose smartphones and does not need any dedicated devices so that more people can use the system. Therefore, we use only general sensors such as a camera and an acceleration sensor equipped in the general-purpose smartphones. Figure 2 shows the architecture of the proposed system. The procedure of proposed system is as follows.

(1) Detection of abnormal car conditions
(2) Detection of driver's fear and surprise emotions
(3) Estimation and record of an unsafe location
(4) Warning to drivers who come by near the unsafe location

In next section, we implement a prototype system using the above procedure.

5. IMPLEMENTATION

In this section, we describe implementation of functions (1) detection of abnormal car conditions, (2) detection of driver's fear and surprise emotions, (3) estimation and record of an unsafe location, and (4) warning to drivers who come by near the unsafe location. The function (4) is out of our research scope but we just implement it.

5.1. Detection of Abnormal Car Conditions

We detect heavy brakes to detect abnormal car conditions. An acceleration sensor equipped in a smartphone is used for the function of heavy brake detection. The function detects a certain measure of acceleration. In the prototype system, a smartphone is mounted on dashboard of a car and senses variation of the acceleration. When the variation value exceeded by -0.25G which is defined by reference to the other existed system for automatic near-miss map, the function recognizes it as a heavy brake.

5.2. Detection of Driver's Fear and Surprise Emotions

For the detection of driver's fear and surprise emotions, we use clmtrackr [22-23] which is a library of facial expression recognition. This library presents each emotion

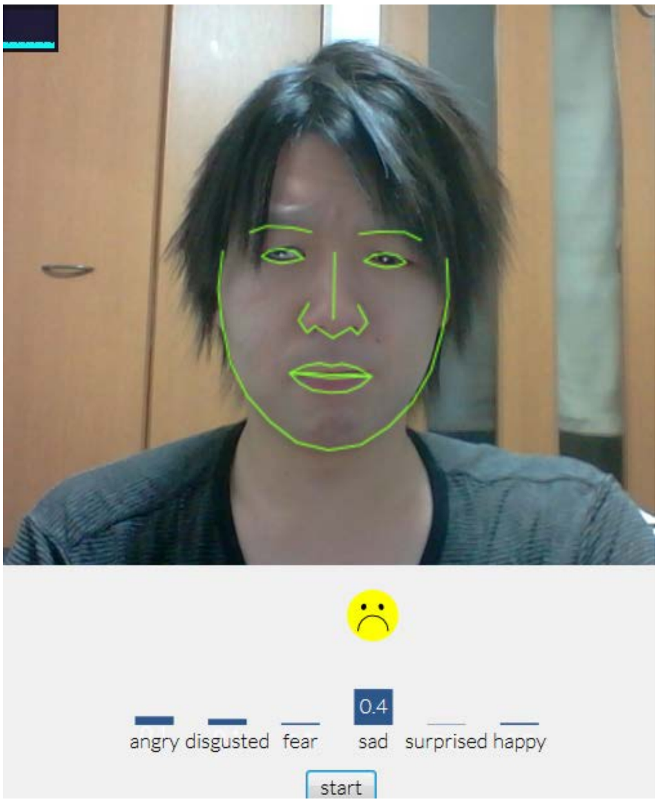


Figure 3: The detection of each emotion by facial expression image in real time

as a probability value from 0.0 to 1.0 by the facial expression. It uses a camera of the mounted smartphone and gets driver's facial expression images in real time. From the image, the library calculate the probability value for each emotion as shown in Figure 3.

5.3. Estimation and Record of an Unsafe Location

We conducted a preliminary experiment to assess a reference probability value of fear and surprise emotion for estimation of unsafe locations and set it to 0.3 with heavy brakes. However, it could not detect unsafe locations when the heavy brake was not detected in a city areas prohibited speeding. In such a case, we use a high probability value of fear and surprise emotions with no heavy brake to estimate unsafe locations in low-speed areas. The high probability value for fear and surprise emotions is used for 0.6 which is twice as much as 0.3 with heavy brakes. After the detection, information of the unsafe location is sent to a server on the Internet via 3G/4G wireless network. The client communicates with the server using WebSocket and the information includes geographical coordinates of an unsafe location, probability values of the driver's emotions and acceleration data.

5.4. Warning to Drivers

Figure 4 shows a user interface of the prototype system on a smartphone in order to detect unsafe locations and give a warning to the driver. The user interface presents a driver's current location on the map by using GPS equipped in the smartphone. The current location is updated in real-time. Unsafe locations are also displayed on the map by getting unsafe location information from the server. The driver's facial expression image is taken from a camera equipped in the smartphone and driver's emotions are detected in real time. The acceleration value is displayed for heavy brakes. The client communicates with the server every 5 seconds. It sends current location information and receives unsafe location information. When the driver comes close to an unsafe location, the client give a warning to the driver with alarmed sound and text. After a fixed time, the alarm is stopped.

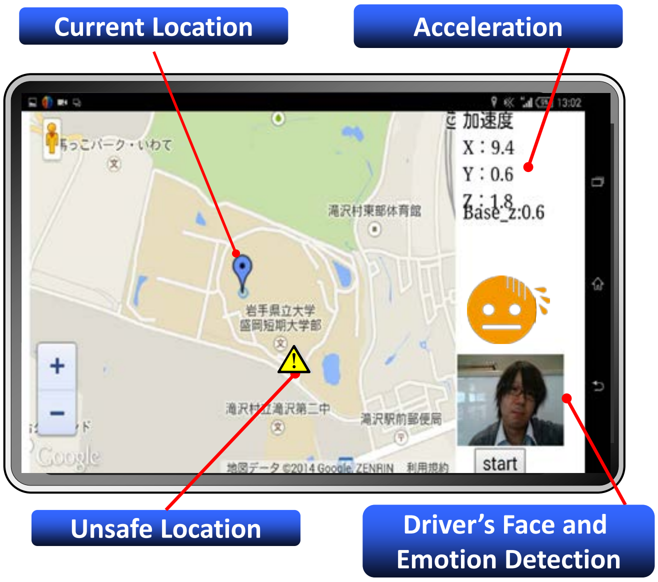


Figure 4: The user interface of the prototype system

6. EVALUATION

We evaluate the prototype system in terms of false-negative and false-positive errors. At first, we explain the experimental methodology. Then, we show the results and discuss about accuracy of the near-miss map.

6.1. Experimental Methodology

In the experiment, we create a manual near-miss map created by a questionnaire and an automatic near-miss map created by the prototype system. After that, we compare them if the automatic near-miss map achieves low false-negative and false-positive errors.

Figure 5 shows the driving route in the experiment. The route includes residential sections around our university, straight road where the driver can put on speed, sloping, and various types of roads. It is 17 km long and takes time for approximately 30 minutes to go around the route.

At first, we created a manual near-miss map by a questionnaire. The subjects are 30 students, who have driver license, of our university. Before conducting the questionnaire, we developed a system for the questionnaire

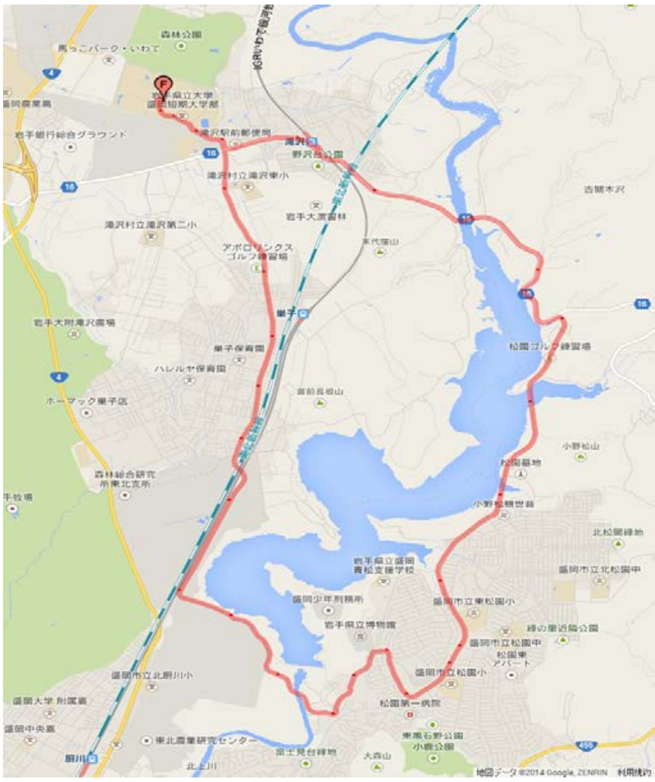


Figure 5: The driving route in the experiment

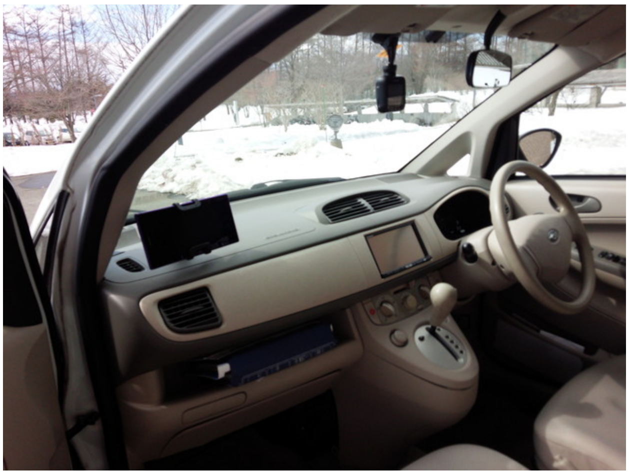


Figure 6: The experimental environment in the car

survey which presents a driving video on the route and current car location on a map. The subjects used the system and answered unsafe locations and their reasons. In this time, there were no dangerous scene in the video and we asked subjects to remember their experience when passed the locations. We defined the unsafe locations where the subjects of ten percent (3 subjects) answered it was unsafe in the questionnaire.

Secondly, we conducted an experiment on the road with a real car which mounted a smartphone on the dashboard for the prototype system. In this experiment, the subjects were 20 students, who have driver license, of our university. We

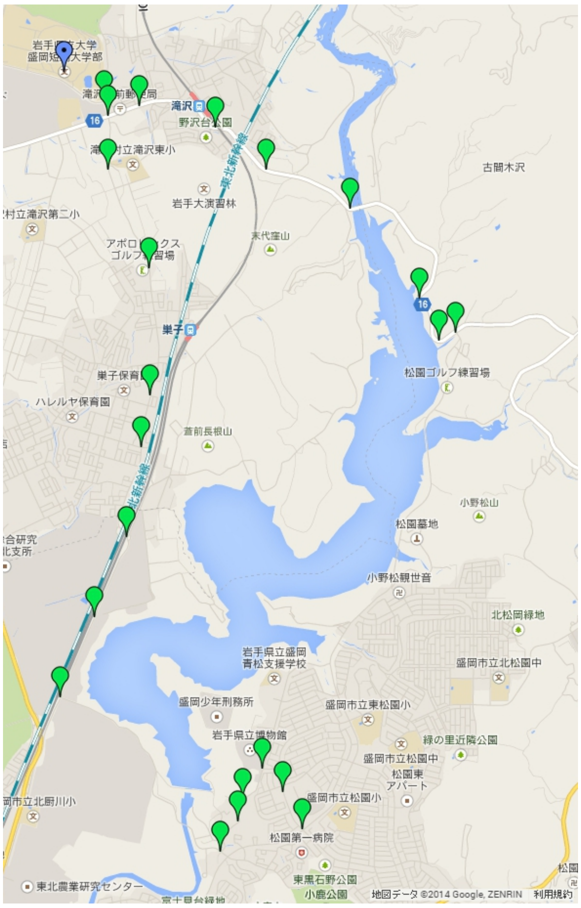


Figure 7: The manual near-miss map created by the questionnaire

suppose the proposed system is utilize driver's emotion. However, for the safety of the subjects, they rode in the front passenger seat of the car. We also record the driving video for analytical use. Figure 6 shows the experimental environment in the car.

6.2. Results and Discussions

Figure 7 shows the result of the manual near-miss map created by the questionnaire and figure 8 shows result of the automatic near-miss map created by the prototype system. Figure 9 shows the comparison result of these two maps. The circles show the locations of false-positive errors and the triangles show the locations of false-negative errors.

The locations of false-positive errors were three points. The one point was detected by heavy brakes and 0.3 probability value of fear/surprise emotions. The other two points were detected by 0.6 probability value of fear/surprise emotions. The rate of the false-positive errors was 14 % which was on 3 errors of 21 detected locations and the prototype system achieved low false-positive error rate.

The locations of false-negative errors were eight points. The rate of the false-negative errors was 36 % which was on 8 errors of 22 actual unsafe locations. The prototype system should be improve in terms of the unsafe locations. Possible reasons for the false-negative errors are low accuracy of the emotion detection. Since the emotion detection of facial

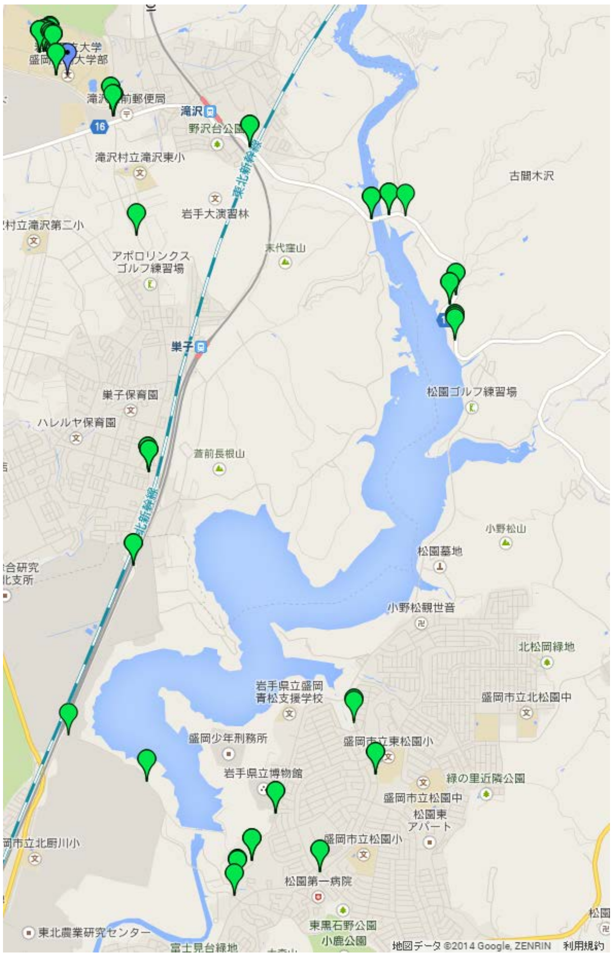


Figure 8: The automatic near-miss map created by the prototype system

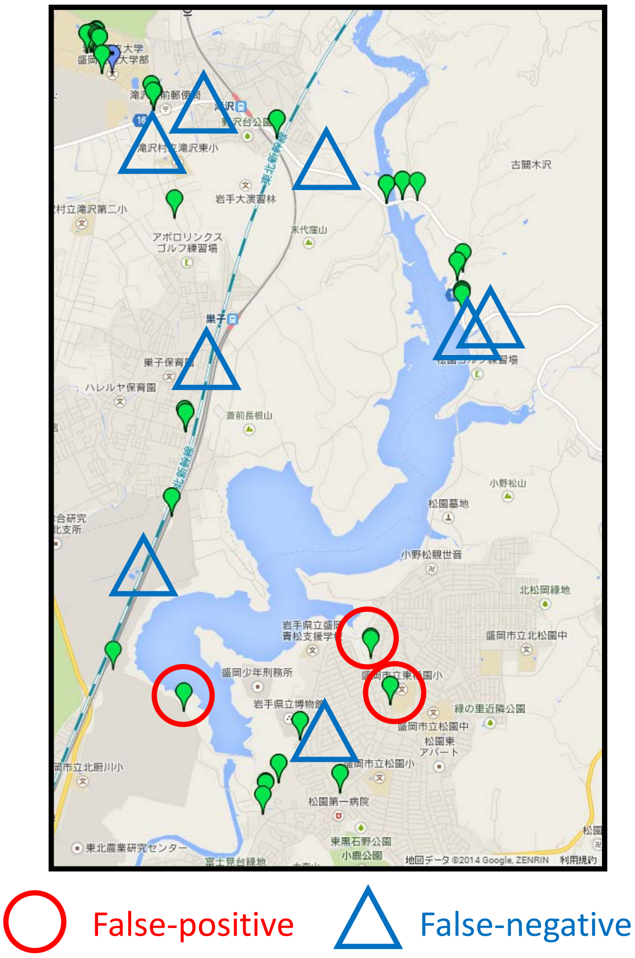


Figure 9: The result compared between the manual and automatic near-miss maps.

expression is difficult and has a limit of accuracy, the prototype system could not detected the subject's fear/surprise emotions. One of the solutions is to use together with other methods for emotion detection such as utilization of driver's heart rate, body temperature and so on.

7. CONCLUSION

In this paper, we proposed a high-accurate near-miss map system utilizing driver's emotions and developed the prototype system. From the evaluation experiment, we found our proposed system could achieve false-positive error rate although it required to improve false-negative error rate. For the future work, we will introduce other methods for emotion detection to reduce false-negative error rate.

REFERENCES

[1] Automobile Inspection & Registration Information Association, https://www.airia.or.jp/publish/statistics/ub83el00000000wo-att/01.pdf

[2] Japan Automobile Manufacturers Association., http://www.jama.or.jp/world/world/index.html

[3] National Police Agency in Japan, http://www.e-stat.go.jp/SG1/estat/List.do?lid=000001150496

[4] Galovski, T. and Blanchard, E.: Road rage: a domain for psychological intervention? Aggressive Violent Behavior, Vol. 9, Issue 2, pp. 105-127 (2004).

[5] Paul Ekman: Facial Expression and Emotion, American Psychologist, pp. 384-392 (1993).

[6] Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu: Automatic Classification of Single Facial Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, No. 12, pp. 1357-1362 (1999).

[7] Dragoş Datcu and Leon Rothkrantz: Facial expression recognition in still pictures and videos using active appearance models: a comparison approach, Proceedings of the 2007 international conference on Computer systems and technologies (CompSysTech), No. 112, pp. 1-6 (2007).

[8] Erica Costantini, Fabio Pianesi and Michela Prete: Recognising emotions in human and synthetic faces: the role of the upper and lower parts of the face, Proceedings of the 10th international conference on Intelligent user interfaces (IUI), pp. 20-27 (2005).

[9] Robert Plutchik: The nature of emotions, American Scientist, pp.344–350 (2001).

[10] W. Parrott: Emotions in social psychology, Psychology Press (2001).

[11] Christian Martyn Jones and Ing-Marie Jonsson: Automatic recognition of affective cues in the affective cues in the speech of car drivers to allow appropriate responses, OZCHI’2005, pp. 1-10 (2005).

[12] Andreas Riener, Alois Ferscha and Mohamed Aly: Heart on the road: HRV analysis for monitoring a driver’s affective state, AutomotiveUI’09, pp. 99-106 (2009).

[13] Paul van den Haak, Rinde van Lon, Jaap van der Meer and Léon Rothkrantz: Stress assessment of car-drivers using EEG-analysis, CompSysTech’10, pp. 473-477 (2010).

[14] Bernhard Anzengruber and Andreas Riener: “FaceLight” – Potentials and Drawbacks of Thermal Imaging to Infer Driver Stress, AutomotiveUI’12, pp. 210-216 (2012).

[15] Myounghoon Jeon, Jung-Bin Yim and Bruce N. Walker: An Angry Driver Is Not the Same As a Fearful Driver: Effects of Specific Negative Emotions on Risk Perception, Driving Performance, and Workload, AutomotiveUI’11, pp. 137-140 (2011).

[16] Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave and Leila Takayama: Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion, CHI’05, pp. 1973-1976 (2005).

[17] A near-miss map for elementary schools of Saitama city in Japan, http://www.city.saitama.jp/kita/001/001/003/p020992.html

[18] A near-miss map of Sodegaura city in Japan, https://www.city.sodegaura.lg.jp/soshiki/doboku-kensetsu/hiyarihattotizu.html

[19] A near-miss map of Shizuoka city in Japan, http://dataset.city.shizuoka.jp/dataset/anhoko

[20] A near-miss map of Minamiuonuma city in Japan, http://www.adclub.jp/common/hiyarihatto/minamiuonuma_niigata.html

[21] Honda SAFETY MAP, http://safetymap.jp/

[22] clmtrackr, https://github.com/auduno/clmtrackr

[23] Jason M. Saragih, Simon Lucey and Jeffrey F. Cohn: Deformable Model Fitting by Regularized Landmark Mean-Shift, International Journal of Computer Vision, Volume 91, Issue 2, pp. 200-215 (2011).

Analysis of Vehicle Information Sharing System by Microscopic Traffic Flow Simulation

Yusuke Takatori [*], Shunsuke Onodera [**] ,Tsubasa Sugiyama [**],Ryozo Kiyohara [***]
Kanagawa Institute of Technology, Japan
[*]takatori@ele.kanagawa-it.ac.jp, [**]{s1212049,s1312085}@cce.kanagawa-it.ac.jp,
[***]kiyohara@ic.kanagawa-it.ac.jp

Abstract – In this paper, the vehicle information sharing (VIS) system for safety driving is analyzed by microscopic traffic flow simulation. To share other vehicles' information, vehicles with an on-board communication unit (for short "OBU vehicle") communicate with other vehicles by vehicle-to-vehicle (V2V) communication technology. In this study, an intersection collision warning system (ICWS) is assumed as the target safety application. In the simulation analysis, two VIS systems are compared; one shares individual information of OBU vehicles (named VIS for ICWS-INDV) and another shares not only individual information but also forward vehicle information that is detected by a forward obstacle detection sensor (named VIS for ICWS-ODSU). The simulation result shows that the performance of the VIS for ICWS-ODSI is superior to that of the VIS for ICWS-INDV.

Keywords: ITS, vehicle information sharing (VIS), V2V communication, Intersection collision warning system (ICWS), forward obstacle detection sensor.

1 INTRODUCTION

In a blind intersection, vehicle-to-vehicle (V2V) communication allows vehicle information sharing (VIS) among vehicles that cannot see each other. However if vehicles with OBU of V2V communication unit (for short "OBU vehicle") and non-OBU vehicle are mixed, naturally, non-OBU vehicle information cannot be shared. Therefore a VIS system that shares not only individual OBU vehicle information but also forward vehicle information obtained by an in-vehicle forward obstacle detection sensor has been investigated [1]. An example of VIS for an intersection collision warning system sharing obstacle detection sensor information (ICWS-ODSI) is illustrated in Fig.1.

In the previous analysis [1], it is assumed that the VIS is conducted under a condition of the ideal communication (no delays and no packet collisions). Meanwhile, a standard of V2V communication which is expected to spread future (IEEE802.11p [2]) considers CSMA/CA as its medium access control (MAC). Because the transmission amount of the VIS that shares the individual driving information and the driving information of forward vehicle detected by an on-board obstacle detection sensor becomes larger than that of VIS that shares only the individual driving information, it is considered that packet collisions of the VIS that shares sensor information increase according to the traffic volume.

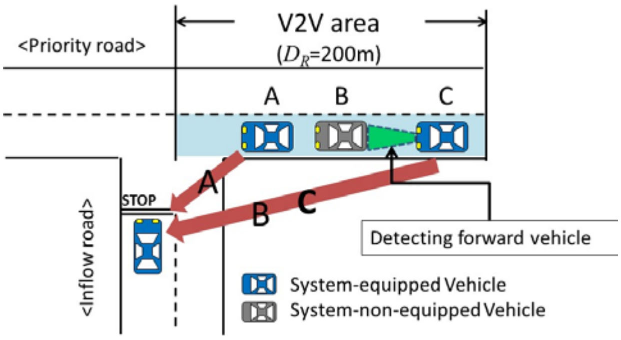


Fig 1: VIS for ICWS-ODSI

In this study, a microscopic traffic simulator in consideration of medium access control is constructed and the performance of VIS is analyzed. In section 2, V2V communication for VIS is explained. In Section 3, a microscopic traffic simulator for VIS performance analysis is described. In Section 4, the VIS performance is analyzed by the microscopic traffic simulator. Finally, Section 5 concludes this paper.

2 V2V COMMUNICATION

2.1 Standard for V2V Communication

Recent years, V2V communication standards of VIS for driving support system are considered in North America, EU and Asia region [3]. As for them, CSMA/CA-based broadcasting communication becomes a candidate for safety driving applications. These standards are summarized in Table 1. Particularly, the standard of North America /EU attracts attention of many other countries. These standards use IEEE802.11p for the physical and data link layer in the OSI reference model.

IEEE802.11p is a wireless communication standards for dedicated short-range communication (DSRC) that includes data exchange between high-speed vehicles and between the vehicles and the roadside infrastructure in the licensed ITS band of 5.9 GHz (5.85-5.925 GHz). Because it is a modified version of IEEE802.11a, many technologies are common. For such a process, IEEE802.11p adopts CSMA/CA for the medium access control. Meanwhile, IEEE 1609.X /WAVE [4] is a higher layer standard based on the IEEE 802.11p. WAVE (Wireless Access for Vehicular Environment) standards defines an architecture and a complementary, standardized set of services and interfaces that collectively enable secure vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) wireless communications. Together these standards provide the foundation for a broad range of applications in the transportation environment, including vehicle safety,

Table 1: vehicle to vehicle (V2V) communication standard

	Japan	North America	EU
standard/committee	RC006	IEEE802.11p/1609.x	IEEE802.11p/1609.x
Frequency band	715～725MHz	5.850～5.925GHz	5.850～5.925GHz
Number of ch.	10MHz×1ch	10MHz×7ch	10MHz×7ch
Modulation	OFDM		
Transmission rate	3～18Mbit/s	3～ 27Mbit/s, / 6～54Mbit/s	3～ 27Mbit/s
Medium access control	CSMA/CA		
Communication mode	Simplex broadcast	Broadcasting without ACK, Multicasting, Unicasting with ACK	

automated tolling, enhanced navigation, traffic management and many others. To share vehicle information, BSM (basic safety message) that is a message set of ITS application is assumed. Basic safety message contains vehicle safety–related information. It is periodically broadcast to surrounding vehicles.

2.2 CSMA/CA

CSMA/CA (Carrier Sense Multiple Access/ Collision Avoidance) is a protocol for carrier transmission in 802.11 networks. The procedure of medium access control by CSMA/CA is explained as follows:

- Carrier Sense: prior to transmitting, a node (vehicle) first listens to the shared medium (such as listening for wireless signals in a wireless network) to determine whether another node is transmitting or not.
- Collision Avoidance: if another node was heard, the node waits for a period of time (usually random) for the node to stop transmitting before listening again for a free communications channel.
- Transmission: if the medium was identified as being clear, it sends the frame in its entirety.

CSMA/CA achieves high-throughput. However, if the number of the nodes or the packet size becomes larger, its performance decreases by increase of the packet collision probability.

2.3 Intersection Collision Warning System sharing On-board Obstacle Detection Sensor Information [1]

Intersection collision warning system (ICWS) discourages a driver in a vehicle that approaches a blind intersection from entering the intersection if another vehicle approaches the same intersection from another direction. In this study, a T-junction model shown in Fig. 1 is configured. The T-junction combines a priority road and an inflow road. The road extending horizontally is called the priority road, and vehicles running on it are called priority vehicles. In the priority road, priority vehicles run on the left-hand side of the road (drivers obey Japanese traffic rules). Furthermore, the road that intersects at right angles is called the inflow road, and vehicles running on it are called inflow vehicles. Inflow vehicles should not tie up priority vehicles that approach the intersection. When the line-of-sight from the inflow road to the priority road is interrupted by obstacles (by a pedestrian, hedge, wall, building, etc.), the driver of the inflow vehicle might pass over with the priority vehicles and get into danger of causing a crossing collision. To avoid this danger, ICWS assists the inflow vehicle in entering the intersection safely by providing driving information (position, velocity, etc.) of the priority vehicles that approach the intersection. In this system, for the priority vehicle to broadcast its individual driving information repeatedly with short intervals, the OBU of the system must have two fundamental functions; a real-time V2V communication function and a real-time positioning function. By using these functions, the inflow vehicle's OBU can acquire other priority vehicles' driving information in real-time. In this paper, we call this ICWS that shares individual vehicle information as ICWS-INDV.

3 MICROSCOPIC TRAFFIC SIMULATOR FOR VIS PERFORMANCE ANALYSIS

Scenargie (Scenario Generation and Management Framework for In-Depth System Analysis and Evaluation) [5] can simulate mobility of nodes and communication of nodes simultaneously. Therefore, the mobility and communication models provided by the Scenargie framework are used for the construction of a microscopic traffic simulator for VIS performance analysis. In this study, a T-junction model shown in Fig. 1 is constructed. Then, the V2V communication module (based IEEE802.11p and

Table 2: Configurations

Road Configuration	T-junction with 2 lanes
distribution of HWD	Exponential distribution
V2V area	200m from the center of the Intersection
Penetration rate	100[%]
transmission interval	100[ms]
Payload size	128[byte]
Vehicle speed	40[km/h]
Average traffic volume	150,300,600[vehicle/h]

IEEE1609 (WAVE)) provided by Scenargie are implemented to the simulator.

To achieve a function that the OBU sends information obtained by in-vehicle obstacle forward detection sensor, it is assumed that the information contained into the message set of the Basic Safety Message (BSM) of WAVE [6]. OBU vehicle makes two packets (the payload size of each packet is 128 byte). One is a packet that contains its individual driving message and another is a packet that contains the driving message of its forward vehicle detected by the forward obstacle detection sensor. They are transmitted separately.

4 SIMULATION RESULTS

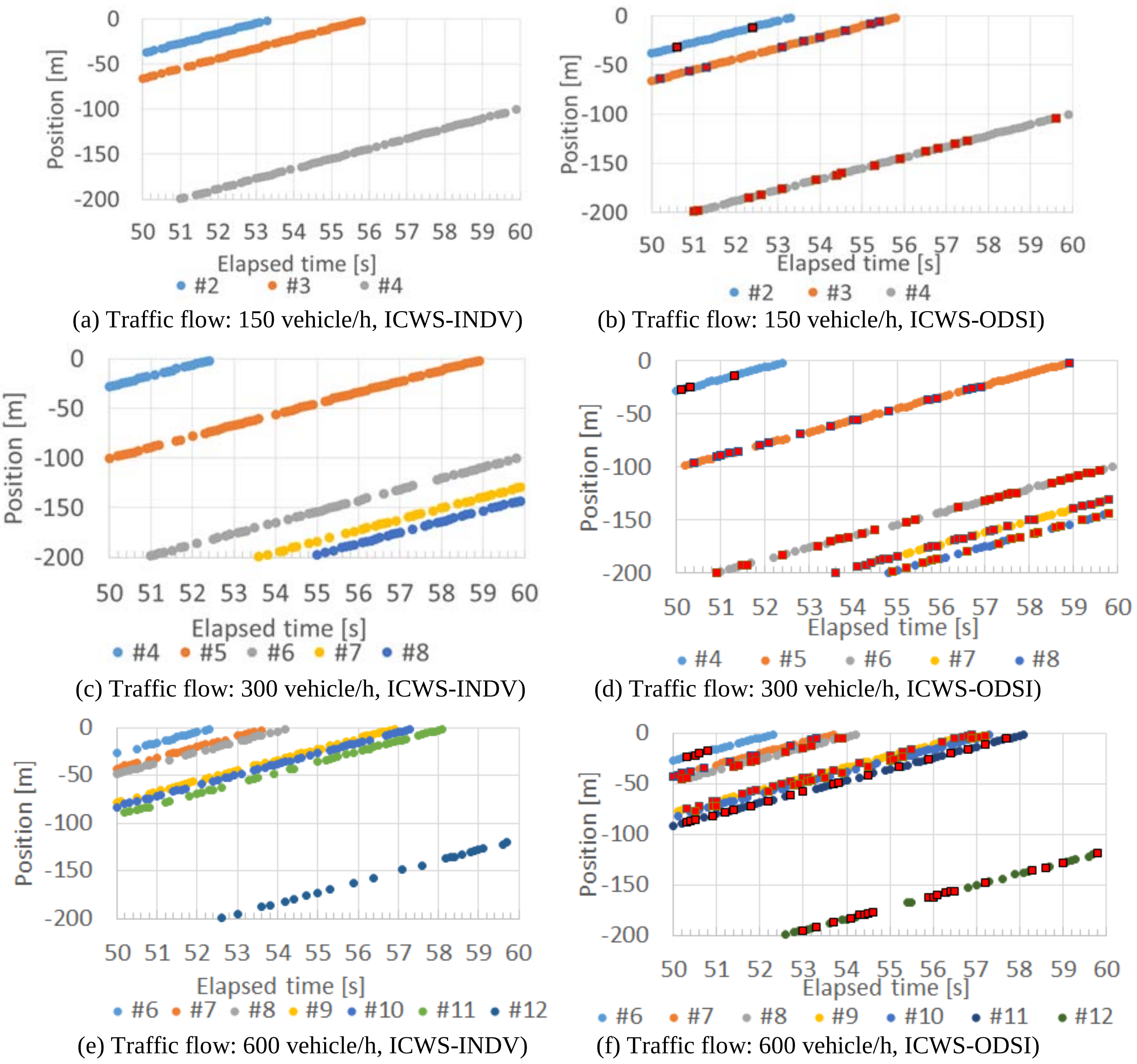
4.1 Simulation Method

In this analysis, left-turn merging situation is assumed. Priority vehicles are arrived at both side of the priority road of the T-junction. The distribution of their arrival interval (corresponds to the headway time distance) follows an exponential distribution. Those vehicles run through the 400m length of priority road and they update its position every 0.1 seconds. In this analysis, two VIS patterns are simulated. One is the VIS for the ICWS-INDV that send only one BSM packet that includes its individual information every 100 ms. Another is the VIS for the ICWS-ODSI that send two BSM packets; individual driving information packet and forward vehicle driving information packet obtained by a forward obstacle detection sensor every 0.1 seconds. The specification and parameters are listed in Table.2. In the table, penetration rate means a rate that OBU-equipped priority vehicles to all the priority vehicles. The simulation data were recorded from 30 seconds to 40 seconds.

4.2 Results

Simulation results are shown in Fig. 2. In these graph, dots indicate that the inflow vehicle obtained entire vehicle information in the V2V area. Besides, some dotted-lines break because of packet collisions or the packet non-transmission by the congestion of the communication channel. Whereas, red square-dots are supplemental information made from the information of a forward obstacle detection sensors and shared via V2V communication. Besides, the cumulative total number of recognized vehicles N_R (corresponds to the total dots in each graph), the cumulative total number of unrecognized vehicles N_U, and the cumulative total number of running vehicles N, the information omission rate (N_U / N) are summarized in Table 3. The information omission rate (IOR) of both of VISs increase with increasing of the average traffic volume. This is because the packet collisions and transmission failure by timeout are increased by the increasing of the average traffic volume.

These result shows that the IOR depends on the traffic flow. That is, the packet collisions and transmission failure by timeout are increased by the increasing of the average traffic volume. Moreover, in each traffic, VIS for ICWS-ODSI shows less IOR than that of ICWS-INDV. The case that a priority vehicle (subject vehicle) with VIS for ICWS-ODSI could not be recognized by the inflow vehicle occurs when both packets from the subject vehicle and the vehicle that follows the subject vehicle are lost. Consequently, it is indicated that negative influence caused by packet collisions and transmission failures is smaller than the positive influence of ODSI sharing.



(a) Traffic flow: 150 vehicle/h, ICWS-INDV)

(b) Traffic flow: 150 vehicle/h, ICWS-ODSI)

(c) Traffic flow: 300 vehicle/h, ICWS-INDV)

(d) Traffic flow: 300 vehicle/h, ICWS-ODSI)

(e) Traffic flow: 600 vehicle/h, ICWS-INDV)

(f) Traffic flow: 600 vehicle/h, ICWS-ODSI)

Figure2: Results of simulation

Table 3: Cumulative total number of vehicles

Traffic flow (vehicle/h)	VIS type	N_R	N_U	N	Information omission rate (%)
150	ICWS-INDV	152	31	183	0.169
	ICWS-ODSI	171	12		0.066
300	ICWS-INDV	215	107	322	0.332
	ICWS-ODSI	249	73		0.227
600	ICWS-INDV	253	152	405	0.375
	ICWS-ODSI	310	95		0.235

5 CONCLUSION

In this study, a simulator for the performance analysis of VIS for ICWS was constructed, which shares vehicle information that is obtained by on-board obstacle detection sensor of OBU vehicles. Moreover, the performance of the VIS for ICWS was analyzed in the situation of 100% of OBU penetration rate. Then two patterns of VIS for ICWS is analyzed; ICWS-INDV and ICWS-ODSI. For the result of ICWS-ODSI shows superior VIS performance that than ICWS-INDV. Furthermore, it is indicated that the VIS performance of both ICWS decrease with increasing the traffic volume because packet collisions and transmission failure by timeout. Meanwhile, it is indicated that negative influence caused by packet collisions and transmission failures is smaller than the positive influence of ODSI sharing. Future work is analysis of the penetration rate properties.

REFERENCES

[1] Y. Takatori, H. Takeo, “Crossing Collision Prevention System Using IVC and In-Vehicle Obstacle Detection Sensor”, Proc. of the IEEE ITSC2013 pp. 911 – 915, (2013)

[2] IEEE 802.11p-2010,” IEEE Standard for Information technology-- Local and metropolitan area networks-- Specific requirements-- Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments”

[3] http://www.soumu.go.jp/main_content/000019515.pdf/.

[4] Roberto A. Uzcátegui,” Wave: A tutorial”, IEEE Communications Magazine, Volume 47, Issue 5, pp.126-133, (2009),

[5] Scenargie: Space-Time Engineering, LLC , http://www.spacetime-eng.com/

[6] SAE J2735,” Dedicated Short Range Communications (DSRC) Message Set Dictionary”, (2009).

A Study of Driver's Behavior with Autonomous and Non-Autonomous Vehicles

Chihiro Miyazaki[†], Seiji Matsuyama[‡], Masashi Saito*, Yuichi Tokunaga**, and Ryozo Kiyohara[†]

[†]Kanagawa Institute of Technology, Japan
[‡]Graduate School of Kanagawa Institute of Technology, Japan
* Kanazawa Institute of Technology
** Mitsubishi Electric Corp.

Abstract - In recent years, many studies have been conducted on autonomous vehicles. These studies could not have been done without information processing and communication technologies (ICT) such as sensing technologies and V2X communication technologies. Many governments expect these technologies to reduce traffic jams, traffic accidents, and so on. In an ideal traffic environment, all vehicles on the road are controlled by autonomous technologies. However, in a real traffic environment, we think that vehicles on the road will be intermingled with autonomous and non-autonomous vehicles. Non-autonomous vehicles may have some functions of autonomous vehicles. In this paper, we focus on a junction of three roads and classify various cases using simple models. Then we clarify the problems in real traffic.

Keywords: ITS, autonomous vehicles, traffic jam, traffic flow

1 INTRODUCTION

Traffic jams constitute a serious societal problem; they cause critical issues such as a significant loss of time and bad influences on the environment. In Japan, 40% of driving time is lost owing to traffic jams. This is an enormous economic loss that is 20% higher than the loss in Europe or America [1]. In recent years, research and development of autonomous vehicles has been actively pursued in order to reduce these problems.

In order to realize an autonomous vehicle, two systems are integrated: an "autonomous type system" and a "collaborative type system." These systems are expected to upgrade[2]. An "autonomous type system" recognizes the driving environment using sensors that are installed on the vehicle. A "collaborative type system" recognizes the driving environment on the basis of acquired information from the outside using V2V and V2R communication technologies.

Many vehicles and traffic infrastructure equipped with DSSS correspond to the collaborative systems. In many cases, collaborative environments result in safe and smooth traffic. Reference [3] proposed a method of reducing right and left turn times at intersections. This method acquired information around the intersection in advance using Inter-Vehicular Communication (IVC) and Road-Vehicle Communication (RVC).

However, in the process of increasing the market for autonomous vehicles, there will be intermixed environments of autonomous vehicles operated by machine, and non-autonomous vehicles operated by humans. In these environments, there will be a possibility that the strong braking by surprising the human who cannot believe the computer and machine perfectly. In such a situation, it is possible that the behavior of autonomous vehicle will cause accidents or traffic jams. One cause is the difference in distance sensation seen by the human eye (see Figure 1) and the safe distance obtained by distance recognition and speed recognition by a machine (see Figure 2).

For example, we consider the following scene: A non-preferred vehicle turns right at a minimum inter-vehicle distance at an intersection where a non-autonomous vehicle that is driving on a preferred road is approaching. In such a situation, the non-autonomous vehicle may sense danger at a short inter-vehicle distance. The vehicle will reduce speed, and the vehicles behind it will similarly reduce speed. Such a situation shrinks the inter-vehicle distance between the following vehicles. This mechanism results in a traffic jam (see Figure 3).

This problem is a difference in judgment between a machine and a human. The point of difference cause the similarly situation between humans. However, humans can make different decisions owing to communication between drivers. The purpose of this study is to realize a driver support system that considers this difference, and to achieve a smooth traffic flow. This paper reveals the problems in an intermixed environment of autonomous vehicles and non-autonomous vehicles by a simulation

2 RELATED WORK

2.1 Problem of Collaborative-Type System

Even if a human and a machine confront a situation in the same way, they do not necessarily perform the same behaviors. If driver can select some behaviors, it occurs confronta-

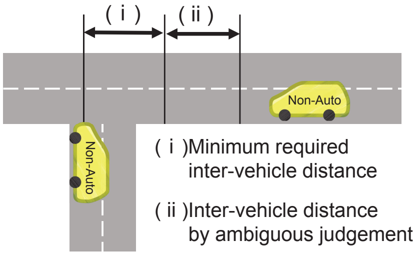

(i)
(ii)
Non-Auto
Non-Auto
(i)Minimum required inter-vehicle distance
(ii)Inter-vehicle distance by ambiguous judgement

Figure 1: Entry judgement by human

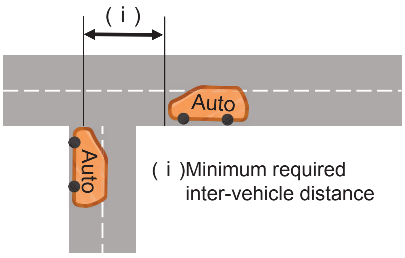


Figure 2: Entry judgement by machine

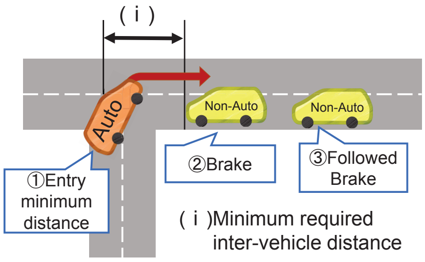


Figure 3: Traffic jam occurrence by autonomous vehicle

tion in between humans and machines [4] . If the human selects a behavior that is different from that of the machine, smooth communication is needed between the human and the machine in order to prevent deterioration of acceptability of human side.

Humans must understand the behavior of the machine and its intent in order to communicate smoothly between human and machine. The following method for this problem is proposed in [5] :

1. Information that the basis for judgment of machine can understand.

2. Information that becomes a clue to understanding the machinefs intentions.

3. Information sharing between human and machine.

4. Information that helps to understand the limits of the capacity of a machine.

If the machine is able to present easy-to-understand information to the human, the human will understand the machine and can accurately rate its ability.

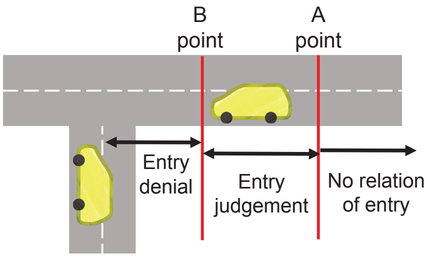


Figure 4: Driver judgement on the priority road

2.2 Problem of V2V System at Intersection

When a driver judges at an intersection using inter-vehicle communication, it is necessary to perform determination along the human behavior. If a driver decides to enter into an intersection by using vehicle communication, the driver needs to define the timing, etc., in advance. In [6] , experiments with a real vehicle and a simulator are conducted from an ”A-point” to a ”B-point.” The ”A-point” is the border where the priority vehicle begins the judgment of acceptance propriety of non-priority vehicle. The ”B-point” is the border where the priority vehicle refuses entry to the nonpriority vehicle. The margin time between the ”A-point” and ”B-point” where the priority vehicle travels at a velocity of $50[km/h]$and $30[km/h]$are $1.5[s]$and $3.4[s]$, respectively. It is considered hard that non-priority vehicle to enter preference road using cueing from this result. If one can indicate his or her intention before the A-point using inter-vehicle communication, can make judgment time of acceptance propriety of nonpriority vehicle.

As results has been and can be expected to achieve secure communication.

3 TRADITIONAL MODEL

In this paper we consider an unsignalized junction that has a crossing with a preference road and non-preference road. Intersection is regarded as road shape that is accident, and traffic jams tend to occur [7] . Smoothing the traffic flow of the intersection leads to problem-solving faced by automobile society. In addition, in provincial cities where means of transport are limited, it is considered that this problem is more serious.

3.1 Problems in Today’s Traffic

In the current environment, where there are no autonomous vehicles, humans perform all driving operations using three phases: cognition, judgment, and operation. It is necessary that the driver correctly performs all phases for safe driving operation.

A survey of the literature [8] indicates that 75% of drivers have made a mistake during the cognition phase. Eighty percent of the information that humans acquire while driving is visual information. Situations in which visual information is limited include those that shut down the information unconsciously by driver prejudice, and those that cause poor visibility by obstacles and weather, and the like. Owing to delays in judgment and unreasonable operation, the risk of an accident increases when information remains insufficient during the cognitive phase.

An unsignalized T-junction has many factors that cause mistakes that result in traffic accidents. One of these factors is the restriction of visual information. Other main factors include the following:

- Static obstacles
  Buildings and walls at the corners of an intersection are representative examples of static obstacles. They limit visual information to the side. Therefore, traffic information in the lateral direction of the lane is insufficient..

- Dynamic obstacles
  Heavy vehicles, including buses and trucks, are representative examples of dynamic obstacles. If these vehicles drive around another vehicle, information obtained from the direction, that exist vehicle, is limited. If heavy vehicles are driving in front of the vehicle, information obtained from the front and side is limited.

- Drivers prejudice
  The prejudice of the driver with regard to a specific location can lead to an accident. If the driver knows that the traffic volume of a nonpriority road is small, that driver is neglected cognitive of nonpriority road. As a result, an accident with the nonpriority vehicle occurs. This accident is caused by driver prejudice.

In order to avoid accidents caused by these factors, the driver requires prudent peripheral confirmation and safe driving operation.

However, there are individual differences in the surrounding confirmation of the status quo. This is not unique to specific drivers. In fact, there are cases in which the results are different for the same driver.

Judging when to enter the intersection is a situation in which individual differences occur. In this situation, determining when to enter is based on the speed of the oncoming vehicle and the distance between the oncoming vehicle and the driverfs own vehicle.

n this research, the speed of oncoming vehicle is v [km/h], and the time that the driverfs own vehicle takes to arrive at the intersection is t [s]. Equation 1 calculates the inter-vehicle distance y [m] in which the driverfs own vehicle is required to turn left or right at the intersection.

$$y = \frac{1000v}{3600}t \qquad \text{(i)}$$

A margin time of 2 s is added to the minimum inter-vehicle distance because 2 s is the inter-vehicle distance that is recommended for proper human judgment [9] .

$$y = \frac{1000v}{3600}(t+2) \qquad \text{(ii)}$$

If the oncoming vehicle is a heavy vehicle, a margin time of 3 s is added to the minimum inter-vehicle distance:

$$y = \frac{1000v}{3600}(t+3) \qquad \text{(iii)}$$

If the speed of the oncoming vehicle and the intersection passing time of the driverfs own vehicle are, respectively, $40km$ and $4.5s$, the resulting calculations for Equation 1, Equation 2, and Equation 3 are shown in Figure 5.

Inter-vehicle distance that is increased as a result of vague judgment can create a new traffic jam. In addition, if pedestrians and bicycles are on the sidewalk, further delays in judgment time occur, and it is expected that the traffic flow will worsen. It is difficult to perform such driving operations, considering safety and impact on traffic flow, in current surrounding recognition.

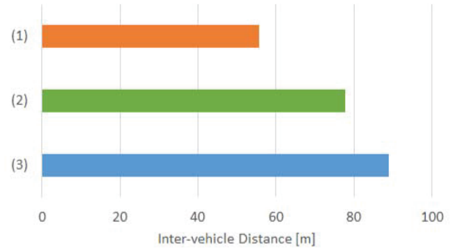


Figure 5: Inter-vehicle distance required for entry

3.2 Ideal Traffic Flow by Autonomous Vehicle

A collaborative system that includes IVC and RVC is said to be effective in solving this problem. By sharing information about other vehicles and infrastructures, information to be used for making judgments is increased. In addition, in an autonomous vehicle, because a machine performs all driving operations with regard to cognition, judgment, and operation. and exact driving operation that does not through of the human ambiguous judgment can be expected.

Reference [3] considered an intersection of low visibility with obstacles at the corners, proposed congestion mitigation techniques that used IVC and RVC, and showed the effectiveness of those techniques through a traffic simulation.

In the proposed method of conventional, share the vehicle information of around the T-junction at between vehicles, driver perform efficient judgment of turn of right or left.

The simulation compared an ITS environment and a non-ITS environment. In the ITS environment, an ITS vehicle communicates by IVC and RVC. In the non-ITS environment, non-ITS vehicles do not communicate with each other.

Figure 7 shows a comparison of the number of completions of right or left turns in this simulation. Figure 8 shows a comparison of the average waiting time of the vehicles in this simulation. In the experiment results for the ITS environment, the average waiting time is shortened, and has increased the right or left turn completion. The average waiting time in the ITS environment increased for left turns. By contrast, the average waiting time in the non-ITS environment was approximately 4 s regardless of right or left turns. From this, in the ITS environment, it is considered performing the effective turn. All vehicles in an ideal ITS environment are assumed to perform cooperatively by communicating with each other. The realization of an ideal traffic flow by performing efficient driving operations can be expected

However, it is difficult to imagine that all vehicles can change into autonomous vehicles that perform efficient driving operations. This is also evident when viewed from the current situation of the automobile society that the vehicle of models of the new and old a variety has come and go. Because the typical lifetime of a vehicle is 10 years, we are at least 10 years from being mandated by legislation to have mixed environments of autonomous and non-autonomous vehicles. In a situation where autonomous and non-autonomous vehicles are mixed in an autonomous vehicle dissemination stage, a system to achieve a smooth traffic flow is obtained that

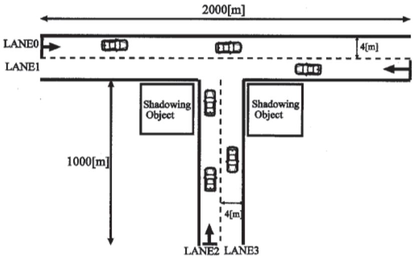


Figure 6: Road model that was used in [3]

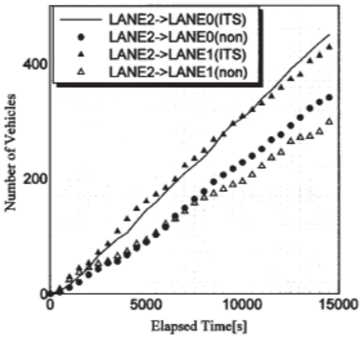


Figure 7: Number of right-turn and left-turn completion vehicle in LANE 2 in [3]

considers the impact that autonomous vehicles have on non-autonomous vehicles.

4 ENTERING THE T-JUNCTION MODEL IN MIXED SITUATION

4.1 Vehicle Model

In the mixed environment that is assumed in this paper, there exist autonomous vehicles, ITS vehicles, and non-ITS vehicles. The autonomous vehicle uses an autonomous-type system and a collaborative-type system, and a machine performs driving operations. The ITS vehicle uses a collaborative-type system, and a human performs the driving operations. The non-ITS vehicle does not use a collaborative-type system, and a human performs the driving operations. Features of these vehicle are listed in Table 1.

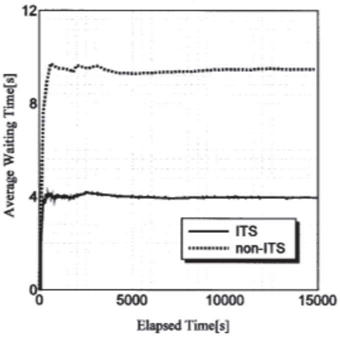


Figure 8: The average waiting time of vehicle to right-turn from LANE 2 to LANE0 in [3]

Table 1: Features of the vehicle

	Communication	Right or Left turn	Operaion subject
Autonomous	o	Formula 1	Machine
ITS	o	Formula 1 Formula 2	human
non-ITS	x	Formula 2	human

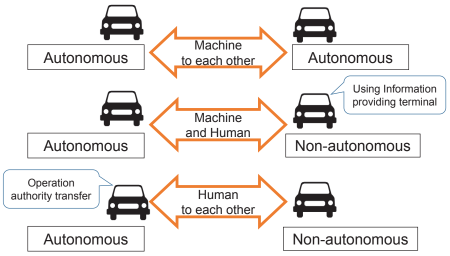


Figure 9: Method of Intention transmission by each vehicle

The driver of an ITS vehicle receives information regarding efficient driving operations in order to perform right or left turns at a minimum inter-vehicle distance. The information is provided by a terminal. However, there is a case of performing a driving operation that takes the long inter-vehicle distance by judgment of driver.

In the original traffic environment, there are pedestrians, bicycles, and heavy vehicles. In addition, if the machine cannot make a decision, machine is not always to perform operation driving in order to transfer driving operation to human. However, in this paper, for the sake of simplicity, we consider an environment in which only these three types are present.

4.2 Concessions in Mixed Situation

Figure 9 shows the operation of each vehicle in a mixed situation. In a T-junction with a mixed situation, an autonomous vehicle must convey its ideas of driving operation to the other driver in order to perform a smooth concession of the right-of-way.

When misses occur, the ITS vehicle and non-ITS vehicle cannot understand the driving operations, and there is the possibility that the traffic flow is stagnant.

If a human drives an ITS vehicle, we can solve this problem by conveying the intent of a driving operation through the information terminal. Therefore, in such a situation, we consider that it is necessary to solve this problem by communicating between human beings.

4.3 Entering the T-junction

It is possible that efficient driving operations of an autonomous vehicle may be dangerous to humans. Thus, effective driving

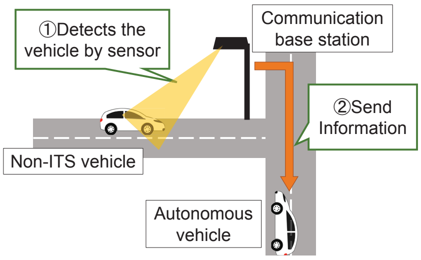


Figure 10: Acquisition of ITS non-equipped vehicle information due to the RVC

operations for smooth traffic flow becomes a factor in traffic jams.

A representative example is the situation shown in Figure 3. In Figure 3, traffic jams are caused by oncoming vehicles that were surprised by autonomous vehicles that approach at the required minimum inter-vehicle distance.

4.4 Acquire Vehicle Information

ITS vehicles acquired information using only IVC and RVC in the traditional model. However, this method cannot acquire vehicle information of non-ITS vehicles.

In this paper, in order to acquire information of non-ITS vehicles that do not have communication equipment, a communication base station is installed at the intersection. This base station detects vehicles using sensors and cameras (Figure 10). Reference [11] was unable to detect all vehicles when using a stereo camera. In addition, bottlenecks in provincial cities are limited. We do not consider installing a sensor to be a problem.

5 EVALUATION BY SIMULATION

5.1 Simulation Method

We ran a simulation using the multi-agent simulator of the traffic simulator ”Scenargie” [10] .

Figure 11 shows a road model that has an unsignalized T-junction. There is a priority road and a nonpriority road.

Priority road length and width are 2000 m and 16 m, respectively, and nonpriority road length and width are 1000 m and 14 m, respectively. Installed building that become departure point and destination at exit of each road. With regard to communication, for simplicity, the vehicles can always transmit and receive.

Three types of vehicle are used: autonomous vehicle, ITS vehicle, and non-ITS vehicle, as described in section 4.1.

Table 2 shows the occupancy rate of each vehicle in the simulation environment. Table 3 lists the specifications for the simulation. For simplicity, we do not distinguish between the driver and vehicle. In reality, however, the characteristics of the driver should be considered because unique local rules often exist.

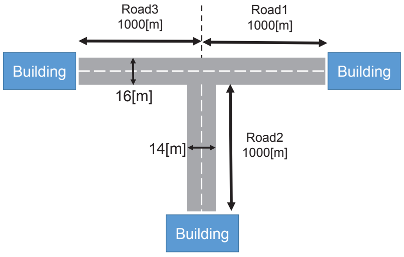


Figure 11: Assumed road model

Table 2: Occupancy rate of each vehicle

Pattern	Percentage		
	autonomous	ITS	non-ITS
A	20	20	60
B	50	20	30
C	80	20	0

5.2 Simulation Results

Figure 12 shows the average longest congestion length of each road when the vehicles occur at 600 [units]. On the priority road, even if the occupancy rate of each vehicle is altered, no significant changes were observed. However, on the nonpriority road, the lower the occupancy rate of autonomous vehicle, the longer the length of the traffic jam.

We consider the problem that, as the occupancy rate of non-autonomous vehicles increases, opportunities for a human to judge a left or right turn increases. In the simulation, when the lead vehicle on the approach road is a non-autonomous vehicle, situation could confirm to increase succession vehicles by take time to right-turn or left-turn.

Figure 13 shows changes in the maximum traffic jam length for different traffic flows and occupancy rates on Road 2. Although there is a difference in the amount of the increase, in every pattern it can be seen that maximum traffic jam length increases with the traffic flow. The higher the occupancy rate isthe smaller the increase in maximum traffic jam length. A traffic jam mitigation effect by diffusion of autonomous vehicle has affecting effectively.

Figure 14 shows the transition in traffic jam length when vehicles occur at 600 units. Figure 15 shows a comparison of the average waiting time at an intersection when vehicles occur at 600 units. In any pattern, the traffic jam length and

Table 3: Simulation conditions

Simulation time	7200[s]
Position update interval	0.1[s]
number of Occurrences Vehicles	150[units], 300[units], 800[units]
Vehicle occurences Interval	exponrential distribution
Definition of Traffice Length	Number of vehicle that has driving at 10[km/h] less before 500[m] of interstion or stopping

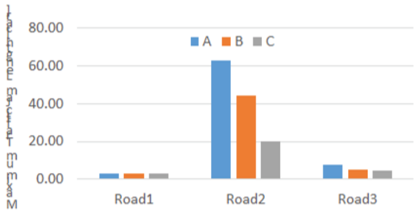


Figure 12: Maximum traffic jam length at occurrence vehicle 600 units

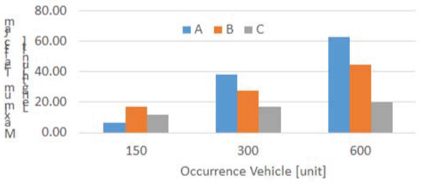


Figure 13: Compare maximum traffic jam length in Road 2

average waiting time increased as the number of vehicle occurrences increased. The peak time of the traffic jam length is in the vicinity of 1500 [s]. However, traffic jam cancellation time and peak time of the average waiting time were different for each pattern.

The times when the traffic jam length started to decrease are as follows: for pattern A, in the vicinity of 2500 [s]; for pattern B, in the vicinity of 2000 [s]; and for pattern C, in the vicinity of 1500 [s]. In addition, the traffic jam lengths are shortened to less than 10 m, and traffic jams are resolved after 500 s from the start of the decrease.

The peak time of the average waiting time was as follows: for pattern A, in the vicinity of 2000 [s]; for pattern B, in the vicinity of 2500 [s]; and for pattern C, in the vicinity of 3000 [s]. This corresponds to traffic jam cancellation times. We consider that the waiting time of a vehicle is reflected in the peak time of the average waiting time.

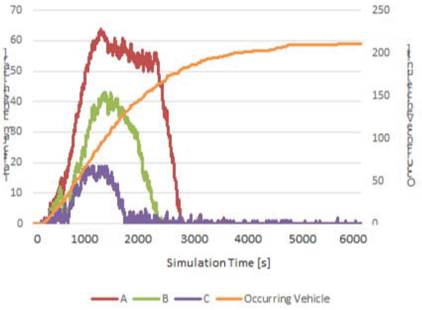


Figure 14: Compare Traffic Jam Length in Road 2

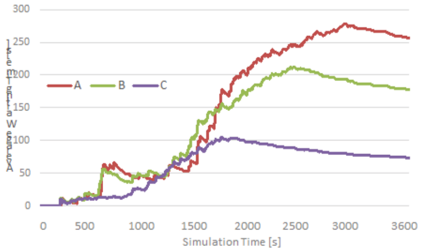


Figure 15: Compose Average Waiting Time in Road 2

5.3 Considerations

The traffic jam length and average waiting time are reduced as the occupancy rate of autonomous vehicles increases. In pattern A that has an occupancy rate of 80%, the traffic jam length and average waiting time are reduced compared with other patterns. Thus, we can say that a traffic jam reduction effect resulting from efficient right-left-turn judgment is applied effectively.

In pattern B and pattern C, the traffic jam length and average waiting time are approximately two to three times those of pattern A. When the occupancy rate of autonomous vehicles is high, the number of vehicles that cannot smoothly judge right or left turns increases. When considering the time loss when operating non-autonomous vehicles, we consider that the traffic jam length and average waiting time are increasing.

When the occupancy of autonomous vehicles is low, if we are unable to achieve a satisfactory traffic jam mitigation effect, the penetration speed to market may could slow down. The premise for the conventional method was that all vehicles operate efficiently. However, it is difficult to solve the problems of a mixed environment by using the conventional method. Thus, we need a new method that considers a mixed environment. An example of this method is shown below:

- Share the on-vehicle equipment information using IVC
  In the conventional method, the infrastructure side detects non-autonomous vehicles, and the result had sending other vehicles. However, in places where a communication base station is not installed, non-autonomous vehicles cannot be detected. Hence, an autonomous vehicle can recognize a non-autonomous vehicle by using sensor and radar, and then share this information between vehicles using IVC. Thus, we can expect to use peripheral vehicle recognition that does not depend on infrastructure equipment.

- Efficient Entry Permit by Autonomous Vehicles
  In simulation of this time, the lead non-autonomous vehicle cannot enter smoothly onto the priority road, and

it causes a traffic jam. To smooth the traffic flow on the entry road, we consider a method where the autonomous vehicle modulates the inter-vehicle distance with the front vehicle, and the non-autonomous vehicle conducts a prompt ingress. To realize this method, need to get swiftly intersection information than conventional method. Therefore, we consider the sharing of non-autonomous vehicle information using IVC.

6 CONCLUSIONS

In this paper, to realize a driver assistance system in a mixed environment with autonomous vehicles and non-autonomous vehicles, we simulated an unsignalized T-junction. This allowed us to clarify the problems in the mixed environment.

As a result, when the occupancy rate of autonomous vehicles is high, a traffic jam mitigation effect was effective. However, when the occupancy rate of autonomous vehicles is low, we are unable to achieve a satisfactory traffic jam mitigation effect. When the occupancy rate of non-autonomous vehicles is high, the number of vehicles that cannot perform a smooth right turn or left turn increases. Be affected by time loss by this operation, became we cannot achieve a satisfactory traffic jam mitigation effect.

From the above, a future task is to consider a mixed environment of autonomous vehicles and non-autonomous vehicles, to devise a method that solves these problems, and to conduct an evaluation. In addition, these problems are very important in local cities because there are limited means of transportation aside from vehicles. Signals are not always installed at T-junctions in local cities. If there is a signal, we consider that a vehicle turning right from a priority road onto a nonpriority road has the same problem. Right turns are important problems in local cities. When many vehicles are waiting to make a right turn, there is a case to block the road straight ahead. To solve this issue, we are considering a method where an approach is performed by a vehicle that is not equipped with a communication device.

In the future, we plan to analyze right turns at general intersections, and to propose a method for communication between drivers that does not use telecommunication.

REFERENCES

[1] Ministry of Land, Infrastructure, Transport and Tourism, WHITE PAPER ON LAND, INFRASTRUCTURE, TRANSPORT AND TOURISM IN JAPAN, 2015(Japanese)

[2] Prime Minister of Japan and His Cabinet, ITS roadmap 2015

[3] Yuta Wakui, Kohei Ohno, Makoto Itami ”A Study on Reducing Traffic Congestion at Intersections using Inter-Vehicle and Road-Vehicle Communications” 18th ITS World Congress, 2011

[4] Yuichi SAITOuDriver Assistance Technology Collaborating with HumansvReliability Engineering Association of Japan : 37(5), 242-249, 2015-09-01(Japanese)

[5] Toshiyuki INAGAKIuSafety and Peace of Mind in Human-Machine Collaborations ? Discussions from Human-Centered Automation Points of View -vThe Japanese Council of Traffic Science,vol.9, No.1, pp-11-20, 2010 (Japanese)

[6] Nakaho NUMATA , Masanori TAKEMOTO, Yasunari KUBOTA , Ryuto TOMINAGA , Seiya SHIDO , Hiroki KITAJIMAuErgonomic study for the vehicle to vehicle communication at intersections without traffic signalsvThe Japan Society of Mechanical Engineers,82(835), 15-00569-15-00569, 2016

[7] Ministry of Land, Infrastructure, Transport and Tourism, WHITE PAPER ON LAND, INFRASTRUCTURE, TRANSPORT AND TOURISM IN JAPAN, 2016(Japanese)

[8] Institute for Traffic Accident Research and Data Analysis, ITARDA InformationCNo.33v(2001)

[9] HondauThe Safety Japan No.434v, 2009-06

[10] Space-Time Engineering, LLC, https://www.spacetime-eng.com/

[11] Takato ShimodaC Yusuke TakatoriCHideya Takeo,A Study on a surrounding vehicle information collection system for VIS, ITS IEE-ITS . ITS 114(508), 11-16, 2015

A Framework to Generate Mobility Traces with Public Travel Survey and Its Application to Evaluation of Wi-Fi Offloading with Vehicles

Hiroki Masano[†], and Tomoya Kitani[‡]

[†]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[‡]Department of Informatics, Graduate School of Integrated Science and Technology,
Shizuoka University, Japan
{h-masano, t-kitani}@kitanilab.org

Abstract - Realistic mobility traces are important for the simulation of a communication system with vehicles. Some of public travel survey data have been available as realistic mobility traces. The temporal granularity of these survey data is hourly or minutely and it is too rough to be used for VANET simulations. VANET simulations need to use finer grain of mobility trace such as movement traces of every second. In this paper, we propose a framework of the achievement method for complement of a travel survey data with a traffic flow simulator: People Flow Data by the University of Tokyo is used as the travel survey data and SUMO is used for the traffic flow simulator. As an application of the generated realistic mobility trace data, we evaluate an opportunistic Wi-Fi offloading with a cellular phone carrier's actual Wi-Fi spot map. In the offloading method, moving vehicles ferry mobile data to the nearest Wi-Fi spot. As a result, we have derived each moving vehicle's encounter frequency and connection time of a Wi-Fi spot. The encounter frequency is estimated about 26 seconds and the connection time is about 23 seconds in a simulation of downtown Tokyo.

Keywords: Mobility trace, Person trip survey data, People flow data, Traffic flow simulator, SUMO, Mobile data offloading, Opportunistic Wi-Fi offloading.

1 INTRODUCTION

These days intelligent transport systems (ITS) are actively researched, and many novel services utilizing traffic flow have been proposed. Such services include vehicle-to-vehicle communication systems and message ferry systems with vehicles. The performance of these systems depends on the physical behavior of vehicles. Accordingly, a realistic mobility trace is necessary to evaluate such services.

To obtain a vehicle's trajectory with GNSS (Global Navigation Satellite Systems) is getting popular. This trajectory is currently the most realistic mobility trace. Cellular phone carriers, car companies providing telematics services, and navigation system companies have collected much amount of vehicles' trajectory data, but such data are not opened to the public or it requires a lot of costs. On the other hand, some public taxi and bus trace datasets are available, but the amount of such data is not enough to simulate a whole city.

Actual temporal-spatial data like an observed data and a social survey data represent crucial information for simulated real-world traffic. However, conventional datasets of trajectories are represented by minutely position data, so it is lacking time step dependent accuracy for VANET simulations. Such temporal-spatial data are generated from statistical OD (Origin/Destination) data and all the passing points along each path are not directly observed. They are just estimated and it may cause a lack of spatial accuracy.

As related work, several methods to gather detailed OD data from such rough temporal-spatial data have been proposed[1][2][3]. In the methods, the rough temporal-spatial data is given as an input to a traffic flow simulator to simulate behavior of vehicles. The mobility traces generated by the methods are regarded as more realistic. The methods use a variety of statistics and datasets to generate trace data. The accuracy of the trace data depends on that of their original data. Each of the methods is designed to use dataset or statistics of a certain city or area as its input, and therefore these methods are not generalized to be applied to a variety of cities.

Public agencies in a variety of countries such as National Household Travel Survey (NHTS) of U.S. Department of Transportation provide Travel Survey data[4][5][6]. OD matrix and travel time are common survey items in a lot of countries like the USA, Japan and Netherlands. We propose a framework for generating complemented mobility traces with travel survey data and a traffic flow simulator. In this paper, we target at the travel survey data of JICA (Japan International Cooperation Agency) and the University of Tokyo, which includes 30 Japanese and Asian cities, as the input dataset. Our method generates mobility traces by complementing such a travel survey data with the traffic flow simulator SUMO[7].

The Center for Spatial Information Science, the University of Tokyo (CSIS) researches a method to consolidate a variety of temporal-spatial data. This method aims at circulation of such temporal-spatial data. CSIS generates detailed individual pseudo mobility-trace data from JICA's temporal-spatial statistics data and provides them named People Flow Data (PFLOW)[8]. JICA's temporal-spatial statistics data is called Person Trip Survey (PT Survey)[6]. The PT Survey is one of the travel survey like the NTHS. At the moment, the PFLOW includes unified grain size data of 26 Japanese cities and 4 Asian cities. Each PFLOW data consists of many individual detailed trace data (a sequence of every minute's location information) for 24 hours, and the amount of traces is, for example, 600,000 for Tokyo 2008. Researchers can access PFLOW free after applied to CSIS[9].

A trip of a conventional PT data consists of the origin and destination location, its hour and its time distance. In each trip of PFLOW, the minutely intermediate points of the trip are estimated and complimented although assuming its mobility as linear uniform motion on the shortest path between the origin and destination. PFLOW has more detail information than the usual travel survey data, but the pathways of minutely geographical position will not be enough for VANET simulations with high-speed vehicles over 30km/h. To use a traffic flow simulator like the related work[1][2] is reasonable for simulating minutely behavior of vehicles on the second time scale. However, such datasets as PFLOW which include less actual traffic volume may not consider traffic jam in the traffic simulator. In this framework, we set up parameters to the traffic flow simulator for considering traffic jam.

As an application of VANET, we evaluate mobile data offloading through Wi-Fi spots by using the proposed framework. Opportunistic Wi-Fi offloading simulations which use moving vehicles have been researched[10]. Moving vehicles ferry mobile data to Wi-Fi spot. This enables the enlargement of Wi-Fi area. In this paper, we simulate such Wi-Fi offloading system based on the proposed framework of the mobility trace. The performance of the Wi-Fi offloading system depends on moving vehicle's encounter frequency and connection time on Wi-Fi spots on the simulation. We use actual Wi-Fi spot map from NTT DoCoMo , the Japanese major cellular phone carrier[13].

2 RELATED WORK

2.1 Synthetic Mobility Models

It is useful to synthesize mobility trace data both from survey data and observed data as a way to presume OD matrix. In this way, the OD matrix can satisfy real-world traffic volume. There are some researches that generate mobility trace from the synthetic mobility models. In these approaches, fine-graded OD matrix is synthesized by a traffic simulator. Two examples about the synthetic mobility model are described as follows.

First, datasets in Luxembourg are used to define each area's weight coefficient[1]. The datasets are about the use of the land map, for example, commercial, residential, industrial and downtown. The amount of the traffic from/to outside of the area is estimated by a roadside measurement unit.

Second, the simulation in Cairn is used three datasets[2]. The datasets are (i) home locations and sociodemographic characteristics, (ii) working places and free-time activities take place, and (iii) the time use patterns. These datasets conjecture 1.2 million ODs satisfied real-world traffic volume in an area of approximately 400 km^2 around the urban agglomeration. In the research, an allocation method of pathway is argued to feed the huge OD matrix into the traffic flow simulator.

An advantage of such synthesized OD matrix is potentially of topping up trip scenarios if the base datasets do not represent the original amount of trips. On the other hand, mock traffic jams are necessary to simulate individual vehicles realistically. To use a real-world traffic volume is a typical method to generate simulated traffic jams. Since a simple pathway interpolation, e.g. the shortest path, of each vehicle's may cause too heavy traffic at specific intersections on simulation, the pathway assignment of each OD trip is important. Road topology on simulators and that of real-world have often different capacity of traffic volume. Accordingly, to guarantee the reality of the pathway is difficult in this method. Furthermore, the same set of datasets are needed if applying these approaches to another city.

In this paper, we generate detailed mobility trace data just from travel survey data. This data are available in a lot of cities around the world.

2.2 Mobility Traces Based On A Travel Surveys

Methods to generate mobility traces based on travel surveys have been researched. For example, a mobility trace is generated from the surveyed OD matrix in Zurich[3]. This OD matrix is surveyed by Swiss regional planning authority. It includes 24-hour zone-level OD matrix. The zone-level OD is not coordinate-OD but summarized-OD in each zone, hence it is impossible to obtain accurate OD points in the zone. This zone-level OD matrix is given into a traffic simulator to output a mobility trace. This simulation has simple road topologies of the queue, and moreover only arterial roadways are covered. In this simulation, 65,000 km^2 area and 260,000 vehicles are included, and however, the amount of the data is less than the original traffic because the data is based on the traffic survey.

On the other hand, individual coordinates of the OD form travel surveys are based on true values, in consequence, the OD's margin of error is smaller than their zone size. That is why the OD matrix from a travel survey has higher accuracy than synthetic mobility models. Travel time is also surveyed in many travel surveys. The travel time enables quantitative evaluations of mobility trace through comparisons with real-world travel time.

Travel survey data are suited for generating mobility traces around the world, because it is published in a lot of cities. This trace generation approach based on travel surveys needs a travel time evaluation, because the survey data includes less traffic volume data and enough traffic jam cannot be considered.

2.3 People Flow Data (PFLOW)

PFLOW[8] is developed from Person Trip Survey (PT Survey)[6], one of the travel survey. OD matrix and travel time are included in the travel survey data. PT Survey is surveyed by Japan local government and Japan International Cooperation Agency. CSIS (the center for Spatial Information Science, the University of Tokyo) aims at circulation of various temporal-spatial data through PFLOW. PT Survey data is the main source of PFLOW. Individual trips in the flow are generated from statistics of survey data.

PFLOW uses population-delimited zone-level of OD matrix, 15,000 peoples on each population group. In PFLOW, the origin point and destination point of a flow is remapped based on the actual distributions of buildings and population of each of the zones. Spatial smoothing in the geocoding is also conducted in order to obtain finer-graded ODs. Each trip of PFLOW is appended to intermediate points as its detailed path between its origin to destination, and these points are calculated considering the mode of transportation and making its path length reasonable. Such pathways are represented by a sequence of coordinates on a minute-by-minute basis.

Departure time and arrival time are surveyed in the each trip of PT Survey. However, survey subjects usually answer rough departure time and arrival time. In the PFLOW, such departure time and arrival time are smoothed.

Advantages of PFLOW are (i) realistic included coordinate-level OD matrix , (ii) unified grain size data of 30 cites in 5 Asian countries, (iii) an application procedure permits access to PFLOW, and (iv) easy cropping due to coordinate-level. Currently, we can access 26 Japanese metropolitan datasets and 4 Asian datasets (Hanoi, Manila, Jakarta and Dhaka).

2.4 Opportunistic Wi-Fi Offloading Which Use Moving Vehicles

Opportunistic Wi-Fi offloading which use moving vehicles is often simulated by the use of various datasets and methods.

In terms of network simulation, a Wi-Fi network model with M/G/1/K queue is used[10]. In this research, trade-off between network ability and Wi-Fi offloading ability is simulated. The network model is realistic, and real-world map information is used in the simulation. However, Wi-Fi spots are deployed randomly, and unrealistic mobility trace which a traffic flow simulator generates automatically is used.

The simulation for 83km^2 of San Francisco shows that we can offloading half of mobile data through the opportunistic Wi-Fi offloading[11]. In the simulation, a real-world mobility trace of taxis is used. Wi-Fi spots are deployed along the main roads adequately.

On the other hand, effective placement of Wi-Fi spots is researched to improve the opportunistic Wi-Fi offloading[12]. In the simulation with traced taxi dataset, the effectiveness of this placement was verified.

As mentioned above, some network models and frameworks are proposed for the opportunistic Wi-Fi offloading simulation. However, there are some cases where unrealistic mobility traces or qualifying datasets like taxi are used. In this paper, we propose the mobility trace to combine it with these network simulations.

3 MOBILITY TRACE GENERATION METHOD FOR VANET SIMULATION

In this section, we mention disadvantages of zone-level OD matrix. Then, the complement method is introduced to conquer the disadvantages. In the proposed framework, we remove all arbitrary area off from PFLOW and selected two area from Tokyo for evaluation. Figure 1 shows the workflow for generating method of mobility trace.

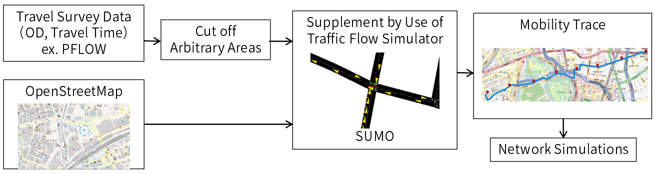


Figure 1: Workflow for Generating Method of Mobility Traces

3.1 Mobility Trace

3.1.1 Prerequisite of Mobility Trace for VANET Simulation

Mobility traces for VANET simulation need to be expressed by moving trajectory of every second based on the actual travel. The performance evaluation of VANET application depends on acceleration and deceleration caused by traffic light waiting and traffic jam. The traffic frequency also has an impact on the performance evaluation. Therefore, mobility traces for VANET simulation need to be based on the actual physical behavior of real-world vehicles. Vehicles usually run several ten meters per second. On the other hand, communication range of VANET is about 100m, so mobility traces need to be expressed by moving trajectory at least of every second.

3.1.2 Disadvantages of Zone-Level OD Matrix

No one can find the true coordinate-level OD points by using zone-level OD matrix. Owing to this, assumed coordinate-level OD points based on zone-level OD matrix have a small margin of error. The longest error and diameter of the zone are same length, since trips between close two zones especially have larger error than usual. This is the common problem of zone-level OD matrix surveyed by travel surveys which is not included travel distance. PFLOW even has such the common problem.

3.1.3 PFLOW Dataset in This Evaluation

In this paper, our framework targets for at PT survey datasets of various cities. PFLOW is more convenient than usual travel survey data and we can access datasets of 30 Asian cities. PFLOW is as versatile as usual travel survey data. Furthermore, it is easy to input PFLOW to the traffic flow simulator, because PFLOW has coordinates of OD and middle path, so we use PFLOW in this paper. PFLOW dataset in this evaluation is from PT Survey data in Tokyo on October 1, 2008. Survey slips were sent to 16,000,000 households. 340,000 households (600,000 people) provided a reply.

A PFLOW dataset covers an extensive area. However, the VANET simulation with such extensive area has huge calculation costs. We cut selected areas in Tokyo (i.e. A downtown and a suburb, 5km square areas) from the PFLOW. The downtown area is *Chiyoda* and the suburb area is *Tachikawa*. The two selected areas are fed to SUMO.

3.2 A Supplement Method of PFLOW by Use of SUMO

CSIS developed the PT survey into easily usable PFLOW. PFLOW's pathways represented by coordinates on a minute-by-minute basis are still rough to simulate VANET applications. In this paper, we make such pathways finer. Traffic simulator SUMO outputs detail pathways represented by coordinates on a second-by-second basis. It is necessary to apply the method which we can hold the reality of travel time of travel survey data.

3.2.1 Simulation Models of SUMO in This Evaluation

In this paper, we feed a free road topology of OpenStreetMap (OSM)[14] into the SUMO. In this simulation, the reality of the road topology and traffic lights depend on capacity of SUMO and OSM. The number of lanes are optimized by SUMO based on tagging road scales in OSM data. These optimized lanes are represented by edges including a pair of node i.e. a starting point and ending point of a short line. In the OSM road topology, the position of traffic lights are given as identical with real-world positions. By contrast, the lights switching are different from real-world, because OSM does not include such changeover timing data. Only one type of general vehicles is used in this simulation and it does not run on the expressway.

3.2.2 Input Method for PFLOW to SUMO

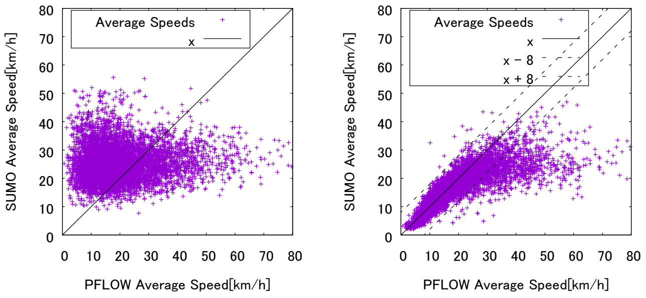
The one method of how to generate a trip scenario is to assign OD and fragmentary middle path. SUMO can calculate the shortest path from assigned information and Dijkstra's algorithm. We assigned the closest pair of edge to PFLOW's OD as the trip scenario's OD. The closest edges to PFLOW's intermediate points was assigned as the trip scenario's fragmentary middle path. This is how to input PFLOW to SUMO.

3.2.3 Velocity Adjustment

PFLOW includes smaller traffic volume than that of real-world, so we cannot consider traffic jams on the SUMO by simple use of PFLOW.

Travel times of PFLOW are based on true values, because they are generated from surveyed PT Survey Data. In addition, most travel distances and average speeds are authoritative, because the average travel distance (17.9 km) is much longer than the general zone size of PT Survey (1km^2-1.5km^2). On the other hand, SUMO simulation with PFLOW dataset and OSM legal speed cannot simulate traffic jams, because PFLOW dataset is smaller than real-world traffic. But for average speed, the simulated speed is faster than PFLOW. Figure 2(a) shows the normal supplemented result of average speeds based on OSM legal speed in downtown Tokyo. The result of the suburb area was similar to the downtown area.

Then, SUMO is re-executed by using speed factors. Those are computed from the first output of SUMO trips. When average speed of the first SUMO trips is $\overline{v_s}$ and average speed of the PFLOW trips is $\overline{v_p}$, a speed factor is $\frac{\overline{v_p}}{\overline{v_s}}$. The maximum value of speed factor is 80 km/h and has potential to excess OSM regal speed.



(a) Before Velocity Adjustment (b) After Velocity Adjustment

Figure 2: The Average Speed Comparisons between PFLOW and Simulated Trip by SUMO (Downtown Tokyo)

3.3 A Supplement Result by Use of SUMO

In each trips, We evaluate coincidence average speed between SUMO and PFLOW. The substantial reason for this evaluation is what survey basis on PFLOW expresses real-world traffic jams. In the comparing between re-executed result and PFLOW, 90% of trips have speed difference within 8km/h. Figure 2(b) shows the coincidence by using the speed factor.

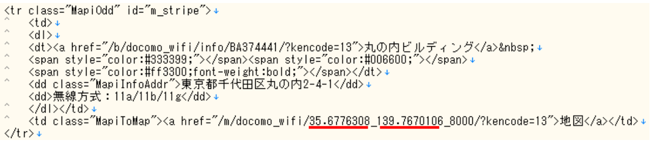
Conversely, 10% of trips data are incorrect by PFLOW's unrealistic fast average speed. Those are common problem of assumed ODs based on zone-level OD matrix.

3.4 Consideration

In this paper, we complemented finer grain of mobility trace based on PFLOW by use of traffic simulator SUMO. PFLOW is a realistic data based on questionnaire survey. SUMO's output is a data which is moving trajectory of every second. It is considered acceleration and deceleration caused by vehicle's waiting for traffic light. In this paper, we also considered vehicle's waiting for caused by traffic jam. Therefore, our proposal method satisfies the requirements of mobility trace for VANET simulation shown at section 3.1.1.

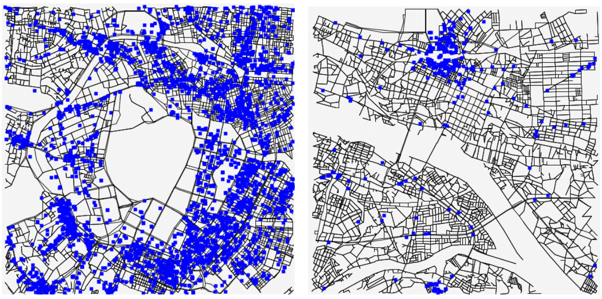
4 AN OPPORTUNISTIC WI-FI OFFLOADING SIMULATION WHICH USE RUNNING VEHICLES

In this section, we evaluate the Wi-Fi offloading system which uses running vehicles. Running vehicle's encounter frequency and connection time of Wi-Fi spot are evaluated by using the proposed mobility trace and a Wi-Fi spot location data. It is necessary to consider detailed network models like packet collision and radio signal propagation model to simulate exact VANET communication. By contrast, we use simple network model with the aim to simulate Wi-Fi offloading statistically. Simulation areas are identical with the evaluation of the generating mobility trace (i.e. The downtown and suburb of Tokyo, 5km square areas).



```
<tr class="MapiOdd" id="m_stripe">↓
^   <td>↓
^   <dl>↓
^   <dt><a href="/b/docomo_wifi/info/BA374441/?kencode=13">丸の内ビルディング</a>&nbsp;↓
^   <span style="color:#333399;"></span><span style="color:#006600;"></span>↓
^   <span style="color:#ff3300;font-weight:bold;"></span></dt>↓
^   <dd class="MapiInfoAddr">東京都千代田区丸の内2-4-1</dd>↓
    <dd>無線方式：11a/11b/11g</dd>↓
^   </dl></td>↓
^   <td class="MapiToMap"><a href="/m/docomo_wifi/35.6776308_139.7670106_8000/?kencode=13">地図</a></td>↓
</tr>↓
```

Figure 3: NTT DoCoMo Part of Document about Wi-Fi Locations



(a) The Downtown Tokyo

(b) The Suburb Tokyo

Figure 4: The Drawing of Extracted Wi-Fi Spots

4.1 The Wi-Fi Offloading Simulation Models

4.1.1 Real-world Wi-Fi Offloading Spot Location Data

NTT DoCoMo, the Japanese major cellular phone carrier has released Wi-Fi spot locations on their website[13]. Figure 3 shows the part of website document. The underline of Fig. 3 shows lat/long by use of Japanese Geodetic Datum 2000 (JGD2000). We got all Wi-Fi spot location data of Tokyo from the website. We extracted all Wi-Fi locations from website and convert the data location from JGD2000 to WGS84 format for used with OSM and PFLOW. The use of only one carrier dataset is justified, because users cannot access Wi-Fi spots installed by other carriers which is not contracted.

Figure 4 shows extracted Wi-Fi spots. There are 2988 spots in the downtown simulation area, and there are 324 spots in the suburb simulation area.

4.1.2 Communication Models

In this evaluation, we use simplified communication models to simulate the upper bound of opportunistic Wi-Fi offloading ability. Wi-Fi communication areas are concentric circles with a radius r in meters with no barriers. Running vehicles connect to the nearest Wi-Fi spot preferentially. There is no no packet collision. The communication probability in Wi-Fi areas is 100%. The negotiation time to get IP address is 0 seconds. The maximum number of concurrent connections at a Wi-Fi spot is not subject to restraint. Communication requests occurs when vehicles outbound from Wi-Fi areas or start to travel. In this simulation, it is no problem whether to complete sending packets or not. However, this offloading system targets at small packets like mail data, synchronous data and system update data. It is possible to apply DTN (Delay Tolerant Networking) to these communication.

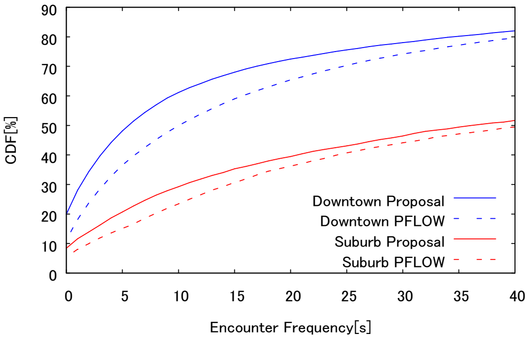


Figure 5: Cumulative Probabilities of The Encounter Frequency（r = 40）

Table 1: Average Encounter Frequency And Connection Time

Data		90% CDF Encounter Frequency	Encounter Frequency	Connection Time
Downtown	Proposal	71s	26.0s	23.6s
	PFLOW	79s	32.0s	22.2s
Suburb	Proposal	291s	107.1s	24.2s
	PFLOW	300s	110.1s	19.4s

4.1.3 Evaluation Items

There are two items conceivable in the Wi-Fi offloading evaluation. One is running vehicle's encounter frequency of Wi-Fi spot. The other is running vehicle's Wi-Fi spot connection time. The send-able packet size depends on Wi-Fi spot connection time. Kinds of send-able content depend on the encounter frequency, because of using DTN and size of send-able packets rely on the connection time of Wi-Fi spot. In the evaluation, we compare proposed mobility trace with normal use of PFLOW through the Wi-Fi offloading simulation.

4.2 Evaluation Result

In general, Wi-Fi spots are estimated to work in 100m radius of communication area without barriers. However, the actuality of Wi-Fi signal is attenuate in long distance. In this comparisons, we set Wi-Fi radius r to 40m as a significant value of parameters .

Figure 5 shows cumulative probabilities of the encounter frequency of two mobility traces. Table 1 shows average encounter frequency and average connection time. The result of using proposed mobility trace has about 10% higher than original PFLOW in case of connection time cumulative probability in downtown Tokyo.

4.3 Consideration

In the framework of generating mobility trace, average speed of SUMO and PFLOW were adjusted. However, the encounter frequency of use of proposed mobility trace is shorter than normal PFLOW (Fig. 5). Generally, Wi-Fi spots are installed in crowded places like mass transport stations. That is why, locations of traffic jams and Wi-Fi spots located at the same position. In contrast, vehicles run speedily on non-dense Wi-Fi areas, hence encounter frequencies between two Wi-Fi spots become shorter than the PFLOW's motion of uniform velocity.

The difference of the connection time between proposed mobility trace and normal PFLOW is explained by same reason. Vehicles run slowly on tight Wi-Fi areas, hence connection times at a Wi-Fi spot become longer than than the PFLOW's motion of uniform velocity.

5 CONCLUSION

In this paper, we proposed the framework of achievement to generate mobility trace from public travel survey data. This Generating of mobility trace aims at vehicle's network simulation. People Flow Data (PFLOW) generated by the Center for Spatial Information Science, The University of Tokyo. It was used in framework evaluation. For this framework, traffic jams were reconstructed based on compensated speed factors of SUMO and adjusted to be average speeds of PFLOW.

In addition, we simulated the opportunistic Wi-Fi offloading system by using proposed mobility trace and original PFLOW. Running vehicles ferry mobile data to Wi-Fi spot for the offloading. The vehicle's encounter frequency and connection time were evaluated in this simulation. In simulation result, there was a difference between proposed mobility trace and normal PFLOW. The necessity of proposed mobility trace was confirmed through evaluation result.

In the downtown Tokyo, the average encounter frequency was 26 seconds and the average connection time was 24 seconds. Accordingly, the data which has delay tolerance like mail data, synchronous data and system update data is capable through this Wi-Fi offloading system.

In this framework, only one PFLOW dataset from private vehicles was simulated. There are improvement of reality by simulation of buses, motorcycles, bicycles and pedestrians in the future task.

Parenthetically, we used very simple network models in the evaluation of Wi-Fi offloading system. Network simulators such as ns-3 has ability to improve the Wi-Fi offloading simulation.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 26280028 (Grant-in-Aid for Scientific Research (B)) and 26330102 (Grant-in-Aid for Scientific Research (C)).

REFERENCES

[1] Y. Pigné, G. Danoy and P. Bouvry, "A Vehicular Mobility Model Based on Real Traffic Counting Data," *Communication Technologies for Vehicles*, pp. 131–142, 2011.

[2] S. Uppoor, O. Trullols-Cruces, M. Fiore, J.M. Barcelo-Ordinas, "Generation and Analysis of a Large-scale Urban Vehicular Mobility Dataset," *Mobile Computing, IEEE Transactions*, vol. 13, no. 5, pp. 1061–1075, 2014.

[3] B. Raney, N. Cetin, A. Vollmy, M. Vrtic, K. Axhausen and K. Nagel, "An Agent-Based Microsimulation Model of Swiss Travel: First Results," *Networks and Spatial Economics*, vol. 3, no. 1, pp. 23–41, 2003.

[4] M. Violland, *Travel/mobility surveys: some key findings*, `http://www.internationaltransportforum.org/statistics/StatPapers/SP201102.pdf`, Organisation for Economic Co-operation and Development, (accessed 2016-02-16).

[5] A. Santos, N. McGuckin, H.Y. Nakamoto, D, Gray, S. Liss, *Summary of Travel Trends: 2009 National Household Travel Survey*, `http://nhts.ornl.gov/2009/pub/stt.pdf`, NHTS, (accessed 2016-02-16).

[6] Tokyo Metropolitan Area Transportation Planning Council, *What is Person Trip Survey? (in Japanese)*, `http://www.tokyo-pt.jp/person/index.html`, (accessed 2015-09-09).

[7] M. Behrisch, L. Bieker, J. Erdmann and D. Krajzewicz, "SUMO - Simulation of Urban MObility ," *The Third International Conference on Advances in System Simulation*, 2011.

[8] Y. Sekimoto, R. Shibasaki, H. Kanasugi, T. Usui and Y. Shimazaki, "PFLOW: Reconstruction of people flow recycling large-scale social survey data," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 27–35, 2011.

[9] Center for Spatial Information Science The University of Tokyo, *Application Procedure* `http://pflow.csis.u-tokyo.ac.jp/?page_id=924` (accessed 2015-12-19).

[10] N. Cheng, N. Lu, N. Zhang, X.S. Shen and J.W. Mark, "Opportunistic WiFi offloading in vehicular environment: A queueing analysis," *Global Communications Conference (GLOBECOM), 2014 IEEE*, pp. 211–216, 2014.

[11] S. Dimatteo, P, Hui, B. Han and Li. V.O.K., "Cellular Traffic Offloading through WiFi Networks," *Mobile Adhoc and Sensor Systems*, pp. 192–201, 2011.

[12] B. Eyuphan, S. K. Boleslaw., "WiFi Access Point Deployment for Efficient Mobile Data Offloading," *Proceedings of the first ACM international workshop on Practical issues and applications in next generation wireless networks*, pp. 45–50, 2012.

[13] NTT DoCoMo, *Wi-Fi Area Search (in Japanese)*, `http://sasp.mapion.co.jp/b/docomo_wifi/`, (accessed 2015-05-21)

[14] M. Haklay and P. Weber, "OpenStreetMap: User-Generated Street Maps," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 12–18, 2008.

Keynote Speech 3:

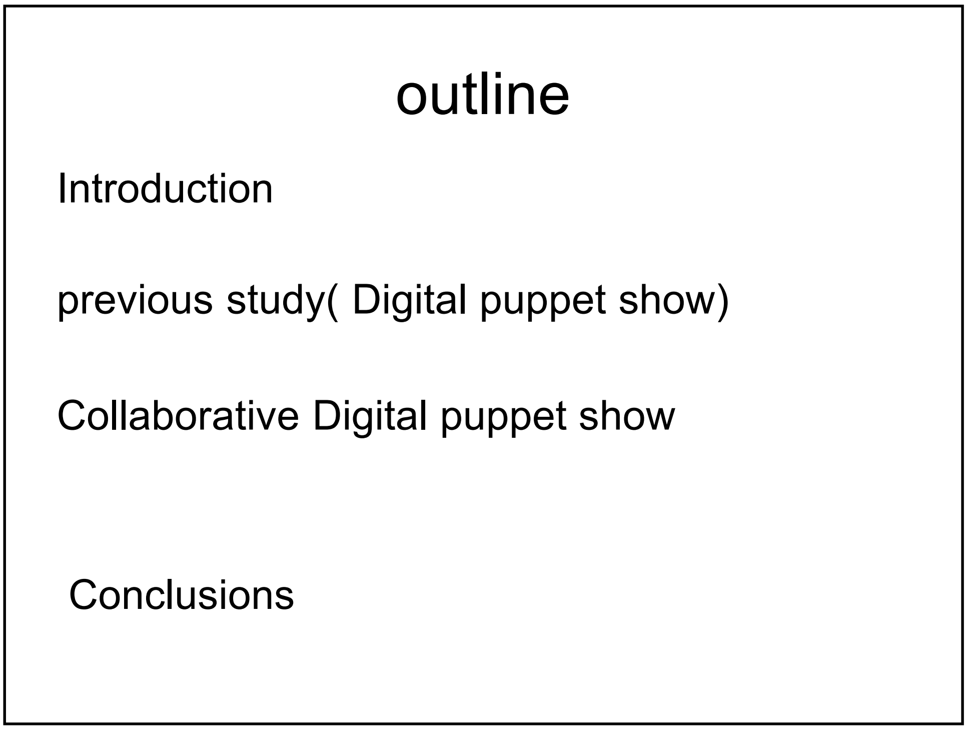
Prof. Fusako Kusunoki

(Tama Art University)

The Design with Information Support Features for the Children with Hearing Disability

Tama art university
Information
design

Fusako Kusunoki



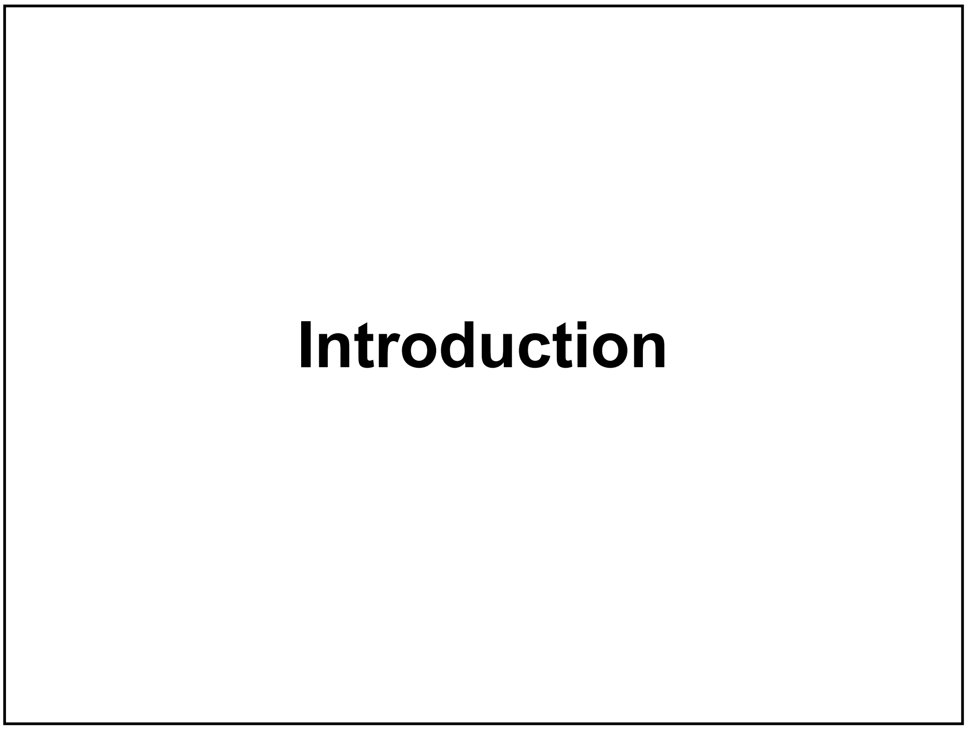
outline

Introduction

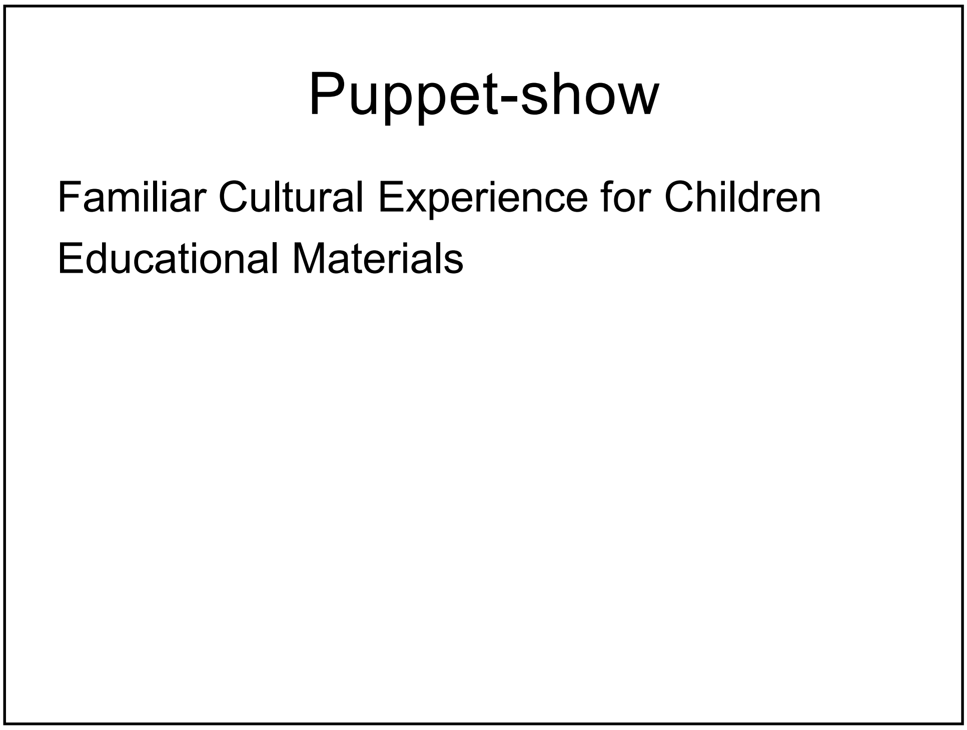
previous study(Digital puppet show)

Collaborative Digital puppet show

Conclusions

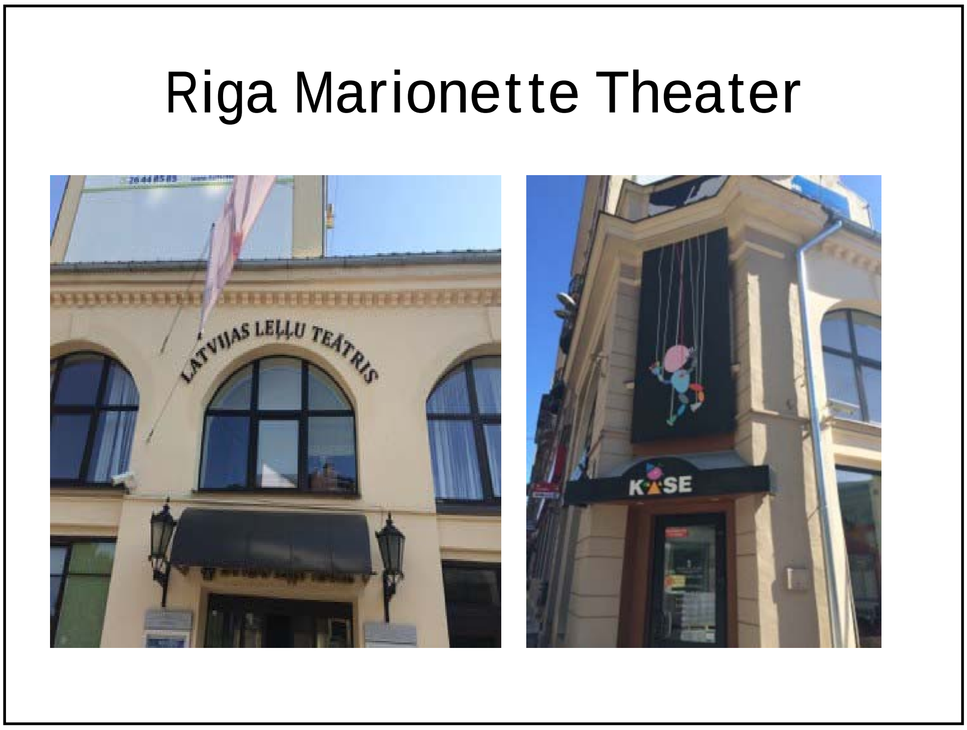


Introduction



Puppet-show

Familiar Cultural Experience for Children
Educational Materials



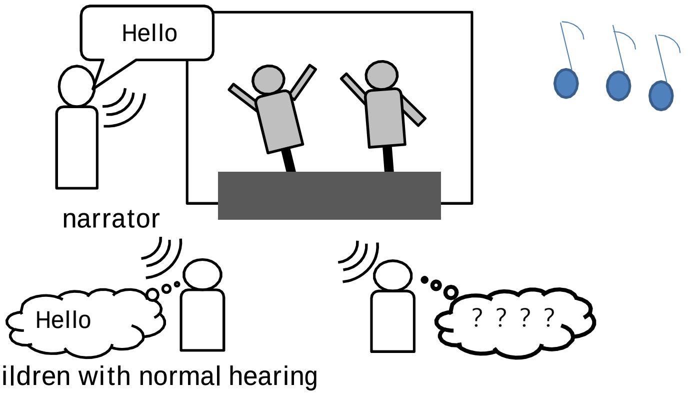
Riga Marionette Theater

Characters

Program (August)

Normal Puppet-Show

Puppet shows are a very familiar cultural experience



the audio information is main

It is difficult for deaf children to enjoy

Problems

Normal Puppet-Show

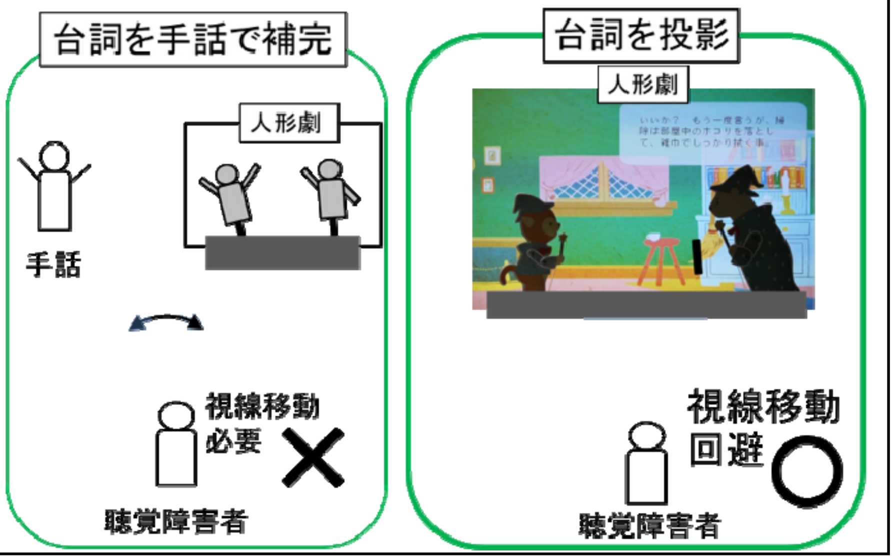
- The Dialogue Information

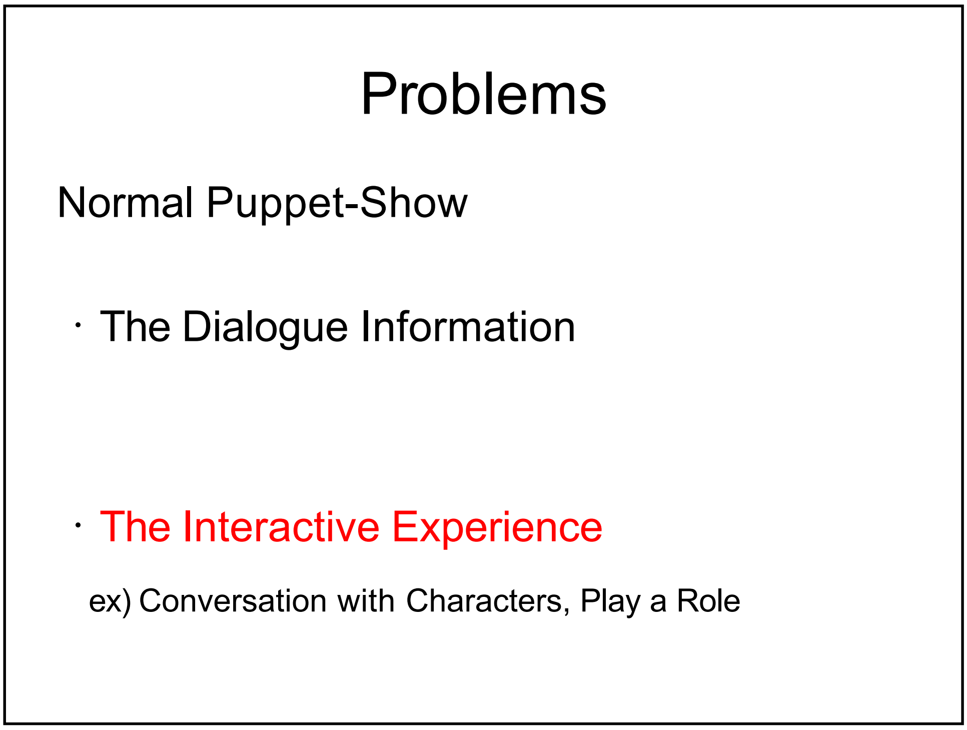
- The Interactive Experience

ex) Conversation with Characters

1. The Dialogue Information

Support the contents of the puppet theater for the deaf
→It is important to get the dialogue content in real time





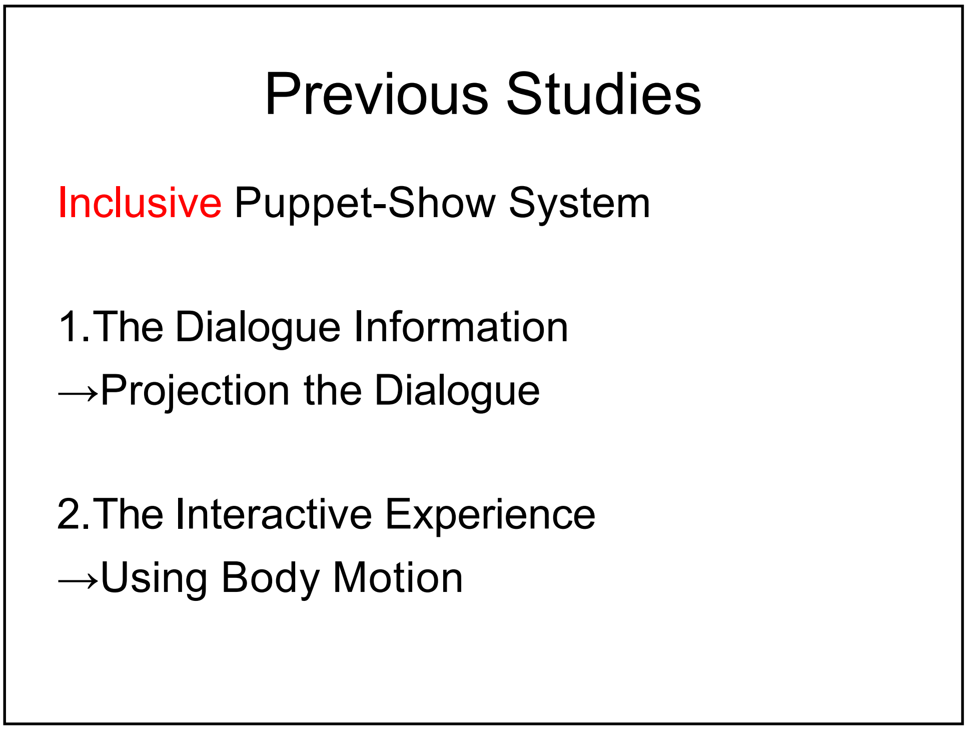
Problems

Normal Puppet-Show

・The Dialogue Information

・The Interactive Experience

ex) Conversation with Characters, Play a Role



Previous Studies

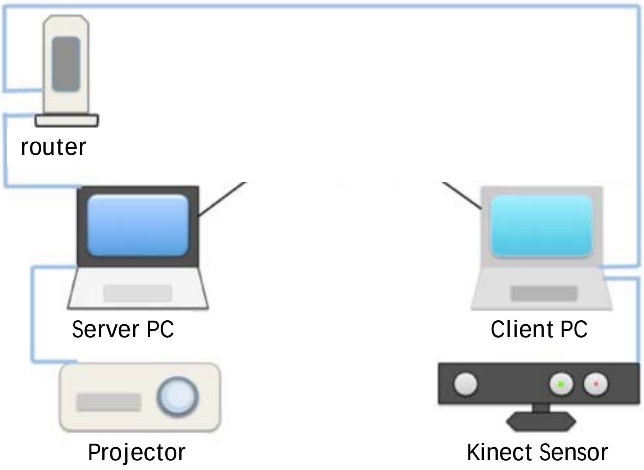
Inclusive Puppet-Show System

1.The Dialogue Information
→Projection the Dialogue

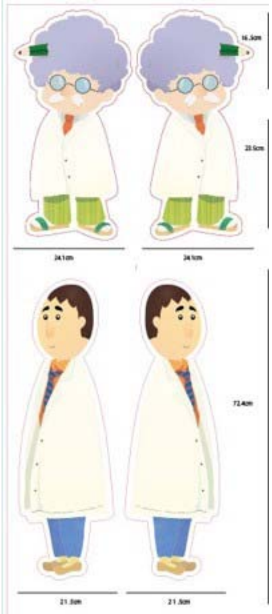
2.The Interactive Experience
→Using Body Motion

Digital Puppet-Show

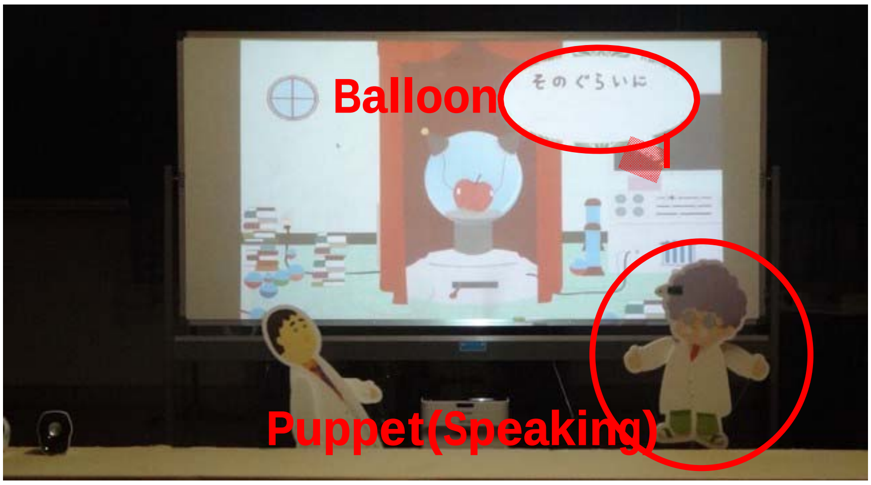
System Framework

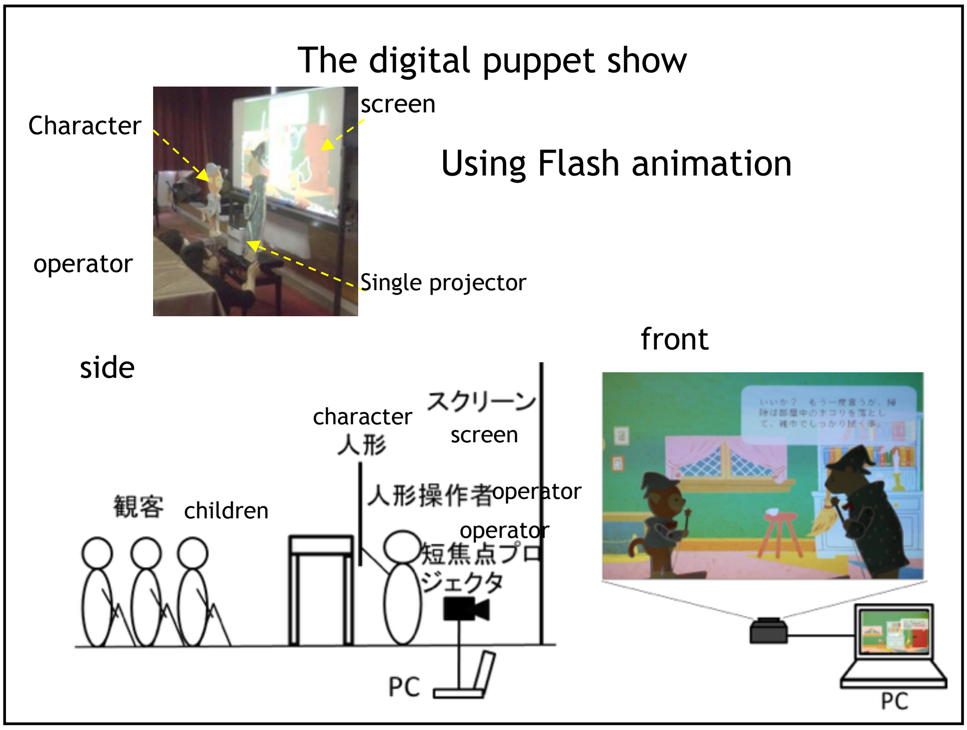


Puppet



The characters’ dialogues





The digital puppet show

Using Flash animation

Previous Studies

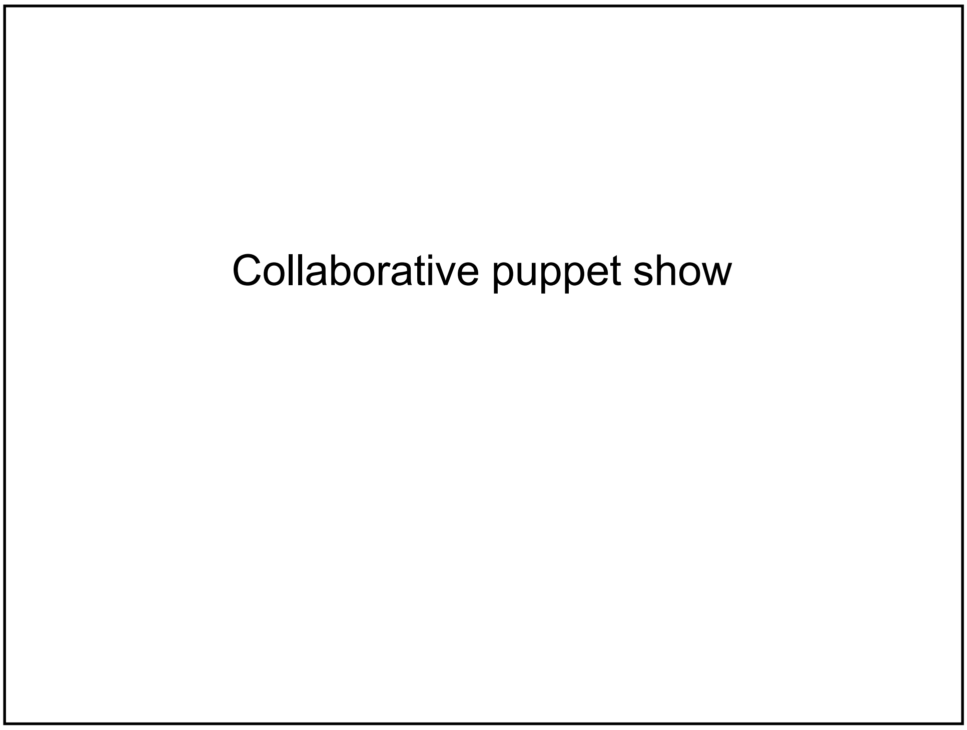
・The Interactive Experience

→Using Body Motion

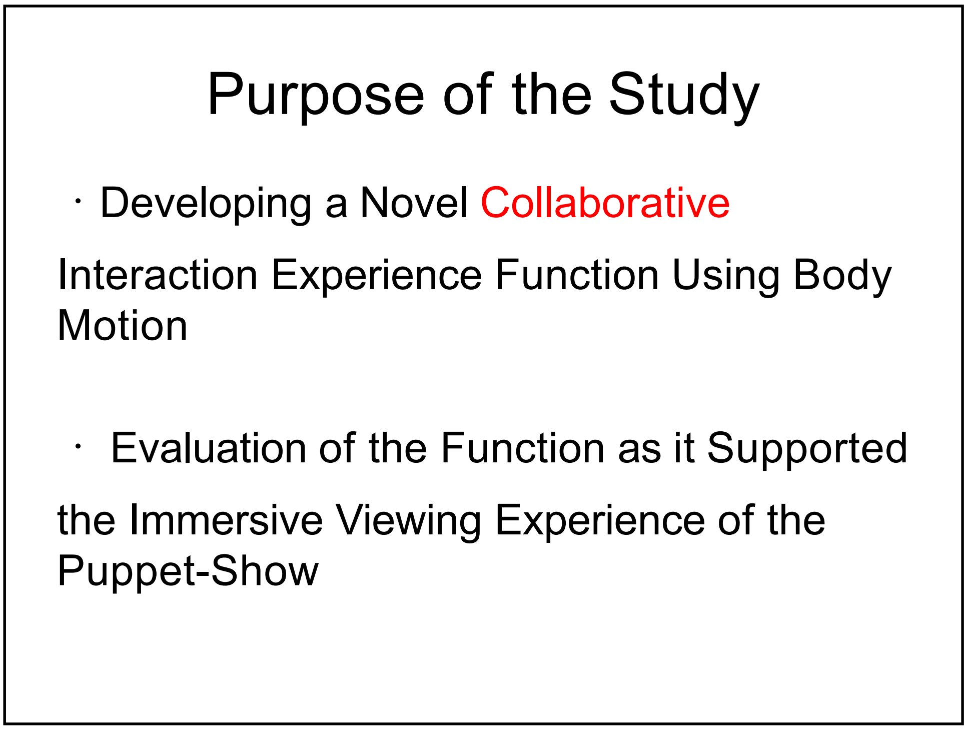
Previous Studies

demo

- video



Collaborative puppet show



Purpose of the Study

・Developing a Novel Collaborative Interaction Experience Function Using Body Motion

・ Evaluation of the Function as it Supported the Immersive Viewing Experience of the Puppet-Show

Story

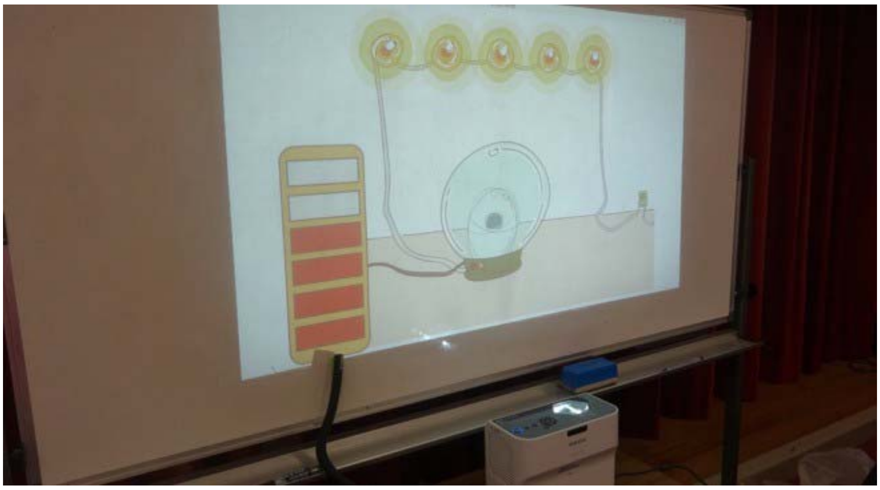
Setting: In Laboratory

Purpose: Repairing Broken Robot

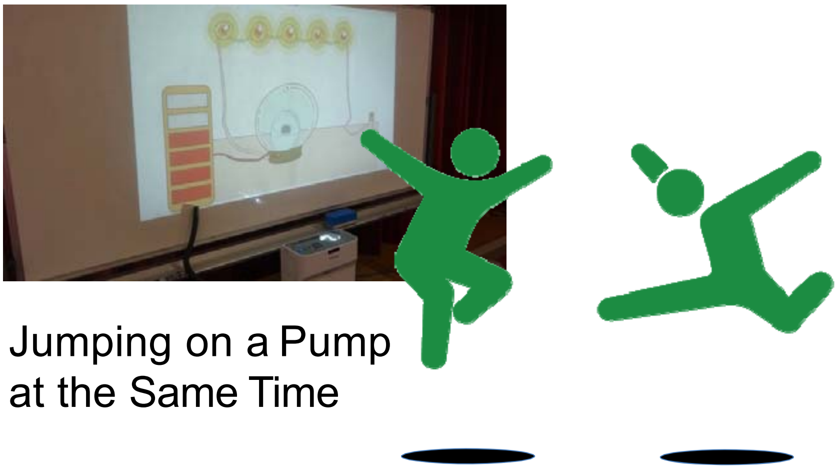
Collaborative Interaction Experience:

Two Interactive Game Like Experiences in the System's Puppet- Show as Problems

1.The Electricity Generation Game

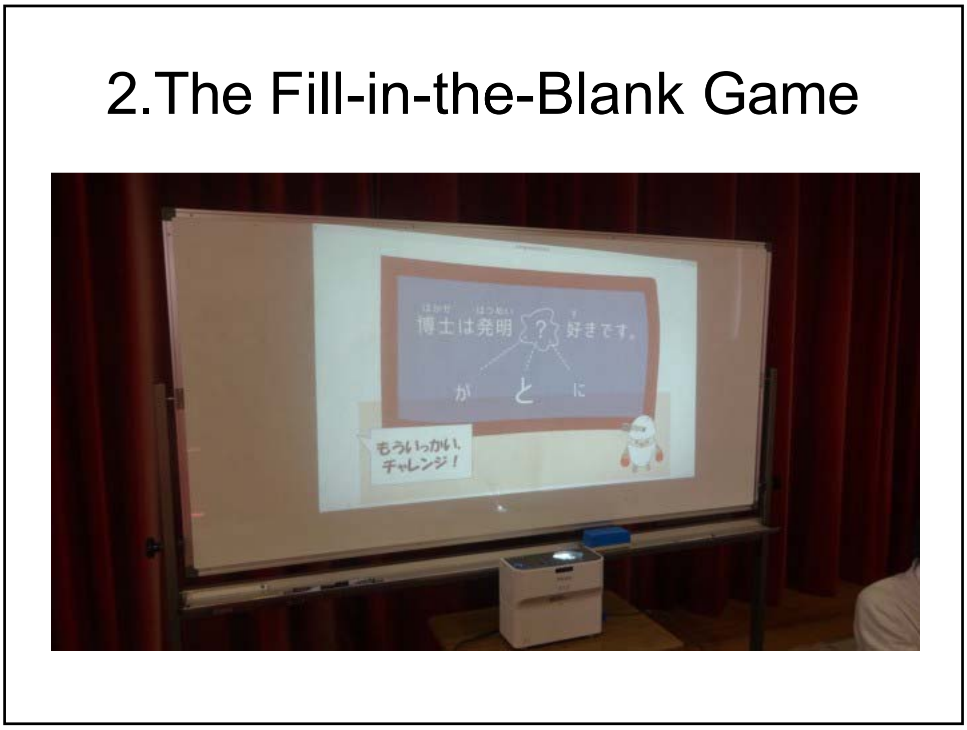


The Electricity Generation Game

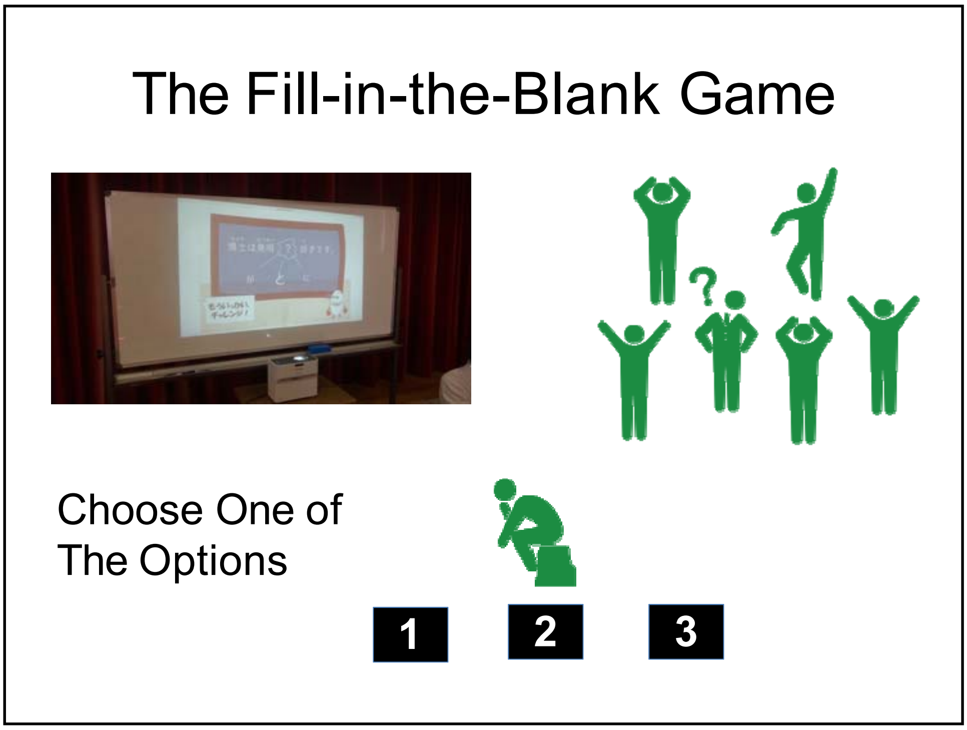


Jumping on a Pump
at the Same Time

- video



2.The Fill-in-the-Blank Game



The Fill-in-the-Blank Game

Choose One of
The Options

The Fill-in-the-Blank Game

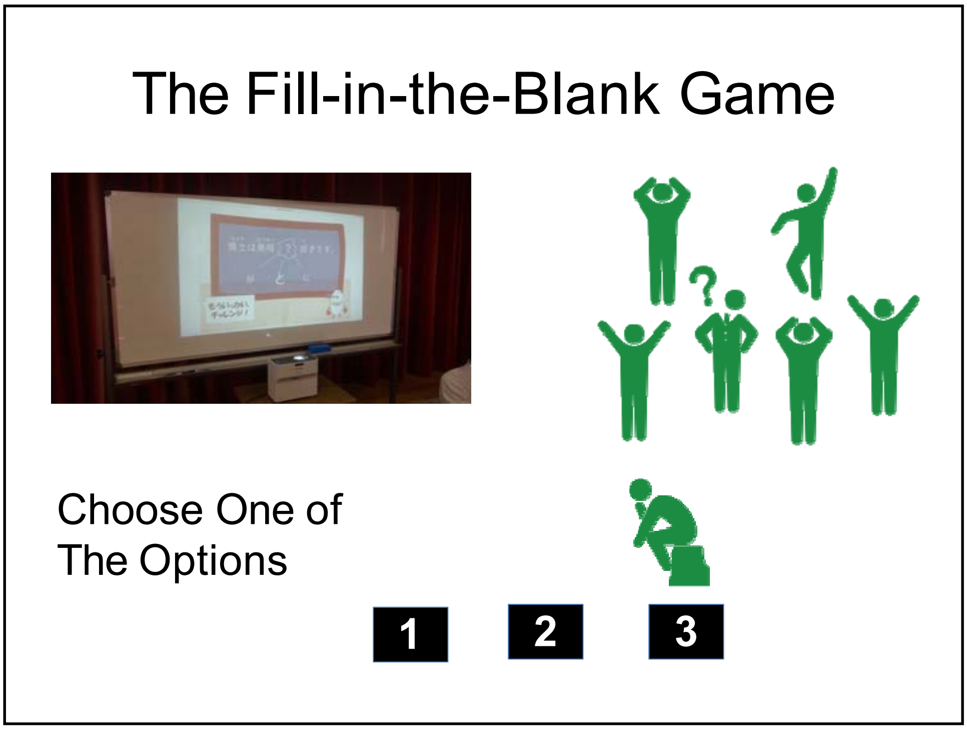
Choose One of
The Options

1 **2** **3**

The Fill-in-the-Blank Game

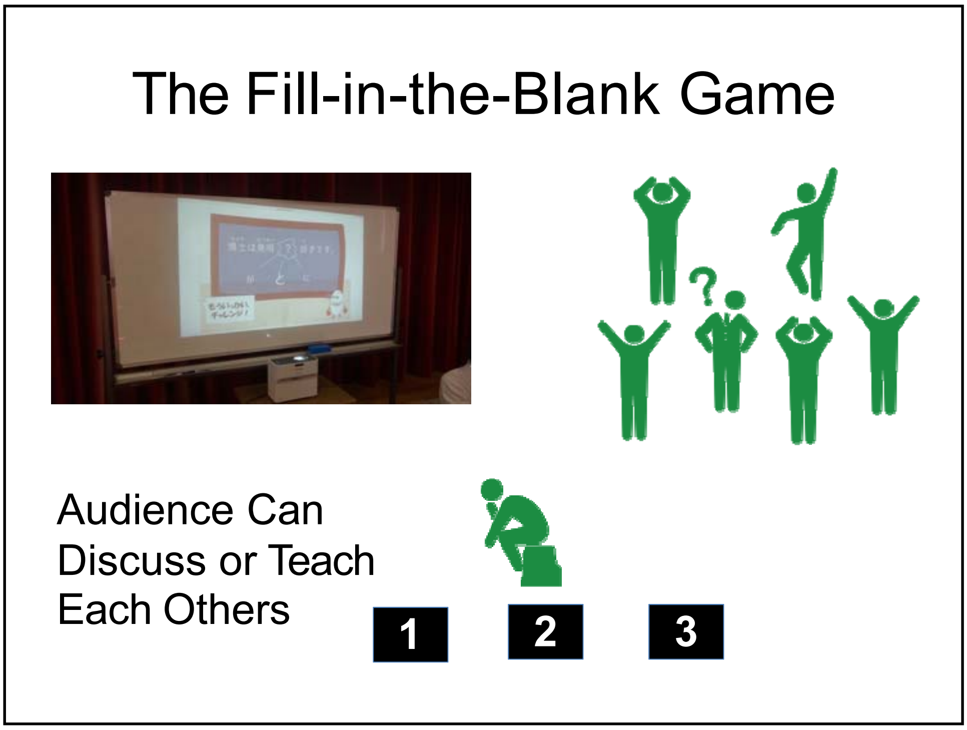
Choose One of
The Options

1 **2** **3**



The Fill-in-the-Blank Game

Choose One of
The Options



The Fill-in-the-Blank Game

Audience Can
Discuss or Teach
Each Others

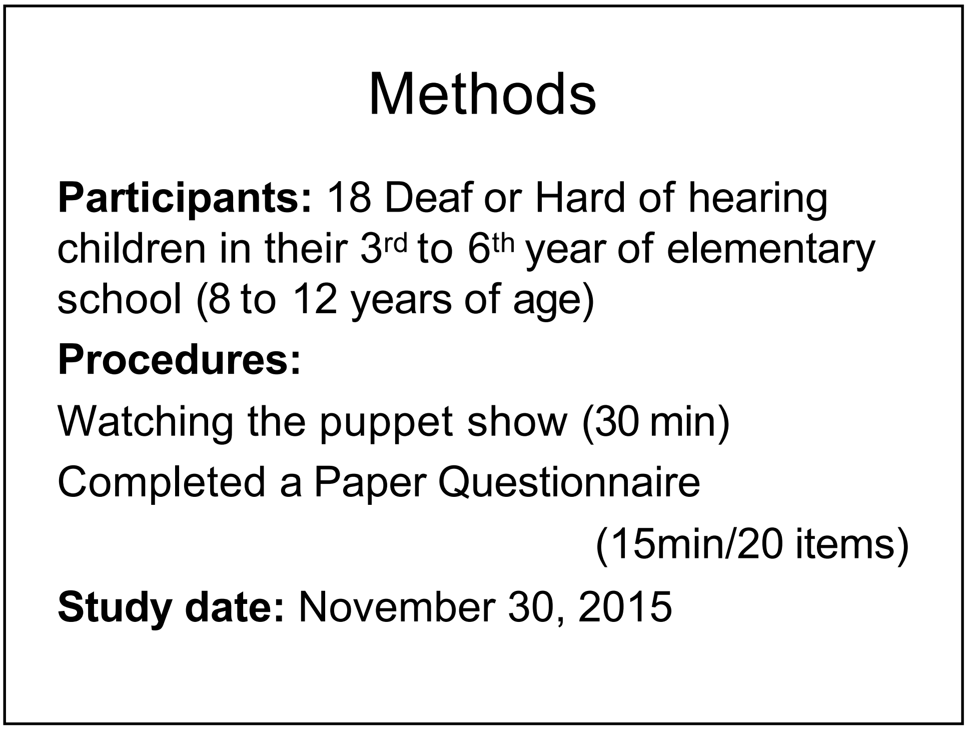
The Fill-in-the-Blank Game

Jumping in Front of the Selection

1 2 3

- video2

Evaluation Experiment



Methods

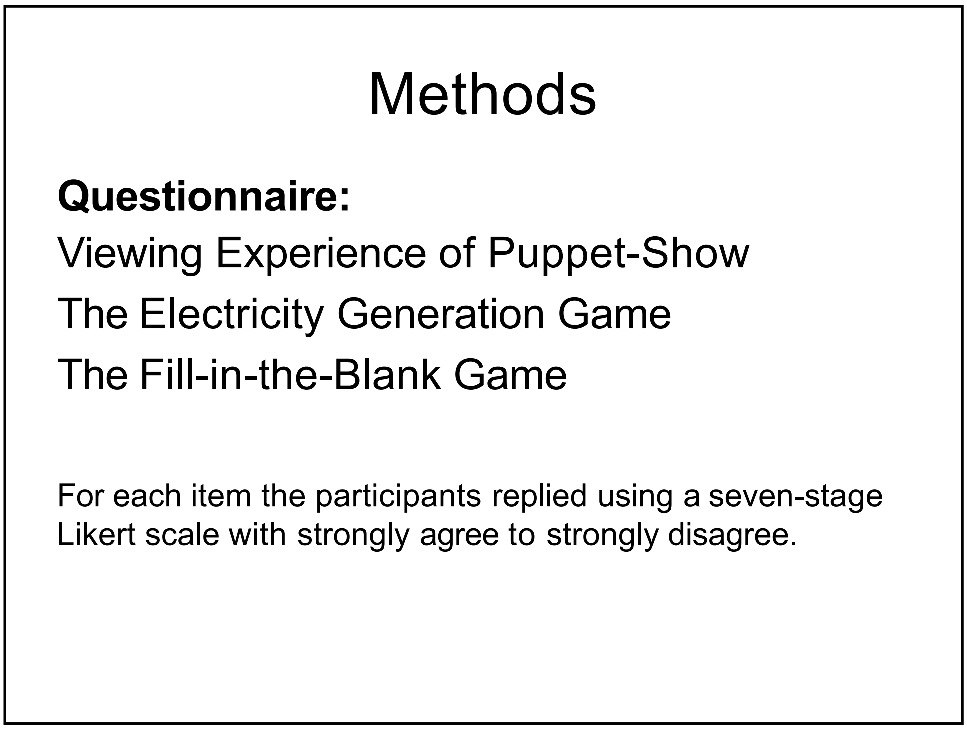
Participants: 18 Deaf or Hard of hearing children in their 3rd to 6th year of elementary school (8 to 12 years of age)

Procedures:

Watching the puppet show (30 min)

Completed a Paper Questionnaire (15min/20 items)

Study date: November 30, 2015



Methods

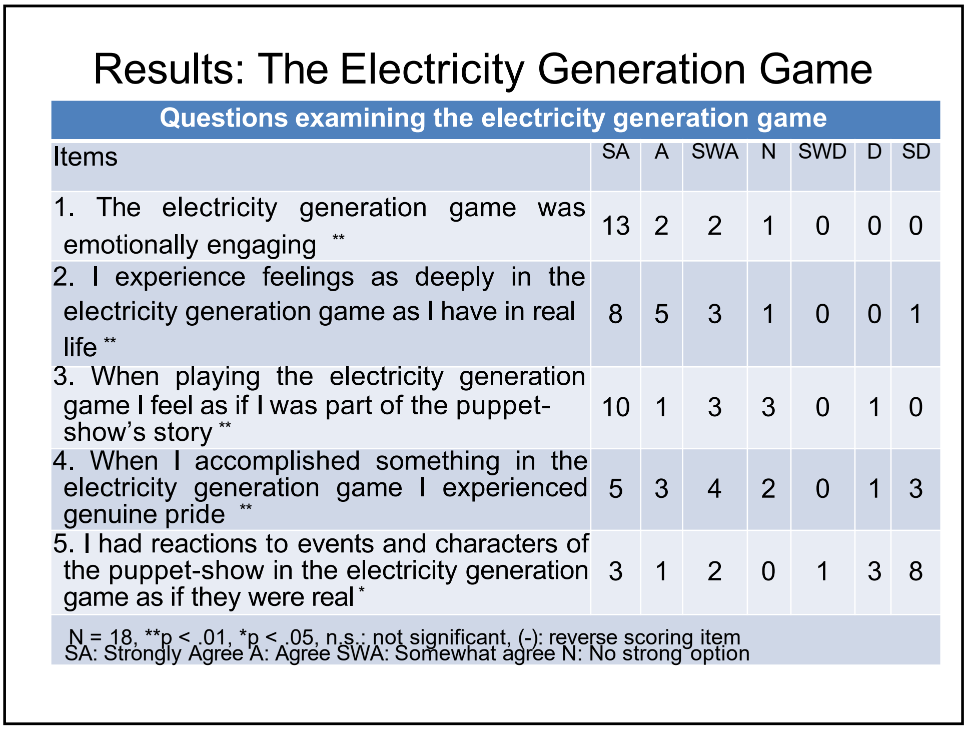
Questionnaire:

Viewing Experience of Puppet-Show

The Electricity Generation Game

The Fill-in-the-Blank Game

For each item the participants replied using a seven-stage Likert scale with strongly agree to strongly disagree.

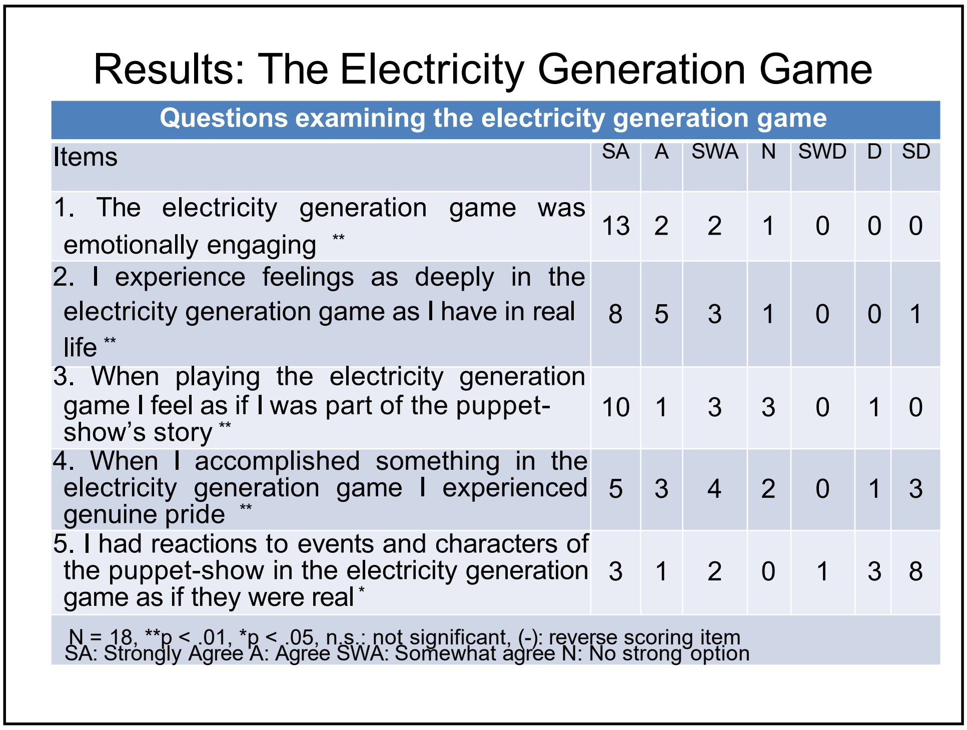


Results: The Electricity Generation Game

Questions examining the electricity generation game

Items	SA	A	SWA	N	SWD	D	SD
1. The electricity generation game was emotionally engaging **	13	2	2	1	0	0	0
2. I experience feelings as deeply in the electricity generation game as I have in real life **	8	5	3	1	0	0	1
3. When playing the electricity generation game I feel as if I was part of the puppet-show's story **	10	1	3	3	0	1	0
4. When I accomplished something in the electricity generation game I experienced genuine pride **	5	3	4	2	0	1	3
5. I had reactions to events and characters of the puppet-show in the electricity generation game as if they were real *	3	1	2	0	1	3	8

N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option

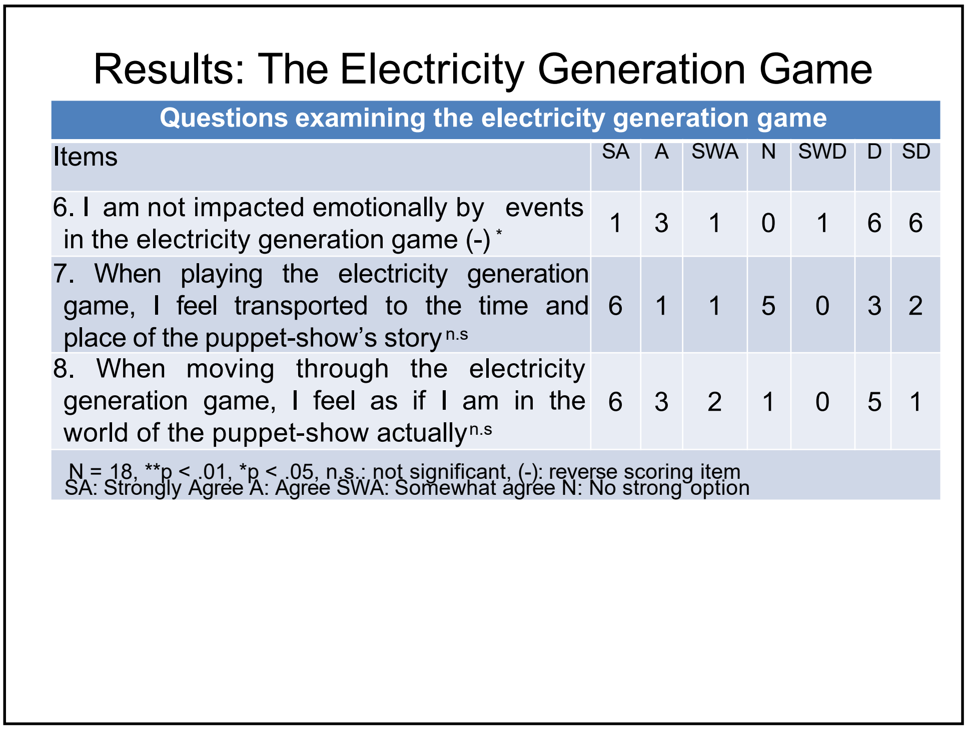


Results: The Electricity Generation Game

Questions examining the electricity generation game

Items	SA	A	SWA	N	SWD	D	SD
1. The electricity generation game was emotionally engaging **	13	2	2	1	0	0	0
2. I experience feelings as deeply in the electricity generation game as I have in real life **	8	5	3	1	0	0	1
3. When playing the electricity generation game I feel as if I was part of the puppet-show's story **	10	1	3	3	0	1	0
4. When I accomplished something in the electricity generation game I experienced genuine pride **	5	3	4	2	0	1	3
5. I had reactions to events and characters of the puppet-show in the electricity generation game as if they were real *	3	1	2	0	1	3	8

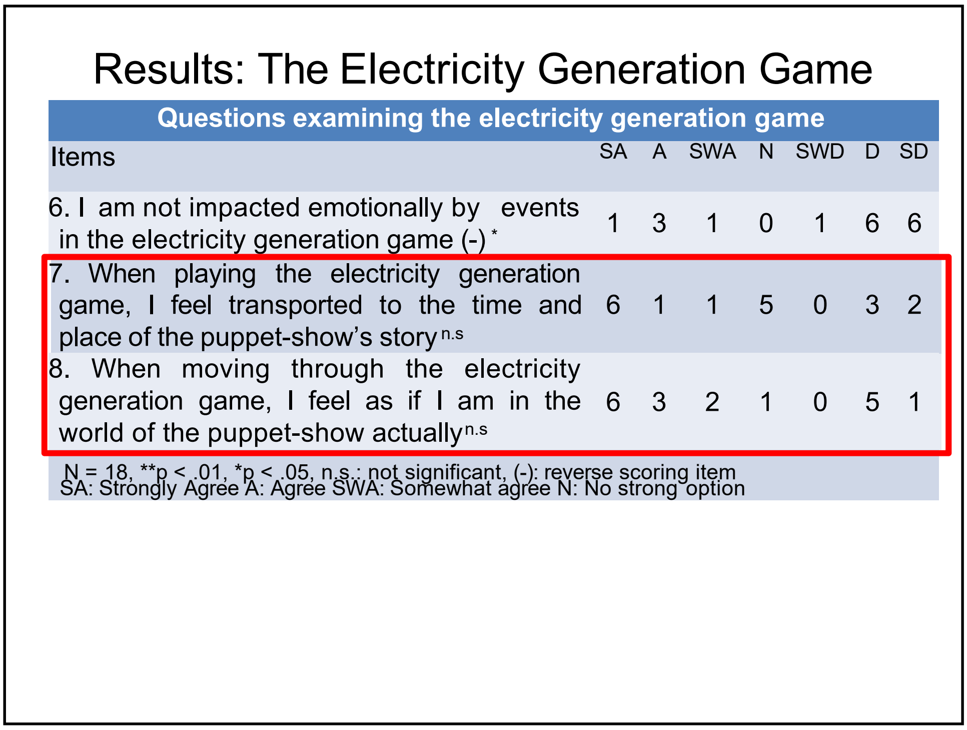
N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option



Results: The Electricity Generation Game

Questions examining the electricity generation game							
Items	SA	A	SWA	N	SWD	D	SD
6. I am not impacted emotionally by events in the electricity generation game (-) *	1	3	1	0	1	6	6
7. When playing the electricity generation game, I feel transported to the time and place of the puppet-show's story n.s	6	1	1	5	0	3	2
8. When moving through the electricity generation game, I feel as if I am in the world of the puppet-show actually n.s	6	3	2	1	0	5	1

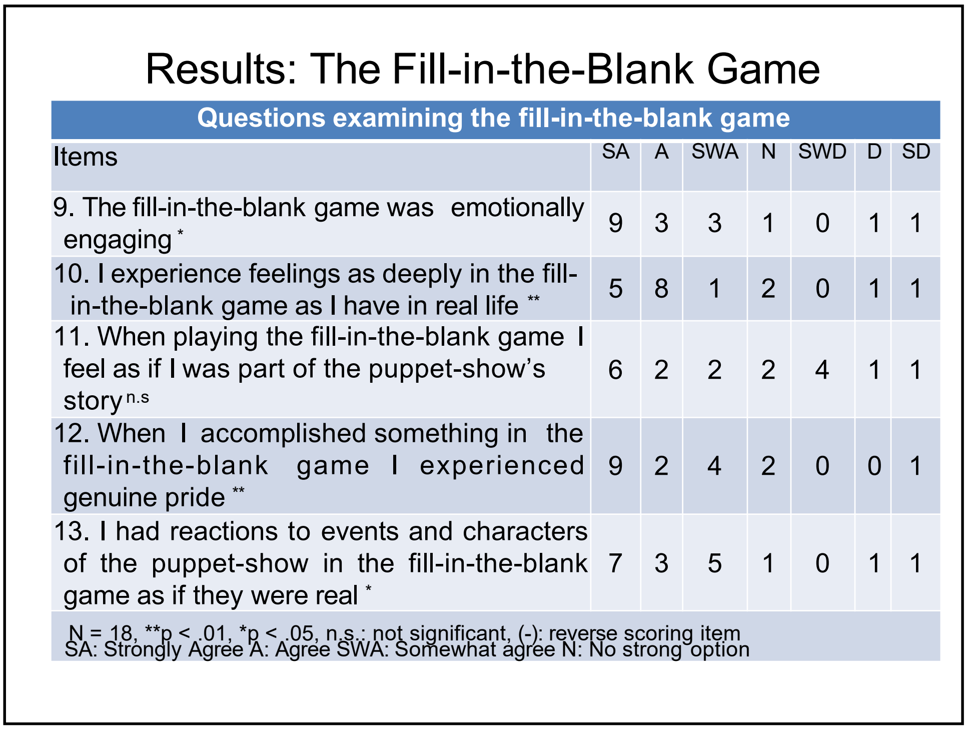
N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option



Results: The Electricity Generation Game

Questions examining the electricity generation game							
Items	SA	A	SWA	N	SWD	D	SD
6. I am not impacted emotionally by events in the electricity generation game (-) *	1	3	1	0	1	6	6
7. When playing the electricity generation game, I feel transported to the time and place of the puppet-show's story n.s	6	1	1	5	0	3	2
8. When moving through the electricity generation game, I feel as if I am in the world of the puppet-show actually n.s	6	3	2	1	0	5	1

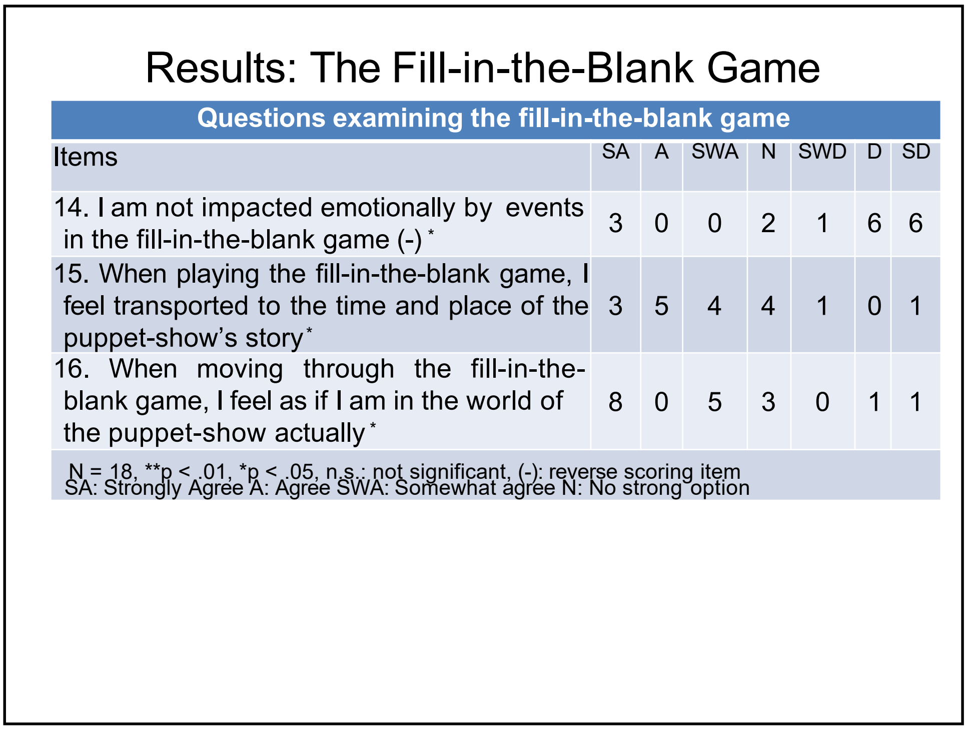
N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option



Results: The Fill-in-the-Blank Game

Questions examining the fill-in-the-blank game							
Items	SA	A	SWA	N	SWD	D	SD
9. The fill-in-the-blank game was emotionally engaging *	9	3	3	1	0	1	1
10. I experience feelings as deeply in the fill-in-the-blank game as I have in real life **	5	8	1	2	0	1	1
11. When playing the fill-in-the-blank game I feel as if I was part of the puppet-show's story n.s	6	2	2	2	4	1	1
12. When I accomplished something in the fill-in-the-blank game I experienced genuine pride **	9	2	4	2	0	0	1
13. I had reactions to events and characters of the puppet-show in the fill-in-the-blank game as if they were real *	7	3	5	1	0	1	1

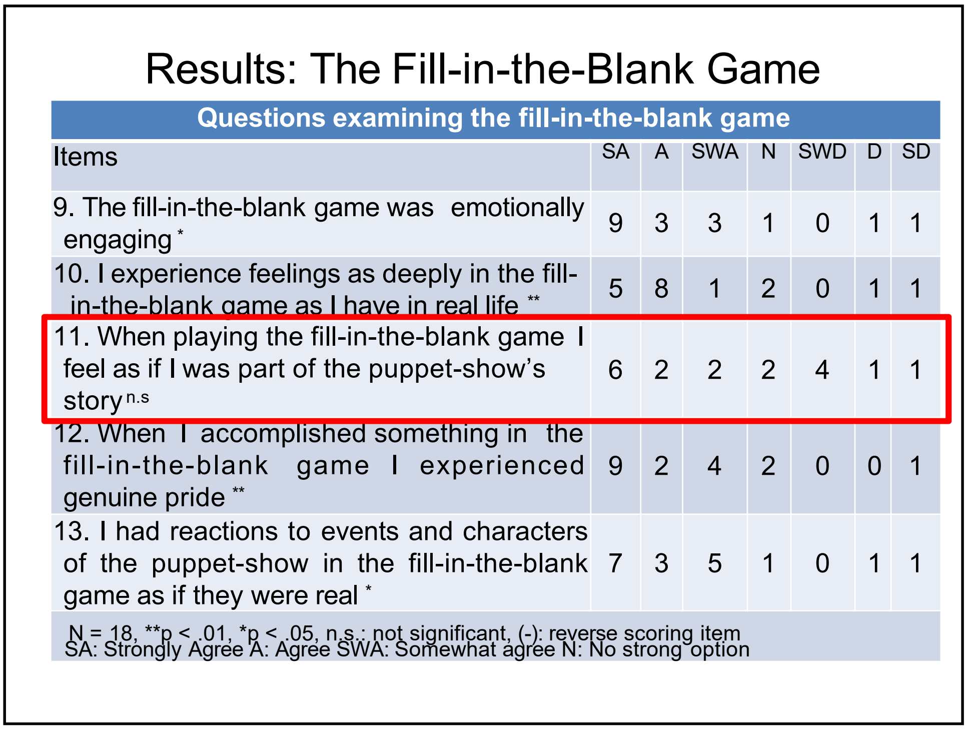
N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option



Results: The Fill-in-the-Blank Game

Questions examining the fill-in-the-blank game							
Items	SA	A	SWA	N	SWD	D	SD
14. I am not impacted emotionally by events in the fill-in-the-blank game (-) *	3	0	0	2	1	6	6
15. When playing the fill-in-the-blank game, I feel transported to the time and place of the puppet-show's story *	3	5	4	4	1	0	1
16. When moving through the fill-in-the-blank game, I feel as if I am in the world of the puppet-show actually *	8	0	5	3	0	1	1

N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option

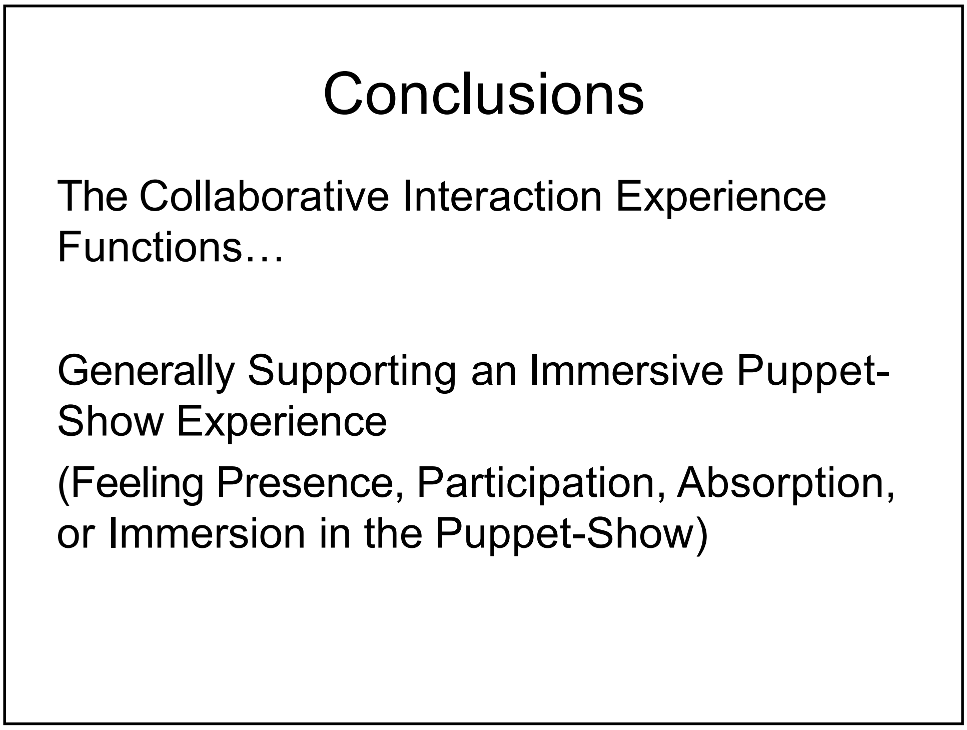


Results: The Fill-in-the-Blank Game

Questions examining the fill-in-the-blank game							
Items	SA	A	SWA	N	SWD	D	SD
9. The fill-in-the-blank game was emotionally engaging *	9	3	3	1	0	1	1
10. I experience feelings as deeply in the fill-in-the-blank game as I have in real life **	5	8	1	2	0	1	1
11. When playing the fill-in-the-blank game I feel as if I was part of the puppet-show's story $^{n.s}$	6	2	2	2	4	1	1
12. When I accomplished something in the fill-in-the-blank game I experienced genuine pride **	9	2	4	2	0	0	1
13. I had reactions to events and characters of the puppet-show in the fill-in-the-blank game as if they were real *	7	3	5	1	0	1	1

N = 18, **p < .01, *p < .05, n.s.: not significant, (-): reverse scoring item
SA: Strongly Agree A: Agree SWA: Somewhat agree N: No strong option

Conclusions

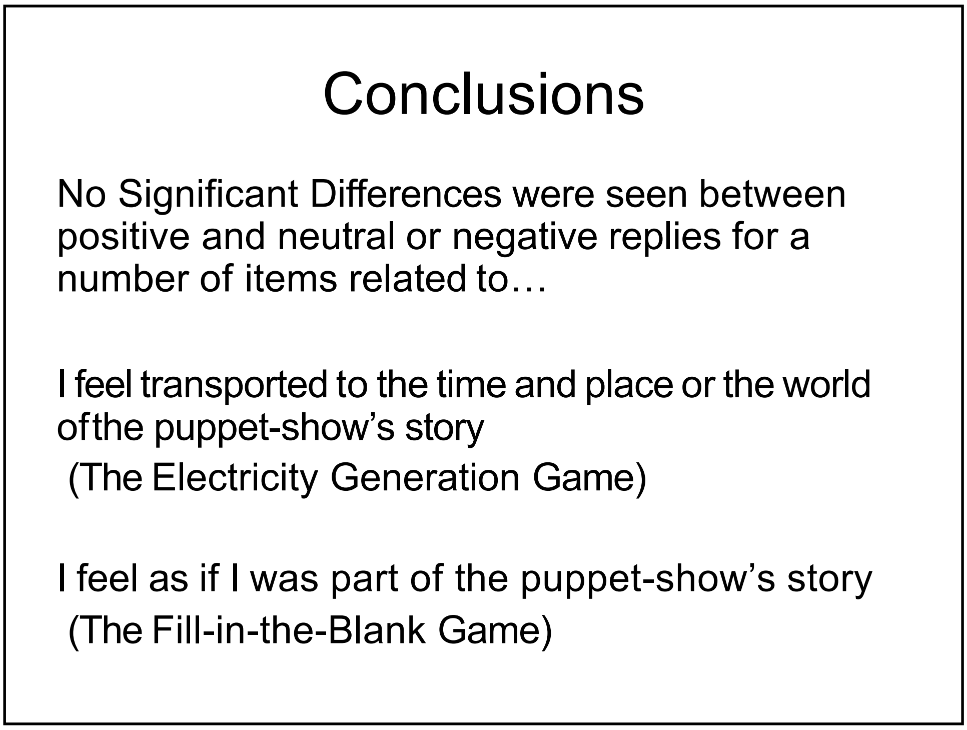


Conclusions

The Collaborative Interaction Experience Functions…

Generally Supporting an Immersive Puppet-Show Experience

(Feeling Presence, Participation, Absorption, or Immersion in the Puppet-Show)



Conclusions

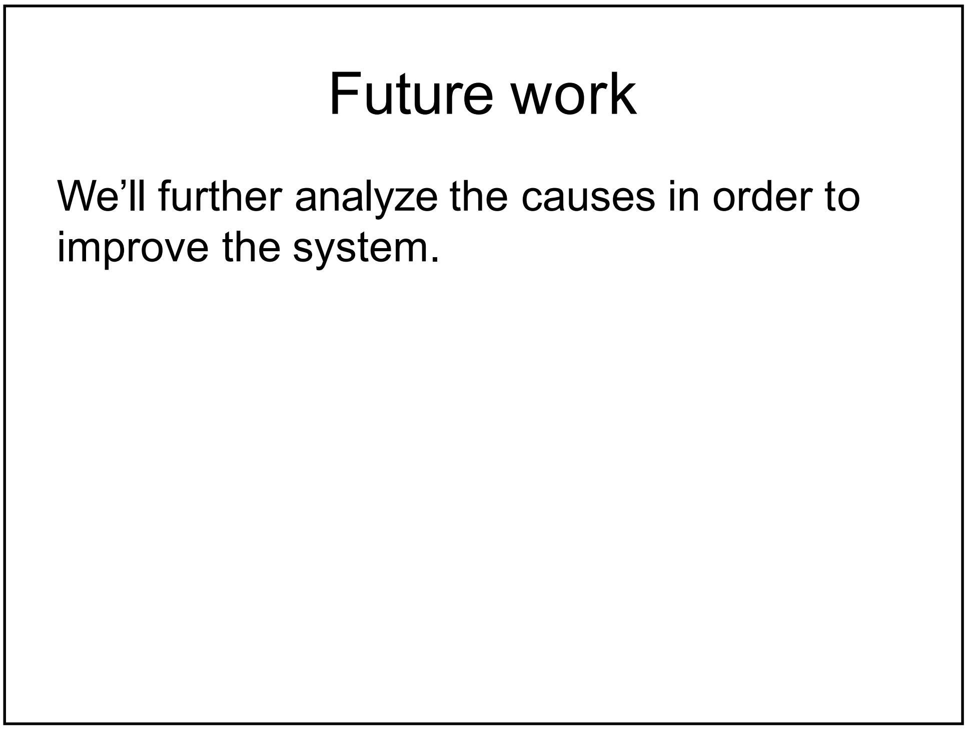
No Significant Differences were seen between positive and neutral or negative replies for a number of items related to…

I feel transported to the time and place or the world ofthe puppet-show's story

(The Electricity Generation Game)

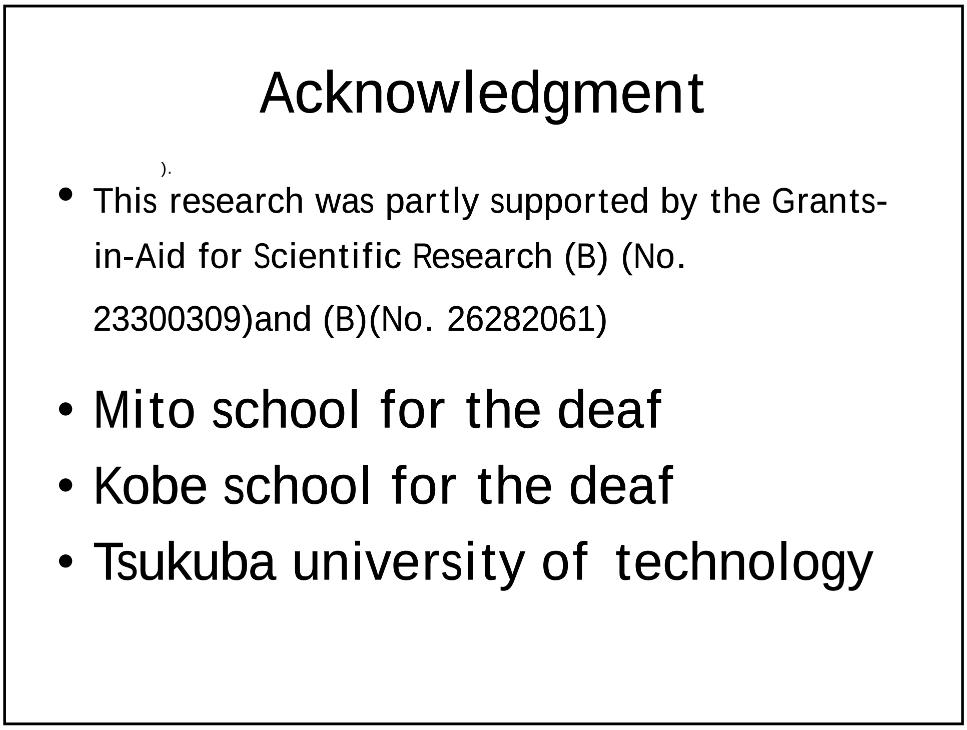
I feel as if I was part of the puppet-show's story

(The Fill-in-the-Blank Game)



Future work

We'll further analyze the causes in order to improve the system.



Acknowledgment

).

- This research was partly supported by the Grants-in-Aid for Scientific Research (B) (No. 23300309)and (B)(No. 26282061)
- Mito school for the deaf
- Kobe school for the deaf
- Tsukuba university of technology

THANK YOU!

Session 4:
Robots Application

(Chair: Yoshia Saito)

A Study on Scalability Analysis of Exploration with Micro-Robots for Search in Rubble

Yuki Koizumi, Minoru Harada, and Toru Hasegawa

Graduate School of Information Science and Technology
{ykoizumi, m-harada, t-hasegawa}@ist.osaka-u.ac.jp

Abstract - To utilize a huge number of small robots, called *micro-robots*, for survivor searches in rubble in disaster areas is considered as a promising approach because of their smallness. How many micro-robots should be deployed is one of crucial research issues for survivor searches with micro-robots. In this paper, we derive theoretical lower bounds of the number of required micro-robots for accomplishing their search mission by modeling rubble as a graph and drawing orbits of micro-robots as paths on the modeled graph. As the first step to analyze the number of required micro-robots, we focus on relations between the number and sizes of rubble. Our comparison between the theoretical lower bounds and simulated results of the number of required micro-robots implies that the searches in rubble with micro-robots becomes difficult as the rubble is large in the vertical direction.

Keywords: Search in Rubble, Micro-Robot, Graph-based Analysis

1 INTRODUCTION

As various robots have been developed [1–4], attempts to use robots for survivor searches in rubble in disaster areas are studied [5, 5]. These robots can enter dangerous areas for humans and perform their tasks. Moreover, using robots may reduce risks of further endangering the survivors and rescuers due to secondary disasters. For these reasons, robots are expected as one of promising means for search and rescue in disaster areas. Since smallness of robots allows them to break into small gaps in rubble, it is more likely to find survivors buried inside the rubble. Therefore, to utilize smaller robots, which are often called *micro-robots* [6–8], for survivor searches in disaster areas is considered [9–11].

Searches in rubble with micro-robots are generally performed on the basis of deployment of enormous micro-robots because of the following two reasons: one is a simple structure of micro-robots and the other is a complex structure of rubble. Since it is difficult to install many functions on small micro-robots, they have minimum capabilities for searching, such as moving and detecting obstacles and survivors [9]. Thus, one approach for achieving a fast completion of the search is deploying enormous micro-robots and searching many possible spaces simultaneously [10]. As another point of view about the difficulty in searches in rubble with micro-robots, Cho and Arnold [11] point out a possibility that micro-robots may fall into holes surrounded with debris inside rubble and they cannot escape from the holes. Therefore, enormous micro-robots have to be deployed so that they accomplish the search in rubble even if a certain amount of them fall into such holes.

To determine how many micro-robots should be deployed to complete a search in rubble with micro-robots is one of crucial issues since the searches are performed with a huge number of micro-robots. However, few researches analyze the number of deployed micro-robots to complete the search in rubble. We simply refer to the number of deployed micro-robots to complete the search in rubble as the number of deployed micro-robots, hereafter. In [10], a method to reduce the number of deployed micro-robots is introduced but the target of the research is two-dimensional spaces with several obstacles. However, the number of deployed micro-robots will strongly depend on complexity of three-dimensional structures of rubble since many holes exist inside rubble of collapsed buildings, as Cho and Arnold point out in [11]. In this paper, we use the number of deployed micro-robots for indicating one of the difficulties in search in rubble with micro-robots and analyze the number of required micro-robots in a three-dimensional structure of rubble.

As the first step to understand difficulty in searches in rubble with micro-robots, we theoretically derive the *minimum number of deployed micro-robots* to complete the searches under an ideal case where all micro-robots move optimally in the rubble on the basis of the complete information about the internal structure of the rubble. We model a pile of rubble as a graph and get the minimum number of micro-robots by deriving the minimum number of paths, which correspond to orbits of the micro-robots, to cover all vertices in the modeled graph. Then, we conduct several simulation experiments assuming that currently developed micro-robots having capabilities of moving horizontally and detecting humans are used. By comparing the theoretical and simulation results , we discuss how searches in rubble with the currently developed micro-robots is difficult.

The rest of this paper is organized as follows. In Section 2, we define the problem of searches in rubble with micro-robots. Then, a method to model a pile of rubble with a graph and a method to derive the theoretical minimum number of deployed micro-robots are described in Section 3 and 4, respectively. We compare the number of deployed micro-robots in the case of the primitive micro-robots with the theoretical results in Section 5. Finally, we conclude this paper in Section 6.

2 PROBLEM FORMULATION OF SEARCH IN RUBBLE WITH MICRO-ROBOTS

In this section, we define the problem of searches in rubble with micro-robots.

2.1 Micro-robots and Searches in Rubble

1) Micro-robots: We assume that low-cost and small, cm-scale, micro-robots are used for searches in rubble in this paper. The micro-robot is equipped with 1) a sensor device to detect humans, 2) a moving mechanism to move inside rubble and 3) a device for notifying findings of survivors. Regarding the human detection in rubble, we assume that micro-robots use a temperature sensor or carbon a dioxide sensor to detect humans. Regarding the moving capability of micro-robots, we assume that they have the capability of moving to horizontal direction since many currently developed millimeter-scale micro-robots have the capability of crawling ground [6–8]. That is, we do not assume that micro-robots have the moving capability toward the upward direction in rubble. Finally, regarding the notifications of the survivor detection, we assume that micro-robots have wireless communication functionality like Wi-Fi or Bluetooth and they can notify the finds of survivors from anywhere inside the rubble to outside the rubble. Hence, we do not make further consideration about notifications of the survivor detection, hereafter.

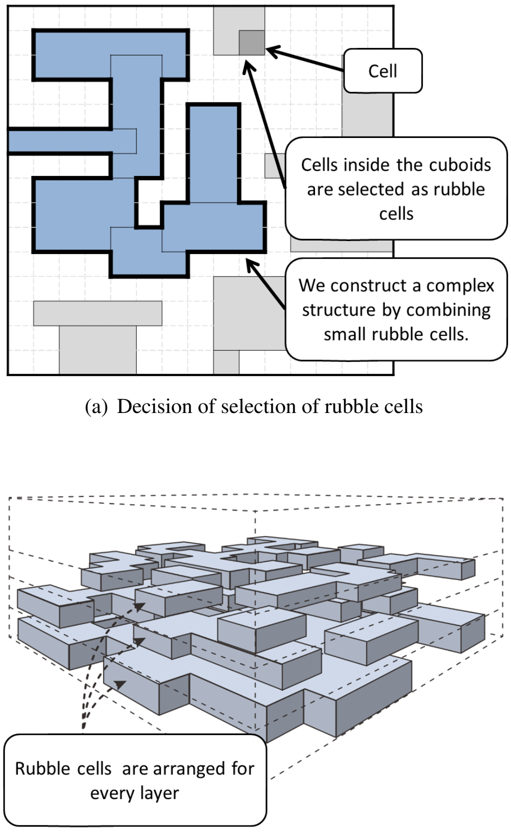
2) Searches in Rubble: A micro-robot finds a survivor if and only if it detects the survivor by using the installed sensor device, i.e., the micro-robots need to approach close enough to the survivor. To find all survivors buried in the rubble, micro-robots need to arrive at all possible *spaces*, which are defined as places without any pieces of rubble, at least once, since no knowledge about locations of survivors is available in advance of the search.

Micro-robots have only the horizontal movement capability. Therefore, once they drop vertically into the lower space, they can never go back to the upper spaces. Since the micro-robots have no way to avoid dropping into *holes*, which are defined as spaces surrounded with walls of debris, micro-robots that drop into holes no longer continue their search missions. That is, it is practically impossible for one micro-robot to reach all spaces in rubble and find all survivors buried there. Hence, for search in rubble with micro-robots, it is necessary to deploy multiple micro-robots and to search all possible spaces simultaneously.

2.2 Building Artificial Rubble

In this section, we explain a method for building rubble to analyze the difficulty in searches in rubble with micro-robots. We refer to rubble built with our method as *artificial rubble*.

We model rubble by piling small and equal-sized cuboids, which are the minimum units of artificial rubble, in a three-dimensional Euclidean space. We divide the three dimensional Euclidean space into small sections. We refer to the three-dimensional Euclidean space and the sections there as *field* and *cells*, respectively. These cuboids represent small blocks of rubble. That is, the cuboids are building blocks of the artificial rubble and we build the complex structure of rubble by piling the cuboids. In this paper, we use a cuboid with sides of 10, 10, and 5 cm for constructing the artificial rubble. The sizes of cells and cuboids are the same. We construct artificial rubble by placing the cuboids into cells. We refer to cells filled with the cuboids, which represent objects of rubble, as *rubble cell* and other cells except for rubble cells as *empty cells*. That is, empty cells represent spaces inside rubble. We handle rubble as a field of given *width*, *depth*, and *height*. To simplify notations, we refer to the lengths in the *x*, *y*, and *z* axes directions in the Euclidean space as width, depth, and height, respectively. Micro-robots discussed above can stay in empty cells having a rubble cell underneath them. We refer to these cells, where micro-robots can stand as *plane cells*. They can move horizontally on plane cells that are adjacent each other on the same height of the field. We refer to a set of adjacent plane cells with the same height as a *plane*.


Cell
Cells inside the cuboids are selected as rubble cells
We construct a complex structure by combining small rubble cells.
Rubble cells are arranged for every layer

(a) Decision of selection of rubble cells

(b) Example of artificial rubble

Figure 1: The overview of constructing artificial rubble

Figure 1 depicts the overview of constructing the artificial rubble. At the initial step, we fill a field of given width, depth, and height with empty cells. That is, the field is empty at the initial step. Then, we fill the field with rubble cells from the bottom toward the top step by step. At each layer of the field, rubble cells are placed until a given ratio, which we call *rubble ratio*, of cells are filled with rubble cells, as shown in Fig. 1(a). At each layer, one cell is selected randomly among empty cells. Then, we make a cuboid of $i \times j$ cells starting from the selected cell and fill with rubble cells in the cuboids. The integer numbers i and j are uniformly distributed random numbers between 1 and 10. Note that we make one constraint where at least one cells underneath the cuboid must be a rubble cell to avoid the situation where rubble cells float. We construct a complex structure by combining small rubble cells in this way, as shown in Fig. 1(b).

2.3 Definition of Searches in Rubble

This section defines the problem of searches in rubble with micro-robots. We define the completion of a search in rubble with micro-robots as the situation where all plane cells in artificial rubble are *covered* by micro-robots. A plane cell is covered when at least one micro-robot reaches the cell. Covering all plane cells means that micro-robots will find all survivors in the rubble. To cover all plane cells, a search in rubble with micro-robots are performed according to the following procedures. First, all micro-robots are deployed on top of the highest rubble cells. Each micro-robot moves independently from other micro-robots. Then, they move to one of adjacent empty cells of the same height in x or y axes directions per unit time. If the cell where micro-robots arrived at is an empty cell, they fall vertically to a plane cell right under the empty cell. Thus, if a plane is horizontally surrounded with rubble cells, they cannot escape from the plane. We refer to such planes as *holes*. We assume that micro-robots do not collide each other. Finally, we assume that micro-robots have enough battery capacity and therefore they do not stop moving till the completion of the search.

2.4 Difficulty in Searches in Rubble

This section explains a metric to indicate difficulty in a search in rubble with micro-robots and discusses a method for analyzing the relation between the size of rubble and the difficulty.

Since a micro-robot in a hole keeps staying there, all plane cells are not able to be covered by one micro-robot. Thus, we have to deploy multiple micro-robots to cover all plane cells. We expect that many micro-robots have to be deployed since many holes exist inside rubble of collapsed buildings. Therefore, the number of deployed micro-robots to complete the search is one of metrics to know how the search is difficult.

To measure the difficulty in the search in rubble with the number of deployed micro-robots, we compare results obtained by simulation with the theoretical lower bounds derived theoretically. In the simulation experiments, we assume that above-mentioned autonomous micro-robots are used, that is, micro-robots having only the horizontally moving capability move autonomously without any knowledge about rubble. We use the *minimum number of deployed micro-robots* to cover all plane cells as a metric to indicate the difficulty in the search in rubble. We simplify refer to the minimum number of deployed micro-robots as the minimum number of micro-robots, hereafter. That is, we use the minimum number of micro-robots as guidelines to measure the difficulty in the search.

To analyze the difficulty, we use a graph that represents a structure of the artificial rubble and derive the minimum number of micro-robots. We define the graph and a method to derive the minimum number of micro-robots from the graph in the following sections.

3 METHOD OF MODELING RUBBLE WITH A GRAPH

3.1 Overview

In this section, we define a search in rubble as a problem traversing a graph which corresponds to the artificial rubble constructed in the previous section. That is, we model the artificial rubble as a directed graph, of which vertices and directed edges correspond to places where micro-robots can be and paths where micro-robots can move, respectively. Then, we derive the minimum number of paths, which correspond to orbits of the micro-robots, to traverse all the vertices in the graph.

We model the artificial rubble as a directed graph according to the following two steps: 1) We simply express all possible places and paths where micro-robots can be and move as a directed graph and then 2) we simplify the graph by summarizing horizontal movements of micro-robots. More precisely, we first derive a graph G_1 by converting all empty cells, where micro-robots can be, as vertices and placing directed edges between all the possible vertices where micro-robots can move horizontally or fall vertically. That is, micro-robots can move from one vertex to another along directed edges in G_1. However, G_1 has many redundant information to know the minimum number of paths to cover all vertices, such as loops and unreachable vertices. Thus, we derive G_2 by summarizing several vertices in G_1 where micro-robots can move each other with horizontal movements to one vertex and removing unreachable vertices. That is, G_2 contains only directed edges that correspond to vertical movements of micro-robots. Consequently, the problem of a search in rubble is equivalent to covering all vertices in G_2 since all plane cell are contained in vertices in G_2 and the minimum number of micro-robots is equivalent to the minimum number of paths to cover all vertices in G_2. The following section mathematical defines the directed graphs and develops several heuristic algorithms to build the graphs.

3.2 Graph Formation

3.2.1 Formulating Rubble as a Graph

Before defining a directed graph, symbols used in this section are defined. The location of the cell in the artificial rubble is expressed by (i, j, k) $(i, j, k \in \mathbb{Z})$, which indicates i-th, j-th, and k-th cell in x, y, and z axis directions, respectively. The cell at (i, j, k) is represented as $c_{(i,j,k)}$. The function $C(i, j, k)$ is defined, which returns 0 if $c_{(i,j,k)}$ is an empty cell, return 1 if $c_{(i,j,k)}$ is a plane cell. X, Y, and Z are the width, depth, and height of the artificial rubble.

Let $G_1 = (V_1, E_1)$ be the directed graph, where V_1 and E_1 are the sets of vertices and edges. V_1 includes all plane cells in the artificial rubble and defined as follows:

$$\begin{aligned} V_1 = \{ v_{(i,j,k)} \mid\ & (i,j,k) \in I \times J \times K \wedge C(i,j,k) = 0 \\ & \wedge (k = 1 \vee (k \neq 1 \wedge C(i,j,k-1) = 1)) \\ & \vee (i,j,k) = (0,0,Z+1) \}, \quad (1) \end{aligned}$$

where I, J, and K represent ranges of x, y, and z axes, respectively. That is, I is represented as $I = \{x \in \mathbb{N} \mid 1 \leq x \leq X\}$ and J and K are represented similarly. The condition $C(i,j,k) = 0 \wedge (k = 1 \vee (k \neq 1 \wedge C(i,j,k-1) = 1)$ represents that $c_{(i,j,k)}$ is plane cell. Therefore, V_1 contains all the plane cells in the artificial rubble. Micro-robots are deployed on top of the rubble. Thus, we add another vertex to express the deployment of micro-robots, *root vertex* v_0, and the root vertex is represented by the constraint $(i,j,k) = (0,0,Z+1)$.

Next, we define the set of directed edges E_1 in G_1. Directed edges express the movement of micro-robots from one plane cell to another plane cell with horizontal movements and free falls and are defined as follows:

$$
\begin{aligned}
E_1 = \big\{ \big(v_{(i,j,k)}, v_{(l,m,n)}\big) \mid \\
(l,m,n) = (i \pm 1, j, k) \vee (l,m,n) = (i, j \pm 1, k) \\
\vee (((l = i \pm 1 \wedge m = j) \vee (l = i \wedge m = j \pm 1)) \\
\wedge \textstyle\sum_{h=n}^{k} C(l,m,h) = 0 \wedge k > n) \\
\vee \left((i,j,k) = (0,0,Z+1) \wedge \sum_{h=n}^{Z} C(l,m,h) = 0 \right) \bigg\}. \quad (2)
\end{aligned}
$$

The condition $(l,m,n) = (i \pm 1, j, k)$ and $(l,m,n) = (i, j \pm 1, k)$ represents the horizontal movements and $(((l = i \pm 1 \wedge m = j) \vee (l = i \wedge m = j \pm 1)) \wedge \sum_{h=n}^{k} C(l,m,h) = 0 \wedge k > n)$ represents movements from one plane cell to another in a free fall through empty cells. To express the deployment of micro-robots, E_1 contains the edge that is link from the root vertex v_0 to vertices of plane cells of the top of the rubble and this is represented by the last condition in Eq. (2).

Next, we present a heuristic algorithm to derive the directed graph G_1 from given artificial rubble in Algorithm 1. First, the sets of vertices and edges V_1 and E_1 are initialized with empty sets at lines 1 and 2. From line 3 to 11, G_2 is constructed from the given artificial rubble according to the definitions in Eqs. (1) and (2). The sub-function SET_EDGE defined at lines 12 to 15 is the function to set edges between vertices.

3.2.2 Summarizing a Graph

Next, we derive the graph $G_2 = (V_2, E_2)$ by summarizing several vertices in G_1 where micro-robots can move each other with horizontal movements to one vertex. That is, vertices in V_2 correspond to planes in the artificial rubble. A plane is denoted by $A_{(i,j,k)}$, which contains the plane cell $c_{(i,j,k)}$, and the constraint to aggregate plane cells and constructing a plane is expressed as

$$A_{(i,j,k)} = \{c_{(i,j,k)} \mid c_{(i,j,k)} \in A_{(i,j,k)s} \wedge A_{(i,j,k)s} = A_{(i,j,k)s-1}\}, \quad (3)$$

where $A_{(i,j,k)\,s}$ is defined using the following constraints.

$$A_{(i,j,k)\,1} = \{c_{(i,j,k)}\} \quad (4)$$

$$
\begin{aligned}
A_{(i,j,k)\,s} = \{c_{(l,m,n)} \mid c_{(l,m,n)} \in A_{(i,j,k)\,s-1} \\
\vee (c_{(l,m,n)} \in Q \wedge \exists c_{(s,t,u)} \in A_{(i,j,k)\,s-1} ((l,m,n) = (s \pm 1, t, u) \\
\vee (l,m,n) = (s, t \pm 1, u)))\} \quad (5)
\end{aligned}
$$

Algorithm 1 Constructing $G_1 = (V_1, E_1)$

Input: The artificial rubble, $c_{(i,j,k)}$

Output: The directed graph, $G_1 = (V_1, E_1)$

```
 1: $V_1 \leftarrow \emptyset$
 2: $E_1 \leftarrow \emptyset$
 3: for $k \leftarrow 1$ to $Z$ do
 4:     for $j \leftarrow 1$ to $Y$ do
 5:         for $i \leftarrow 1$ to $X$ do
 6:             if $C(i,j,k) = 0 \wedge C(i,j,k-1) = 1$ then
 7:                 $V_1 \leftarrow V_1 \cup \{v_{(i,j,k)}\}$
 8:             end if
 9:         end for
10:     end for
11: end for
12: for all $v_{(i,j,k)} \in V_1$ do
13:     SET_EDGE($v_{(i,j,k)}, c_{(i\pm1,j,k)}$)
14:     SET_EDGE($v_{(i,j,k)}, c_{(i,j\pm1,k)}$)
15: end for

16: function SET_EDGE($v_{(i,j,k)}, c_{(l,m,n)}$)
17:     if $C(l,m,n) = 0$ then
18:         if $n = 1 \vee C(l,m,n-1) = 1$ then
19:             $E_1 \leftarrow E_1 \cup \{(v_{(i,j,k)}, v_{(l,m,n)})\}$
20:         else
21:             $z \leftarrow n$
22:             while $C(l,m,z) = 0$ do
23:                 $z \leftarrow z - 1$
24:             end while
25:             $E_1 \leftarrow E_1 \cup \{(v_{(i,j,k)}, v_{(l,m,z+1)})\}$
26:         end if
27:     end if
28: end function
```

$$
\begin{aligned}
Q = \{c_{(i,j,k)} \mid (i,j,k) \in I \times J \times K \wedge C(i,j,k) = 0 \\
\wedge (k = 1 \vee (k \neq 1 \wedge C(i,j,k-1) = 1))\} \quad (6)
\end{aligned}
$$

To express the relationship between a plane and a vertex v_a that constitutes the plane a, we introduce S_{v_a}, where S_{v_a} is a set of plane cells that constitute the plane a. S_{v_a} satisfies $\forall c_{(i,j,k)} \in S(v_a) \forall c_{(l,m,n)} \in S(v_a)$ $(A_{(i,j,k)} \in A_{(l,m,n)})$. S_{v_0} is an empty set.

Next, we explain how a set of directed edges E_2 of directed graph G_2 is constructed. Edges in E_2 represent the vertical movements in free falls from one plane to the next plane. The edge from v_a to v_b must satisfy $\exists\, c_{(i,j,k)} \in S(v_a)$ $\exists\, c_{(l,m,n)} \in S(v_b)$ $(((l = i \pm 1 \wedge m = j) \vee (l = i \wedge m = j \pm 1)) \wedge k > n)$. Since micro-robots are deployed via the root vertex v_0, we connect the root vertex to all other planes which can reach from the top of the artificial rubble, i.e., directed edges are placed from v_0 to all v_a that satisfies the constraint $\exists\, v'_{(i,j,k)} \in S(v_a)$ $\sum_{h=n}^{Z} C(l,m,h) = 0$.

Then, we present a heuristic algorithm to derive the directed graph G_2 from the given rubble and the directed graph G_1 built with Algorithm 1. In The algorithm consists of three parts: initializing variables from lines 1 to 3, constructing vertices from lines 4 to 14, and constructing edges from lines 15 to 25. To distinguish which plane vertices belong to, we assign identifiers to all plane and the identifier of plane where

Algorithm 2 Constructing $G_2 = (V_2, E_2)$

```
Input: $G_1 = (V_1, E_1)$ and $c_{(i,j,k)}$
Output: $G_2 = (V_2, E_2)$
 1: $V_2 \leftarrow \emptyset$
 2: $E_2 \leftarrow \emptyset$
 3: $id \leftarrow 1$
 4: for $k \leftarrow 1, Z$ do
 5:     for $j \leftarrow 1, Y$ do
 6:         for $i \leftarrow 1, X$ do
 7:             if $C(i,j,k) = 0 \wedge C(i,j,k-1) = 1 \wedge c_{(i,j,k)}.id = 0$ then
 8:                 $V_2 \leftarrow V_2 \cup \{v_{id}\}$
 9:                 CLUSTER$(id, c_{(i,j,k)})$
10:                 id $\leftarrow id + 1$
11:             end if
12:         end for
13:     end for
14: end for
15: for $k \leftarrow 1, Z$ do
16:     for $j \leftarrow 1, Y$ do
17:         for $i \leftarrow 1, X$ do
18:             if $C(i,j,k) = 0 \wedge C(i,j,k-1) = 1$ then
19:                 $id \leftarrow c_{(i,j,k)}.id$
20:                 SET_EDGE2$(v_{id}, c_{(i\pm1,j,k)})$
21:                 SET_EDGE2$(v_{id}, c_{(i,j\pm1,k)})$
22:             end if
23:         end for
24:     end for
25: end for
```

$c_{(i,j,k)}$ belongs to is stored to $c_{(i,j,k)}.id$. The initial value of $c_{(i,j,k)}.id$ is zero. From lines 4 to 14, each plane in the given artificial rubble is converted to a vertex. That is, all adjacent plane cells of the same height are aggregated toe one plane. To construct planes, we use sub-function CLUSTER, which aggregates plane cells recursively assign the identifier to the plane cells. From lines 15 to 25, directed edges are placed from one plane to another where micro-robots can move vertically in free falls. Finally, we construct G_2' by removing vertices that are unreachable from v_0 and edges to the removed vertices since micro-robots cannot reach such spaces in the rubble.

4 THE MINIMUM NUMBER OF MICRO-ROBOTS

The theoretical minimum number of micro-robots is equivalent to the number of paths in the minimum path cover of the modeled graph G_2'. A path cover is a set of directed paths such that every vertex in the graph belongs to at least one of the path. If the path cover consists of the minimum number of paths, the path cover is referred to as the minimum path cover. In this section, we first explain the reason why the minimum number of micro-robots is equivalent to the number of paths in the minimum path cover and then how to derive the minimum path cover of G_2'.

Algorithm 3 Sub-functions for constructing $G_2 = (V_2, E_2)$

```
function SET_EDGE2$(v_{id}, c_{(i,j,k)})$
    $z \leftarrow k$
    if $C(i,j,z) = 0 \wedge C(i,j,z-1) = 0$ then
        while $C(i,j,z) = 0$ do
            $z \leftarrow z - 1$
        end while
        $id2 \leftarrow c_{(i,j,k)}.id$
        $E_2 \leftarrow E_2 \cup \{(v_{id}, v_{id2})\}$
    end if
end function

function CLUSTER$(id, c_{(i,j,k)})$
    if $C(i,j,k) = 0 \wedge C(i,j,k-1) = 1 \wedge c_{(i,j,k)}.id = 0$ then
        $c_{(i,j,k)}.id \leftarrow id$
        CLUSTER$(id, i \pm 1, j, k)$
        CLUSTER$(id, i, j \pm 1, k)$
    end if
end function
```

A search in the rubble with micro-robots is equivalent to traverses of vertices leaving from the root vertex along with directed edges on the directed graph derived from the rubble. Specifically, we have modeled a given pile of artificial rubble as a directed graph. Passing through a vertex in the graph is equivalent to surveying the space in the rubble that correspond to the vertex. In the similar way, moving along with a directed edge is equivalent to moving from one space to another in the rubble. Therefore, orbits of micro-robots in the rubble can be expressed as paths starting from the root vertex v_0 in the directed graph G_2'. The minimum number of deployed micro-robots is equivalent to the minimum number of paths, which are originating from v_0, to cover all vertices in G_2'.

Next, we derive the minimum number of paths originating from the root vertex v_0 for covering all the vertices in G_2'. A path P is defined as an ordered set of vertices. We denote a set of paths using $\mathscr{P}$, hereafter. The set of paths to cover all vertices is also called as *path cover* and it is defined as follows: $\mathscr{P}$ is a path cover of $G = (V, E)$ if $\mathscr{P}$ is a set of paths of G such that every $v \in V$ is included at least one path $P \in \mathscr{P}$ [12]. That is, a path cover must satisfy the following condition:

$$\forall v \in V \ \exists P \in \mathscr{P} \ v \in P. \tag{7}$$

The minimum path cover is a path cover $\mathscr{P}$ such that $|\mathscr{P}|$ is minimum.

The orbits of micro-robots can be expressed by paths but the paths must start with v_0 since micro-robots are deployed on top of the rubble. That is, the paths of micro-robots must satisfy the following constraint:

$$\forall P \in \mathscr{P} \ (v_0 \in \mathscr{P}) \wedge \forall v \in V \ \exists P \in \mathscr{P} \ (v \in P) \tag{8}$$

Such a path cover is called a single starting point path cover. Therefore, the constraints of paths of the path cover problem and the search in rubble with micro-robots are slightly different, as shown in Eqs. (7) and (8). The minimum single starting point path cover is a single starting path cover $\mathscr{P}$ such that $|\mathscr{P}|$ is minimum. Though algorithms to deriving the minimum path cover have been already developed but no algorithms to derive the minimum single starting point path cover.

If the number of elements of the minimum single starting point path cover is equal to that of the minimum path cover, the minimum number of micro-robots can be derived by solving the problem of the minimum path cover of G_2'. We prove that the number of elements of the minimum single starting point path cover is equal to that of the minimum path cover as follows: Since for all vertices $v \in V_2'$ in G_2' at least one path that originates from v_0 and reaches to v exists, a set of paths $\mathscr{P}$ exists such that $\mathscr{P}$ satisfies the following condition, $|\mathscr{C}| = |\mathscr{P}| \wedge \forall P \in \mathscr{P}\ (v_0 \in P) \wedge \forall P' \in \mathscr{C}\ \exists P \in \mathscr{P}\ (P' \subseteqq P)$, where $\mathscr{C}$ is the minimum path cover of G_2'. That is, the set of paths $\mathscr{P}$ is the minimum path cover and the all paths in $\mathscr{P}$ start with v_0. The number of elements of the minimum path cover is equal to that of the minimum single starting point path cover. Hence, the minimum number of micro-robots can be derived by solving the problem of the number of elements of the minimum path cover of G_2'.

Finally, we explain a method to derive the minimum path cover in G_2'. G_2' is a directed acyclic graph (DAG) [12]. It is a well-known fact that the minimum vertex-disjoint path cover in DAG can be derived by solving the maximum matching problem by converting the DAG into a bipartite graph [13]. The minimum vertex-disjoint path cover is a set of the minimum number of elements in vertex-disjoint path covers. The vertex-disjoint path cover is a set of paths $\mathscr{P}$ such that for every $v \in V$ in G there exists at exactly one path $P \in \mathscr{P}$ including v. The vertex-disjoint path cover must satisfy another constraint that no paths in the set cannot share vertices in addition to the minimum path cover. However, if the target graph G is a DAG, the minimum path cover of G is equal to the minimum vertex-disjoint path cover of the transitive closure of the graph G_{clo} [12]. Therefore, we can have the minimum path cover of G_2' by deriving the minimum vertex-disjoint path cover of transitive closure of G_2'. In this way, the minimum number of micro-robots deployed to complete the search in the rubble can be derived.

5 DIFFICULTY IN SEARCH IN RUBBLE WITH MICRO-ROBOTS

In this section, we analyze the difficulty in searches in rubble with micro-robots. First, we describe our simulation environments. Then, we analyze the number of deployed micro-robots and the minimum number of deployed micro-robots by change the width/depth and height of the artificial rubble.

5.1 Simulation Conditions

We set parameters of the artificial rubble as follows: To express complicated shape of rubble, we set the size of one cell to 10 × 10 × 5 cm. Since we suppose that micro-robots search inside highly dense rubble, where a large robot cannot enter, we set the percentage of rubble cells in the artificial rubble to reasonably high, i.e., 0.65. In our simulation experiments, micro-robots are deployed randomly on top of the artificial rubble. In the rubble, the micro-robots move horizontally on planes according to Lévy Flight mobility model [14, 15], which is know as one of efficient mobility patterns to search targets, and they fall in free falls when they reach an empty cell. In this paper, we ignore situations where micro-robots stop working during the search due to several issues, such as failures of locomotion or sensor devices. We define that a plane cell is surveyed if at least one micro-robot enters the plane cell. The search will be finished within 24 hours since the probability of humans under rubble being alive decreases rapidly 24 hours after they are buried. We compute the number of deployed micro-robots to cover 90% of plane cells in the artificial rubble within 24 hours. We compute the average of results obtained from 30 simulation trials. In the following section, we investigate how the search in rubble is difficult by evaluating the number of deployed micro-robots by changing the size of the artificial rubble.

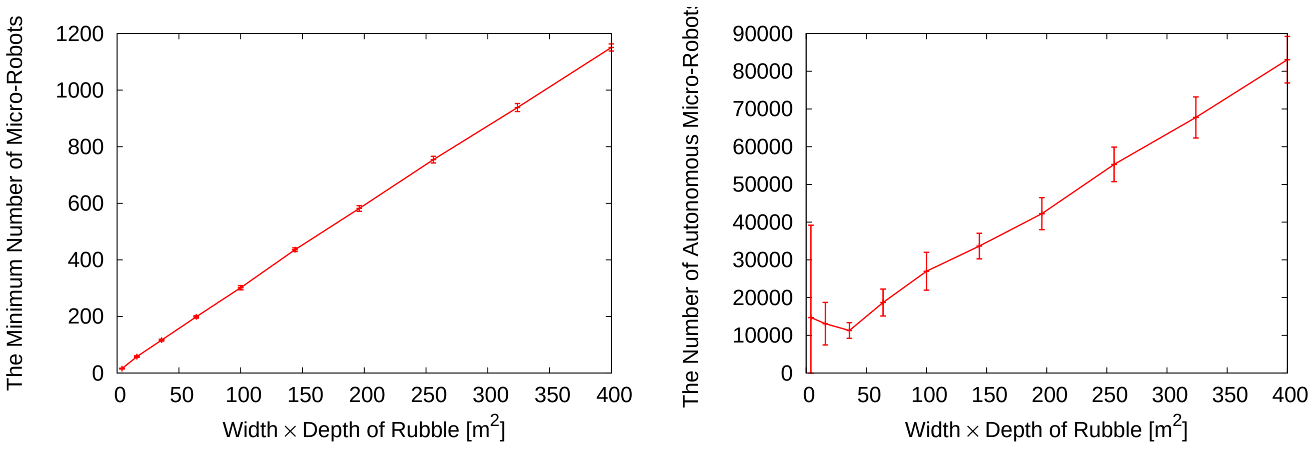
5.2 The Number of Deployed Micro-robots

First, we observe the effects of the area size, i.e., the width multiplied by the depth, of the artificial rubble on the number of deployed micro-robots. Figure 2 shows relations between the number of deployed micro-robots and the area size of the rubble. In this simulation, the height of the rubble is set to 1.5 meters. The horizontal axes of the figures are the area size, which is defined as the width multiplied by the depth. The minimum number of micro-robots derived theoretically is shown in Fig. 2(a) and the number of deployed micro-robots obtained through simulations is shown in Fig. 2(b). The error bars in Fig. 2(b) indicate the 95% confidence intervals. Comparing the absolute values of the analytical and simulation results is nonsense. We compare the tendency of the results in this paper. Both the minimum number of micro-robots derived theoretically and the number of deployed micro-robots obtained through simulations increase almost proportionally to the area size. These results imply that primitive micro-robots, which have only the horizontal movement function as their locomotion function, can be used for search in wide range of rubble.

Next, we observe the effects of the height of the artificial rubble on the number of deployed micro-robots in Fig. 3. The horizontal axes indicate the height of the artificial rubble. The vertical axes are the same as those in Fig. 2. The minimum number of micro-robots derived theoretically is shown in Fig. 3(a) and the number of deployed micro-robots obtained through simulations is shown in Fig. 3(b). In this simulation, both the width and depth of the artificial rubble is set to 10 meters. In contrast to the minimum number of micro-robots derived theoretically, which increases almost proportional to the height, the number of deployed micro-robots obtained through simulations increases more sharply. The results suggest that the search in the rubble with primitive micro-robots is quite difficult if the rubble is high.

6 CONCLUSION

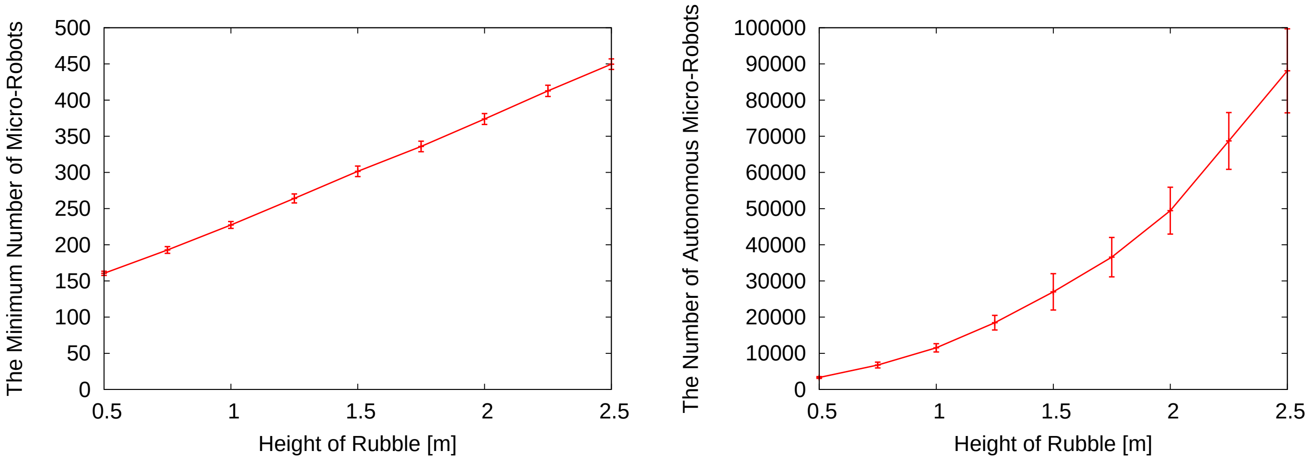
In this paper, we analyze the number of deployed micro-robots to complete the search in the rubble. To investigate the difficulty in the search, we derive the theoretical lower bounds of the number of deployed micro-robots, which can be used as guideline to measure the difficulty in the search in the rubble. As the first step to analyze the difficulty in



(a) The minimum number of micro-robots derived theoretically

(b) The number of deployed micro-robots derived by simulations

Figure 2: The minimum number of micro-robots and the number of deployed micro-robots in the case that the area size of the rubble is changed.



(a) The minimum number of micro-robots derived theoretically

(b) The number of deployed micro-robots derived by simulations

Figure 3: The minimum number of micro-robots and the number of deployed micro-robots in the case that the height of the rubble is changed.

the search in the rubble, we compare the number of deployed primitive micro-robots, which have only the horizontal movement function as their locomotion function, with the minimum number of micro-robots derived theoretically by changing the area size and the height of the rubble. In contrast to the minimum number of micro-robots derived theoretically, which increases almost proportional to the height, the number of deployed micro-robots obtained through simulations increases more sharply. The results imply that searches in rubble with primitive micro-robots get difficult as the rubble gets large in the vertical direction.

REFERENCES

[1] K. Osuka and H. Kitajima, "Development of mobile inspection robot for rescue activities: Moira," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003*, vol. 4, pp. 3373–3377, IEEE, 2003.

[2] M. Arai, Y. Tanaka, S. Hirose, H. Kuwahara, and S. Tsukui, "Development of "souryu-iv" and "souryu-v: " serially connected crawler vehicles for in-rubble searching operations," *Journal of Field Robotics*, vol. 25, no. 1-2, pp. 31–65, 2008.

[3] J. Casper and R. Murphy, "Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 33, no. 3, pp. 367–385, 2003.

[4] B. Yamauchi, "Packbot: A versatile platform for military robotics," in *Proceedings of SPIE*, vol. 5422, p. 229, 2004.

[5] R. R. Murphy, S. Tadokoro, D. Nardi, A. Jacoff, P. Fiorini, H. Choset, and A. M. Erkmen, "Search and rescue robotics," in *Springer Handbook of Robotics*, pp. 1151–1173, Springer, 2008.

[6] E. Edqvist, N. Snis, R. C. Mohr, O. Scholz, P. Corradi, J. Gao, A. Dieguez, N. Wyrsch, and S. Johansson, "Evaluation of building technology for mass producible millimetre-sized robots using flexible printed circuit boards," *Journal of Micromechanics and Microengineering*, vol. 19, June 2009.

[7] B. R. Donald, C. G. Levey, C. D. Mcgray, I. Paprotny, and D. Rus, "An untethered, electrostatic, globally controllable MEMS micro-robot," *Microelectromechanical Systems*, vol. 15, pp. 1–15, Feb. 2006.

[8] T. Ebefors, J. U. Mattsson, E. Kalvesten, and G. Stemme, "A walking silicon micro-robot," in *Proceedings of the 10th International Conference on Solid-State Sensors and Actuators*, pp. 1202–1205, June 1999.

[9] S. Dubowsky, J. S. Plante, and P. Boston, "Low cost micro exploration robots for search and rescue in rough terrain," in *Proceedings of IEEE International Workshop on Safety Security and Rescue Robotics*, Aug. 2006.

[10] D. K. Sutantyo, S. Kernbach, P. Levi, and V. A. Nepomnyashchikh, "Multi-robot searching algorithm using lévy flight and artificial potential field," in *Proceedings of IEEE International Workshop on Safety Security and Rescue Robotics*, pp. 1–6, July 2010.

[11] J. H. Cho and M. G. Arnold, "Survivor search using a quasi-2D-parallax algorithm with massive microrobot swarms," in *Proceedings of the 14th WSEAS International Conference on Systems*, pp. 522–525, 2010.

[12] M. Kowaluk, A. Lingas, and J. Nowak, "A path cover technique for LCAs in dags," in *Algorithm Theory–SWAT 2008*, pp. 222–233, Springer, 2008.

[13] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (Instructor's Manual Second Edition)*, ch. 26. CreateSpace Independent Publishing Platform, Jan. 2014.

[14] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, S. Havlin, M. G. E. D. Luz, E. P. Raposo, and H. E. Stanley, "Lévy flights in random searches," *Journal of Physica A: Statistical Mechanics and its Applications*, vol. 282, no. 1, pp. 1–12, 2000.

[15] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. G. D. Luz, E. P. Raposo, and H. E. Stanley, "Optimizing the success of random searches," *Nature*, vol. 401, no. 6756, pp. 911–914, 1999.

Labor Service by Own Copy Robot

Osamu Yuki*

*Canon Inc.

yuuki_osamu@yahoo.co.jp

Abstract – In recent year, people can easily take a picture of around a person by such as a drone equipped with a camera. Moreover, three-dimensional (hereinafter abbreviated as 3D) image data can be generated from these pictures by using the algorithms such as SFM or SIFT. Then, the shape as a 3D model is possible to make from 3D image data by a 3D printer. I think that people will become making own copy robot in a personal factory in the future. Such a factory come true with digital imaging technology, 3D reconstruction technology, image recognition technology, 3D printing technology and robot control technology. Image recognition technology recognize the hand, the foot, the body and the head from the 3D image data. These 3D image data are shaved spaces for mounting of the mechanisms, the control electronics and CPU. And these parts are produced robot shapes by a 3D printer. The people will own a 3D printer. And a person will get possible to make the own copy robot in one's private factory. In this paper, I propose to make an own copy robot in order to change the work style. The robot assemble the parts, such as a actuate mechanisms, the control electronics and the artificial intelligence (hereinafter abbreviated as A.I.) . In proposed making robot method, an robot owner can obtain the copyright of the it. The value chain form the work service by offering this robot. It is important that AI, a robot and a person evolve together changing each role in the economic system.

Keywords: labor service, own copy robot, work style

1 INTRODUCTION

The work style has been changing by the development of information equipment and infrastructure. For example, there is the providing of labor forces by using the video, audio and data to the company from the home and external office the like. People are freed from constraints of time and place by such a work style. In order to maximize the working ability of the individual, it is necessary to slide into activating the sharing of information to be exchanged within the group using a variety of devices in the workplace. Such a work style was made possible by the spread of ICT technologies such as "smart devices, communication network by wired or wireless, social networks and cloud". It is possible to share information through efficient communication on a global basis.

On the other hand, changes in work style has also been advanced by the use of A.I. is known such as a "WATSON" of IBM [1].

However, if these changes has been progressed, the employment of people may be replaced to A.I.. In addition, The information sharing between people will be replaced to the information sharing between A.I.s in the network.

In case of the workforce provided, its income is divided by the company providing the software of A.I. . We need to search for a method of providing a labor force that build a value chain can benefit for people. In this paper, I propose that people provide labor services by an own copy robot. The value chain will solve the singularity problem.

2 RELATED WORKS

Android is appearing in TV commercials. Professor Hiroshi Ishiguro is director of the Intelligent Robotics Laboratory at Osaka University, Japan. A notable development of the laboratory is the Actroid, a humanoid robot with lifelike appearance. Technology of making androids, robots closely resembling the human form, has been established. For example, Matsukoroid [2] is often referred to by the media. In the case of Matsuko, there was the possibility to scan her to obtain data. They had to take a trial and error process using three-dimensional Computer aided design system (hereinafter abbreviated as CAD). Matsukoroid can record the sound of Matsuko Deluxe and makes a voice synthesis. Therefore, Matsukoroid is able to talk various texts with the voice of Matsuko Deluxe. Photograph of a Matsukoroid is shown in Figure 1.

On the other hand, Robi is a robot created by Tomotaka Takahashi. Design of Robi has been focused on communication and emotions [3]. Robi has no practical function, except the pleasure of those who share their time with him. Robi can turn its head and pelvis 360 degrees. Microphones in its ears detect the location of a speaker and it will automatically turn its head towards him or her. Robi is the 35-cm tall and 1kg weight toy robot. It has a vocabulary of 250 English words and can understand more than 250 phrases and command. It also can dance and show emotion by changing the color of its eyes. It will take some time to complete making Robi, but he will gain functions at each stage. De Agostini ModelSpace expert team breaks it down in Five Easy Build Phases.

Photograph of a small robot ROBI is shown in Figure 2.

Following are its functions.

(1) Speech-recognition Board

Robi has a sophisticated speech recognition board programmed to understand many different English phrases and reply appropriately.

(2) Microcontroller Board

The high-performance microcontroller board controls important functions such as Robi's motion and reactions.

(3) Speaker

Robi’s chest includes a miniature speaker for its spoken responses.

(4) Motion Sensors

Robi’s motion sensors detect human presence and will turn its face in response to movement.

(5) Servo Motors

Robi’s movements are controlled by a total of 20 servos. They use the Robi-servo command system to ensure smooth movement, and have simple, robust wiring connections.

In a news paper, researchers at MIT’s Computer Science and Artificial Intelligence Laboratory (CSAIL) present the first-ever technique for 3-D printing robots that involves printing solid and liquid materials at the same time [4]. A photograph of a 3-D hexapod robot is shown in Figure 3. The new method allows the team to automatically 3-D print dynamic robots in a single step, with no assembly required, using a commercially-available 3-D printer. This 3-D hexapod robot moves via a single motor, which spins a crankshaft that pumps fluid to the robot’s legs. Besides the motor and battery, every component is printed in a single step with no assembly required. Among the robot’s key parts are several sets of “bellows” 3-D printed directly into its body.

The people will own a 3D printer. And a person will get possible to make the own copy robot in one’s private factory.

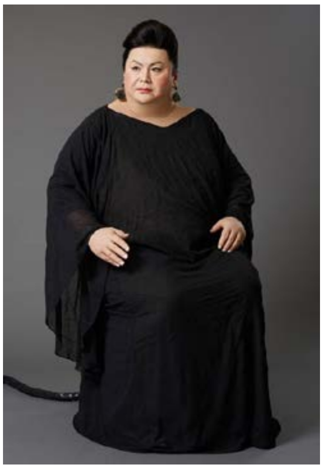


Figure 1: Matsukoroid

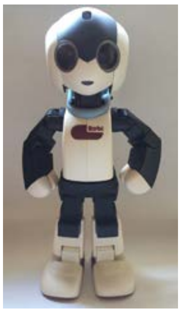


Figure 2: Small robot ROBI

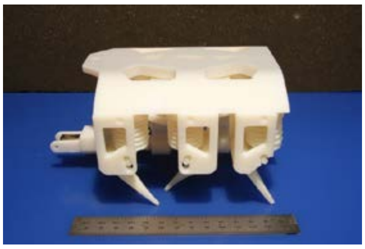


Figure 3: 3-D hexapod robot

It is important to install A.I. for an autonomous robot.

Basically, A.I. would be the ultimate version of Google. So we have the ultimate search engine that would understand everything on the Web. It would understand exactly what we wanted, and it would give us the right thing. That's obviously artificial intelligence, to be able to answer any question, basically, because almost everything is on the Web, right? We're nowhere near doing that now.

However, we can get incrementally closer to that, and that is basically what we work on. And that's tremendously interesting from an intellectual standpoint [5].

And, we know “WATSON” as the A.I. supplied by IBM. Watson detects variants of the same answer and merges their feature scores together. Watson then computes the final confidence scores for candidate answers by applying a series of Machine Learning models that weight entire feature scores to produce final confidence scores [1]. The process of WATSON is shown in Figure 4. (Referring to Dave Mobley, 2014)

However, Stephen Hawking warns artificial intelligence could end mankind. Prof Stephen Hawking has said that efforts to create thinking machines pose a threat to our very existence. He told "The development of full artificial intelligence could spell the end of the human race." [6].

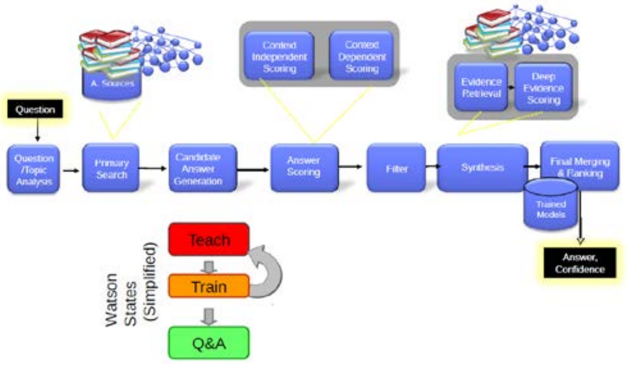


Figure 4: Watson

3 EXPERIMENT AND THE RESULTS

I explain the making flow of the own copy robot in figure 5.

First, I used the multi-copter to take 2D photographs. In order to take 2D photographs from the multi-angle, I could use the multi-copter phantom 3 manufactured by DJI equipped with camera. In addition, because the multi-copter was equipped with gimbals to keep the camera horizontally, it could take stabilized 2D photographs even if the aircraft was shaking. By using this multi-copter, I took 2D photographs of myself from the over-head to the feet while changing an angle of the camera. However, the camera has a wide angle lens. Therefore, the lens has the lens distortion. 2D photographs were distorted, because I did not correct the distortion of 2D photographs. 2D photographs are shown in Figure 6. Reconstructed 3D text mapping pictures were also distorted. Reconstructed 3D pictures are shown Figure 7. IN future study, I will take photographs by using a multi- copter again. And, I will correct the lens distortion of photographs.

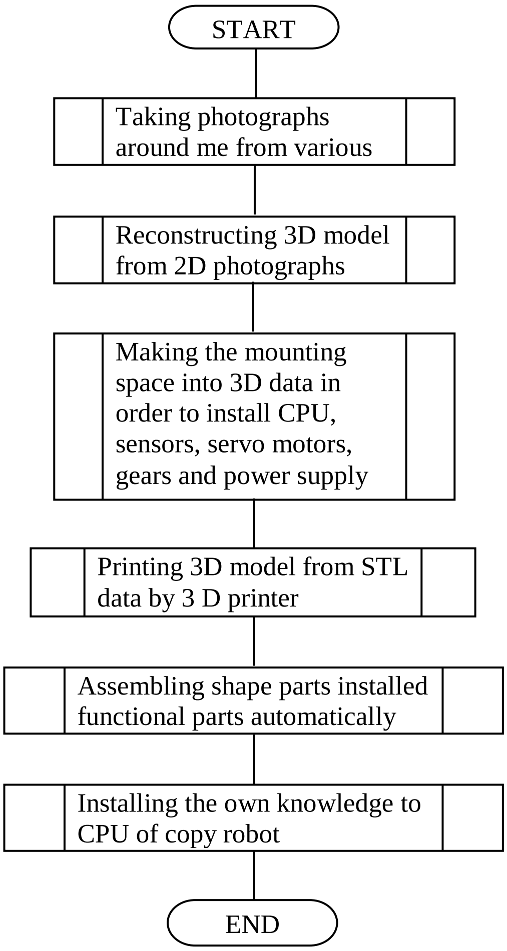

START
Taking photographs around me from various
Reconstructing 3D model from 2D photographs
Making the mounting space into 3D data in order to install CPU, sensors, servo motors, gears and power supply
Printing 3D model from STL data by 3 D printer
Assembling shape parts installed functional parts automatically
Installing the own knowledge to CPU of copy robot
END

Figure 5: Making flow of own copy robot

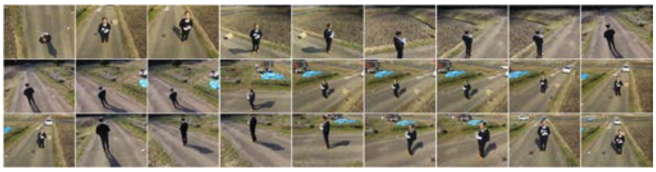


Figure 6: 2D photographs by multi-copter

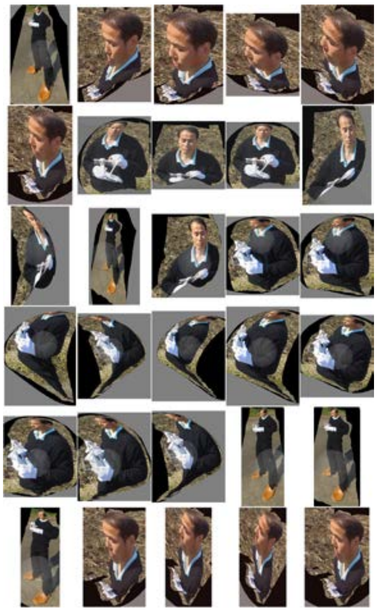


Figure 7: 3D pictures with lens distortion

In next step, I prepared the compact digital camera to take 2D photographs from the multi-angle. Some of 2D photographs are shown in Figure 8.

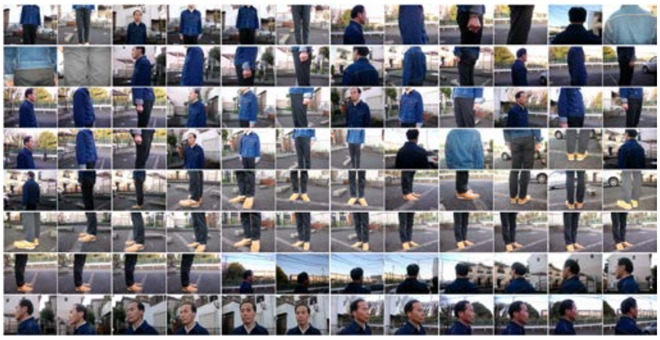


Figure 8: 2D photographs by compact digital camera

These 2D photographs were reconstructed into a 3D image by PhotoScan using SfM algorithm [7][8][9]. PhotoScan is has been made by Agisoft Company. First, PhotoScan perform alignment of 2D photographs. Camera positions and angles in 3D view are shown in Figure 9. The alignment of 2D photographs is performed in order to analyze the position that took 2D photograph. PhotoScan analyze the position of camera from EXIF information buried in photographs and the overlap of each photograph automatically.

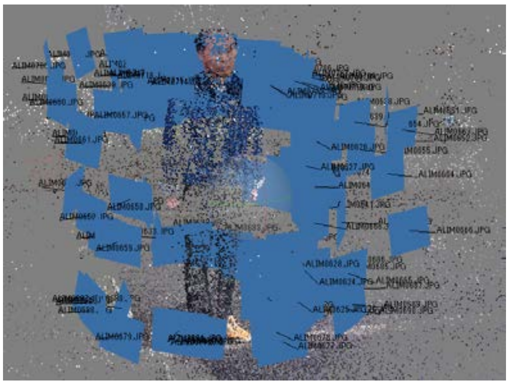


Figure 9: Camera position and point in 3D view

When the camera model is pinhole model is pinhole model, the SfM parameters at each camera position is given as (1) and (2).

Assuming the origin of the image plane is the image center, the three dimension point m :

$$m = (m_x, m_y, m_z) \tag{1}$$

can be converted into the image point p :

$$p = (p_x, p_y) \tag{2}$$

from following equitions (3), (4), and (5) [10].

$$m' = R(m - c) \tag{3}$$

$$p = (-fm'_x / m'_z - fm'_y / m'_z) \tag{4}$$

$$p = (1 + k_1|p'|^2 + k_2|p'|^4)p' \tag{5}$$

Here,

(f): Camera focal length

(c) : Camera central location

(k_1, k_2) : Distortion of radius parameter

(R) : Rotation matrix

Reconstructed 3D images are shown in Figure 10.

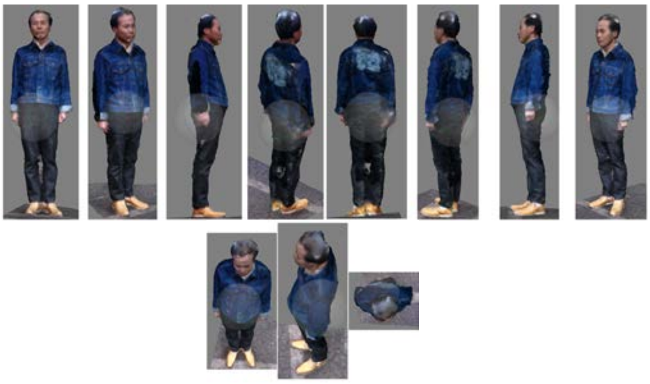


Figure 10: Reconstructed 3D images

The following things are proposed by the desk study. The 3D object can be constructed from the data of the STL format by a 3D printer. 3D data of the STL format consists of many triangle mesh shape. Triangle facets are described with three vertexes coordinates and normal vectors. PhotoScan can output the STL format data. Therefore, 3D model can be made from the 3D data immediately. However, separating the head, the hand, the body and the leg is needed in order to assemble a moving robot. The robot has to actuate these parts. After output process of the 3D model is completed, the 3D model has to be driven with the designed degrees of freedom. Therefore, their parts are shaved the mounting space off, in order to install a CPU, servo motors, gears and a power supply.

However, it is difficult to assemble the robot by this method full automatically. On the other hand, the image recognition technique by machine learning has been known. We have recognized the face by machine learning. The face recognition from the 2-dimensional image is shown in Figure 11. Furthermore, we have collected the correct image of the photograph of a servomotor. Then, we have created learning pattern from these photographs. And, we have gotten feature vectors from these patterns. We have classified the servomotor pattern from a 2-dimensional photograph. We show this result in Figure 12. The result was correct recognition. Pattern recognition algorithm of Figure 11 and Figure 12 is shown following.

Pattern Recognition Algorithm

We apply adaptive boosting (hereinafter abbreviated as AdaBoost) algorism to enhance the learning capability. AdaBoost algorithm [11][12][13] detects a face and a servo-motor.

Given example images

$(x_1, y_1), \ldots, (x_n, y_n)$

Where

$y_i = 0{,}1$ for negative and positive examples respectively.

Initialize weight

$w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0{,}1$ respectively,

Where m and l are the number of negatives and positives respectively.

For $t = 1, \ldots, T$:

1. Normalize weights,

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$

So that w_t is a probability distribution.

2. For each feature, j , train a classifier h_j which is restricted to using a single feature. The error is evaluated with respect to

$$w_t, \varepsilon_j = \sum_i w_i \,|\, h_j(x_i) - y_i \,| .$$

3. Choose the classifier, h_t, with the lowest error ε_t.

4. Update weights:

$$w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$$

Where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ Otherwise, and $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$.

The final strong classifier is:

$$h(x) = \{ \begin{matrix} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T}\alpha_t \\ 0 & Otherwise \end{matrix}$$

Where $\alpha_t = \log\frac{1}{\beta_t}$

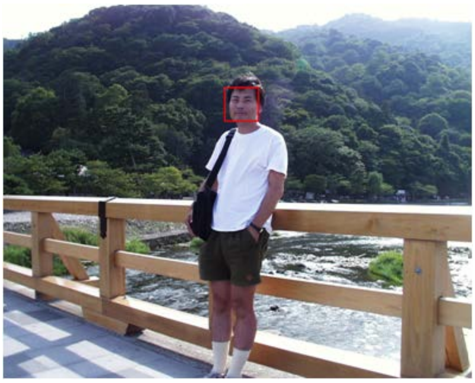


Figure 11: Face recognition

Figure 12: Recognition of servo motor

By using the Figure 11 method, we can classify the head, the hand, the body, and the leg from 3D shape data. And, the Figure 12 method can be applied to classify the servo motor from robot parts.

In case of applying machine learning to 3D shape recognition, it is able to increase the correct recognition rate more than the 2D image recognition. Because of image recognition technology can recognize the head, the hand, the body and the leg from the 3D image data, each outer shape as parts can be made by a 3D printer by using the recognition results. In addition, the mounting space in order to install a CPU, servo motors, gears and a power supply can be made by processing the 3D image data of parts. These parts using image recognition are made by a 3D printer from the STL format data, too. The parts of head, hand, body, leg, sensor, and actuator are picked up from the parts box by using the image recognition technique, and the robot is assembled by using the image recognition technique. When these processes are finished, the hardware making process of the own copy robot is completed.

The software of machine learning is installed to the memory devices of the robot hardware. The software of machine learning learns the personal knowledge. This robot is able to take action based on the personal experience and knowledge. An artificial neural network is shown in Figure 13. An artificial neural network has generally had the problem of slow training multi-layer neural networks. For example, a neural network for face recognition is defined by a set of input neurons which may be activated by the pixels of an input image [14]. Therefore, A.I. should learn many faces. It is possible to use alternative computer architectures to speed up the processing. However, this learning process usually needs long time even though uses GPU-based implementations. Usually the learning rule will depend on the activities of the neurons. The learning rule depends on the values of the target supplied by the teacher or on the current value of the weights. If I have the past database of target faces, A.I. can learn them from the database. If I do not want to wait for the long time, I can use A.I. by unsupervised learning. The learning time becomes shorter with the recognition which gradually goes up. The learning process of the A.I. is able to be switched into on or off. In the off state, the weights of the interconnections are fixed. When the robot finishes learning photographs, texts, speeches, etc. from my information devices (For example, PC), the robot stops the learning.

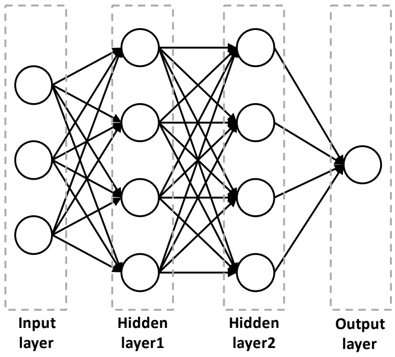


Figure 13: Artificial neural network

In proposed making robot method, an robot owner can obtain the copyright of it. The value chain form the work service by offering this robot. The value chain is shown in Figure 14. It is important that a robot and a person evolve together with each different ole in the economic systems.

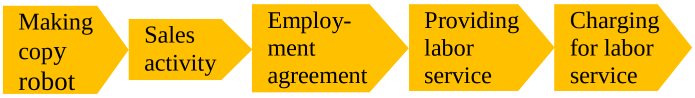


Figure 14: Value chain of labor service

4 CONCLUSION

As discussed above, the person can earn income by incorporating the own copy robot in flow of value chain labor services. In addition, physical body can be used to gain knowledge or creative activities without spending time for work.

When the own copy robot is laid off by one’s employer or A.I. of the copy robot became inefficient for elapsed time, the person can be added the new knowledge A.I.. The proposed labor service is important so that a robot coexists with the human race.

In the future, I will continue the research for increasing of accuracy of individual elemental technology. Especially, I research how to import efficiently my memorized data to my own robot from my information devices. And I want to make the proposed service come true.

REFERENCES

[1] How Watson Works, available from http://www.cs.uky.edu/~raphael/grad/keepingCurrent/HowWatsonWorks.pdf (accessed 11-Apr-2016).

[2] Matsukoroid, available from http://prnavi.jp/wp-content/uploads/2015/07/150716matsukoroid.pdf (accessed 09-Apr-2016).

[3] TOP INTERVIEW - Tomotaka Takahashi, available from https://rikeinavi.com/16/contents/magazine/img/pdf10s/10s_0205.pdf#search='TOMOTAKA+TAKAHASHI+PDF' (accessed 04-Apr-2016).

[4] First-ever 3-D printed robots made of both solids and liquids, available from http://news.mit.edu/2016/first-3d-printed-robots-made-of-both-solids-and-liquids-0406 (accessed 17-Apr-2016).

[5] ANITA WASILEWSKA CSE 352 ARTIFICIAL INTELLIGENCE, available from http://www3.cs.stonybrook.edu/~cse352/G19Google.pdf (accessed 11-Apr-2016).

[6] Stephen Hawking warns artificial intelligence could end mankind, available from http://www.bbc.com/news/technology-30290540 (accessed 11-Apr-2016).

[7] R. A. Andersen and D. C. Bradley., “Perception of three-dimensional structure from motion,” Trends in Cognitive Sciences, Vol. 2, No 6, pp. 222 – 228, 1998

[8] N. Saveley , M. S. Seitz. and R. Szeliski., “Modeling the World from Internet Photo Collections,” International Journal of Computer Vision, Vol. 8, Issue 2, pp. 189 – 210, 2007.

[9] C. Tomashi and T. Kanade. , “Shape and motion from image streams under orthography: factorization method,” International Journal of Computer Vision, Vol. 9, Issue 2, pp. 137 – 154, 1992.

[10] Generation of Arbitrary Viewpoint Images from Image Compilation by Estimation of Position and Pose of Mobile Camera, available from http://www.robot.t.u-tokyo.ac.jp/~yamashita/paper/E/E175Final.pdf (accessed 17-Apr-2016).

[11] Paul Viola and Michael J.Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", IEEE CVPR, 2001.

[12] Rainer Lienhart and Jochen Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", IEEE ICIP pp.900-903, Sep. 2002.

[13] Yuuki, O., Yamada, K., Kubota, N., “Trajectory Tracking for a Pitching Robot based on Human-like Recognition”. In: IEEE CIRA 2009, Dec.2009.

[14] Chris Edwards, “Growing Pains for Deep Learning”, Communications of the ACM, Vol. 58, No. 7, pp.14-16, July. 2015

Session 5:
Data Analysis

(Chair: Tomoya Kitani)

A Multimedia Control and Processing Framework for IoT Application Development

Daijiro Komaki[*], Shunsuke Yamaguchi[*], Masako Shinohara[*], Kenichi Horio[*], Masahiko Murakami[*], Kazuki Matsui[*]

[*]Fujitsu Laboratories Ltd., Japan
{komaki.daijiro, yamaguchi.shun, m-shinohara, horio, mul, kmatsui}@jp.fujitsu.com

Abstract – When creating Internet-of-Things (IoT) applications, it is difficult to deal with multimedia data captured from cameras and microphones installed at field sites since it requires a wide variety of knowledge of topics such as codecs, protocols and image processing. To solve this problem, therefore, we propose a framework that makes it easy to deal with multimedia stream data in IoT application development. Our framework has three main features as follows: (1) Virtualization of multimedia input/output devices; (2) Distributed execution of multimedia processing pipeline between gateways and a cloud; and (3) Simple service description using a graphical flow editor. In this paper, we present some of the proof-of-concept applications we created and discuss the effectiveness of our framework from the perspective of complexity, productivity and ease of trial and error.

Keywords: Internet-of-Things; Multimedia; Framework; Web API

1 INTRODUCTION

We have entered the era of the *Internet-of-Things* (IoT), where not only computers but also physical objects (i.e., *things*) such as vehicles and home appliances are connected to the internet and interact with each other, or with systems, services and people. When creating such IoT applications, it is important to make devices such as sensors and actuators already installed at field sites (e.g., classrooms, concert halls, building entrances) available for various applications, rather than to install devices at a field site for a specific purpose [2, 5].

In addition, not only sensory data (e.g., temperature and acceleration) but also multimedia data (i.e., audio and video) captured from devices such as cameras and microphones installed at field sites are important for IoT application development, since we can offer many beneficial applications that utilize multimedia data as sensory data by using computer vision technologies (e.g., *detecting a suspicious person at a building entrance*), or that utilize sensory data as an input to process multimedia data (e.g., *adding effects to a live video stream of a concert event according to the mood of audiences there*).

On the other hand, there are some multimedia frameworks such as Kurento[1] [4] and Skylink[2] that make it easy to create applications utilizing multimedia data. These frameworks provide multimedia server programs and client libraries, and developers can easily create their applications without worrying about the differences in codecs and formats of audio/video contents by using the provided libraries. For example, developers can easily create applications equipped with multimedia features (e.g., VoIP, augmented reality) simply by connecting multimedia processing blocks as a pipeline. However, even when using these frameworks, problems remain when considering the characteristics of IoT application development, as follows:

- Since there may be many different types of devices at field sites, developers need to know the detailed specifications (e.g., an interface to start/stop capturing media data, to establish a media session between a device and a media server) in advance, and need to create applications according to the specifications.

- If all multimedia stream data generated at field sites are continuously transferred to a media server, they consume large amounts of network bandwidth.

To solve these problems, we designed and implemented a framework to simplify the process of multimedia IoT application development. Our framework has three main features as follows:

Virtualization of multimedia input/output devices

Our framework provides a mechanism to virtualize multimedia input/output devices (e.g., cameras and speakers) to obscure the differences in heterogeneous device specifications. This mechanism means that developers no longer need to consider the details of devices already installed at field sites, and applications once created can be adapted for another field site.

Distributed execution of multimedia processing pipeline between gateways and a cloud

Our framework provides a mechanism to distributedly process single multimedia stream data by coordinating multiple multimedia servers running independently on gateways (installed at field sites) and on a cloud, respectively. This mechanism enables developers to easily create multimedia IoT applications that can save network bandwidth usage and can serve immediate detection and response to field sites.

[1] Kurento: https://www.kurento.org/

[2] Skylink: http://skylink.io/

Simple service description using a graphical flow editor

Our framework provides a web-based graphical flow editor tool to simply define a distributed multimedia processing pipeline by connecting input/output device blocks and filter blocks. This tool enables developers to easily use trial and error by replacing and relocating each block.

In this paper, we present some of the proof-of-concept applications we created and discuss the effectiveness of our framework from the perspective of complexity, productivity, and ease of using trial and error.

2 RELATED WORK

2.1 Multimedia frameworks

There are several frameworks that make it easy to create applications that utilize multimedia stream data. GStreamer [3] is an open-source framework for creating multimedia applications that handle audio, video and any kind of data flow in a modular way. The basic idea of GStreamer is to link together various plug-in elements (e.g., sinks/sources, encoders/decoders, filters) on provided pipeline architecture to obtain a stream that meets the desired requirements. This seems to be effective for developers who are not familiar with multimedia processing or multimedia networking. However, even when using GStreamer, developers are required to know which type of devices, which protocols, and which codecs to use in advance, in order to define a pipeline.

Kurento Media Server is an open-source multimedia server based on GStreamer that supports WebRTC. Developers can easily create web-based multimedia streaming applications (e.g., VoIP, video conference, augmented reality) using provided APIs. Since Kurento Media Server provides the mechanism to absorb the differences in media codecs and formats, even developers unfamiliar with multimedia processing (e.g., web application developer) can create multimedia streaming applications by linking media processing modules (e.g., image processing, event detection) via the provided web APIs. In addition, this framework provides a way to implement a new media processing module using OpenCV [4](Open-Source Computer Vison). Computer vision experts can create their new modules independently from the application development process.

Although it becomes easy to deal with multimedia stream data by using such framework technologies, these frameworks are not necessarily suitable for creating IoT applications (that make use of devices already installed at field sites) since developers need to create their applications according to device type, protocols, codecs and so on.

2.2 IoT application development platforms

On the other hand, there are several cloud-based platforms for creating and deploying IoT applications that utilize multiple sensors and actuators installed at field sites. Kii Cloud [5], a *Backend-as-a-Service* for IoT application development, provides the functionalities to virtualize devices on the cloud. IoT application developers can create their IoT applications by combining multiple virtualized device functionalities by using provided APIs, so they need not be concerned about the differences in detailed specifications such as communication protocols.

IBM Bluemix [6] also provides a way to create IoT applications using virtualized device functionalities on the cloud. Bluemix provides a graphical flow editor (called Node-RED [7]) to create interactive, near real-time IoT applications by simply connecting things and services. Blackstock [3] focused on the fact that many IoT scenarios require the coordination of computing resources across networks: on servers, gateways, and devices, and extended Node-RED in order to create distributed IoT applications that can be portioned between servers and gateways. MyThings [8] (provided by Yahoo! Japan) enables users to create IoT applications that link various devices to various web services by simple IF-THEN rules.

Owing to such platforms, it becomes easy to create IoT applications that connect multiple devices and services to each other. However, if developers attempt to create an IoT application that deals with multimedia streams generated from field sites, they have to use multimedia frameworks such as those mentioned above.

There have been a few efforts to simplify the creation of IoT applications that can handle both sensory data and multimedia data in combination. ThingStore [1] provides the mechanisms to virtualize any type of device as a thing that generates Boolean data (i.e., Boolean value represents whether a certain event occurs or not). Owing to this abstraction, application developers can deal with media input devices as sensor devices and can simply create IoT applications that coordinate both sensory and multimedia stream data. However, since ThingStore abstracts multimedia stream data as Boolean data, it is not suitable for dealing with end-to-end multimedia stream data transferred from a device to another device.

2.3 Target of our framework

We aim to make it easier to deal with not only sensory data but also multimedia stream data among devices in such cloud-based IoT application development platforms. In this paper, we propose a framework that makes it easy to create such multimedia IoT applications by extending Kurento Media Server. Our framework focuses on IoT applications

[3] GStreamer: https://gstreamer.freedesktop.org/

[4] OpenCV: http://opencv.org/

[5] IoT Cloud Platform Kii: https://en.kii.com/

[6] IBM Bluemix: https://www.ibm.com/developerworks/cloud/bluemix/

[7] Node-RED: http://nodered.org/

[8] myThings: http://mythings.yahoo.co.jp/

where sensory stream data and multimedia stream data affect each other interactively.

3 KURENTO MEDIA SERVER

In this section, we describe Kurento Media Server, which is the basis of our framework implementation. Kurento Media Server is an open-source software media server that makes it simple to create web applications equipped with multimedia features (e.g., VoIP, video conference, augmented reality). Kurento Media Server provides endpoint modules (i.e., elements to input or output the multimedia stream) and filter modules (i.e., elements affecting the media stream or detecting events from the media stream) shown in Table 1. Application developers are simply required to take the modules needed for an application and to connect them, without worrying about differences in codecs and formats of audio/video data. Figure 1 shows an application example: the video stream captured by the web browser is sent to the media server, then *FaceOverlayFilter* detects faces from frames of the video stream and puts a specified image on top of them, and finally the face-overlaid video stream is sent back it to the web browser, while recording it on the media server.

Kurento Media Server provides Java and JavaScript client libraries. Developers can use the functionalities Kurento Media Server offers on their applications. Figure 2 shows the required procedures to construct the pipeline shown in Figure 1. Firstly, a web browser-side script creates a Session Description Protocol[9] (SDP) offer and sends it to the web application server. Then, a server-side script creates the pipeline by using the provided client library, makes *WebRtcEndpoint* process the SDP offer in order to get an SDP answer, and sends the SDP answer back to the client. Finally, the client processes the received SDP answer to establish a WebRTC session with the *WebRTCEndpoint* on the media server. When the client starts sending the video stream, the processed video stream data is sent back to the client. Our framework also uses the provided client library to control Kurento Media Server in the same manner as described.

The filters and endpoints that Kurento Media Server provides have their own methods for clients to change inner parameters. Moreover, events raised by filters are subscribable by client-side scripts. For example, in Figure 3, an application can change the image path to overlay and can subscribe an event that a barcode is detected in a frame of video stream. Our framework makes use of such features of Kurento Media Server and implemented some additional functionalities from the perspective of the IoT scenario.

4 FUNCTIONAL REQUIREMENTS

Typical scenario cases we assume as multimedia IoT applications are as follows.

Table 1: Modules provided by Kurento Media Serer

Endpoints (Inputs/Outputs)	
WebRtcEndpoint	Send and receive WebRTC media flow
RtpEndpoint	Send and receive RTP media flow
PlayerEndpoint	Read media from a file or URL
RecorderEndpoint	Store media flow to a file or URL
Filters (Processing/Detecting…)	
FaceOverlayFilter	Recognize face areas and overlay picture on that area.
ZBarFilter	Detect barcode and QR code
GStreamerFilter	Use filters of GStreamer

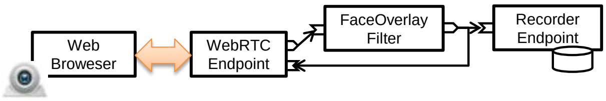


Figure 1: Example of connecting endpoints and filters

[Case 1] *Capture live video stream from a certain camera at a field site and transfer it to a screen installed at the same place.*

[Case 2] *Obtain text segment data from live audio stream captured from a certain microphone by speech recognition, overlay the text on live stream video captured from a certain camera, and project the text-overlaid video stream on a nearby screen.*

[Case 3] *Count the number of people from live video stream captured by a certain camera, and detect an event according to the change of that number.*

[Case 4] *Record live stream video captured from a certain camera installed at a field site only when the temperature there is higher than a threshold value.*

In order to make it easy to create such multimedia IoT applications, we extract the requirements of functionalities that the framework should provide as follows:

(1) Virtualization of multimedia input/output devices

Considering the above scenarios, it is desirable to deploy an IoT application once created to many various field sites rather than to create an IoT application for a specific field site. However, there may be different types of devices (e.g., IP camera that supports RTSP, USB camera) at field sites and they may communicate using different protocols (e.g., WebRTC, RTP, HTTP). Such heterogeneity does not become a problem when creating conventional web applications since developers already know which type of device to use (i.e., devices are virtualized on the HTML5 layer on the browser side). However, from the perspective of IoT scenarios, it is assumed that many IoT applications utilize the same devices already installed at field sites together. Therefore, the framework should obscure such device heterogeneity so that developers do not need to consider it.

[9] RFC 4566 - SDP: Session Description Protocol: http://tools.ietf.org/html/rfc4566.html

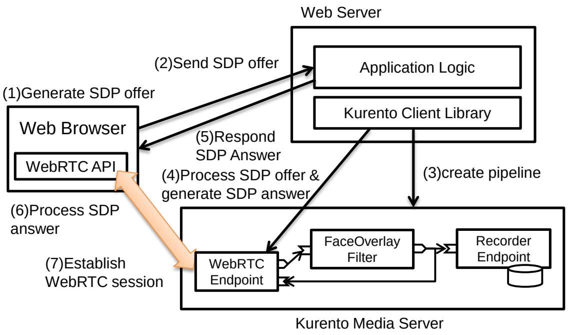
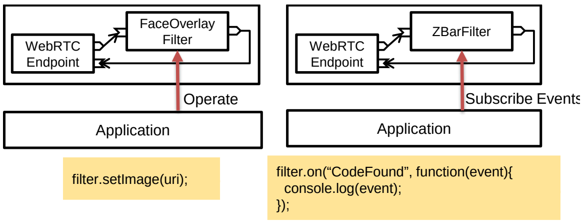


Figure 2: Procedure to establish media session



```
filter.setImage(uri);
```

```
filter.on("CodeFound", function(event){
  console.log(event);
});
```

Figure 3: Interaction between application and filters

(2) Distributed execution of multimedia processing pipeline between gateways and the cloud

Since multimedia stream data is far larger than sensory data, it consumes a large amount of network bandwidth if transferring all multimedia stream data generated at field sites to the cloud. Considering the case where the results of event detection from video stream data captured from a field site are fed back to the same field site (Case 2), or the case where an IoT application needs only metadata extracted from multimedia streams (Case 3), Processing multimedia stream data on a computational resource near the field site is effective in saving network bandwidth usage and responding to the field site quickly.

To realize this, it is effective to process a single multimedia stream distributedly between a gateway (installed at the field site) and a cloud. However, in order to do that, it requires laborious procedures such as opening ports for sending/receiving multimedia stream on multiple media servers and establishing a media session between them. Therefore, the framework should enable the processing of single multimedia stream distributedly without considering media session establishment.

(3) Cooperation with external system

We assume not only the case where the framework control and process end-to-end multimedia stream are sent from one device to another (Case 1 and 2) but also the case where the framework detects an event using time series metadata extracted from a multimedia stream (Case 3) and the case where the framework controls the media stream according to the changes in sensory data (e.g., temperature) (Case 4). To do that, the framework should provide a way to easily cooperate with an existing IoT platform that provides the functionalities of time series data analysis or complex event processing.

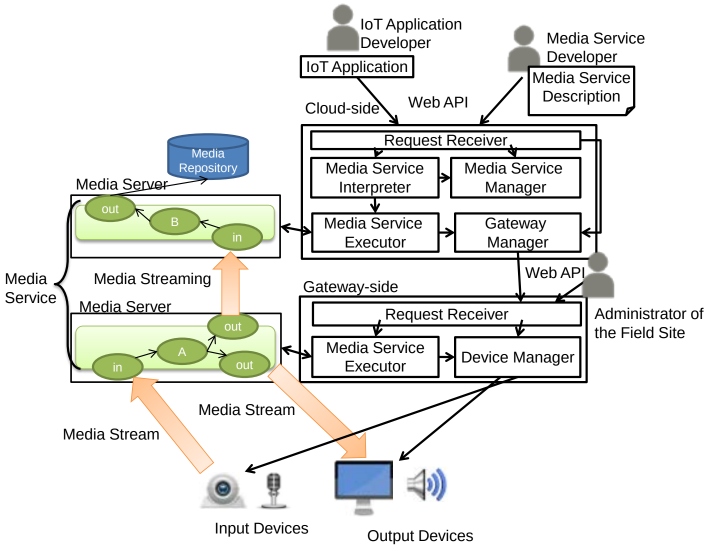


Figure 4: Architecture of our framework

(4) Simple description of media processing pipeline

By using media framework technologies such as Kurento Media Server, developers can easily create multimedia web applications by connecting multiple endpoints and filters as a pipeline, and can easily use trial and error by replacing or reconnecting each block. The framework should inherit this feature to simply implement a media processing pipeline, while satisfying the above three requirements ((1)-(3)).

5 ARCHITECTURE

In order to meet the above requirements, we designed the architecture of our framework (Figure 4). Here, we define the term media service as a set of an entity to process multimedia stream data running on media servers, and an IoT application as an entity to use media service(s) by using the APIs that our framework provides. Our framework uses multiple media servers running independently on the gateway(s) and the cloud, respectively, and deploys corresponding modules to cooperate with multiple media servers. The framework forms star topology, where the cloud-side module aggregates all gateway-side modules. Noted that we adopted Kurento Media Server as the media server, but other media servers are adaptable to realize such architecture. In the following, we describe the behavior of each component.

[Request Receiver]

This component receives the requests from the clients (e.g., registration of devices, gateway, media services, and operation of media services) via web APIs (shown in Table 1). The person who installs the devices uses the web APIs on the gateway side, while media service developer and IoT application developer use those on the cloud side.

[Media Service Manager]

This component manages the media service descriptions written in JavaScript Object Notation (JSON) format. Since media services to be executed on the gateway side are

Table2: Web APIs

URI	Method		parameters
/service	GET	Get a list of registered media service descriptions	
/service	POST	Register new media service description	id service
/service/:id	GET	Get the media service description specified by id	id
/service/:id/create/:params	GET	Create media service instance using specified media service description by assigning specified parameters.	id params
/pipeline	GET	Get a list of executed media service instances	
/pipeline	POST	Create media service instance	service:
/pipeline/:id	GET	Get the media service instance specified by ID	id
/pipeline/:id/:method	GET	Operate media service specified by ID (start | pause | stop | release)	id method
/gw	GET	Get a list of registered gateways	
/gw	POST	Register a new gateway	key uri
/device	GET	Get a list of registered devices	
/device	POST	Register a new device	key uri

deployed at the time of execution, media service descriptions are centrally managed on the cloud side.

[Media Service Interpreter]

This component converts media service descriptions into executable ones for the **Media Service Executor**; this component divides media service description into cloud-side and gateway-side media services.

[Media Service Executor]

Based on converted media service descriptions (described above), this component initializes and controls media processing modules on the corresponding media server. In addition, this component establishes media sessions between devices and gateways and between gateways and the cloud. The cloud-side **Media Service Executor** cooperates with the **Gateway Manager** in order to deploy a media service to the specified gateway and establish the session between media services, while the gateway-side one cooperates with the **Device Manager** in order to establish a media session between the device and the endpoint on the gateway-side media server.

[Gateway Manager]

This component manages the relationships between ID of each gateway (i.e., keyword to specify a field site) and their URLs and provides an interface to control gateways. When receiving the requests from **Media Service Executor**, this component forwards it to the specified gateway-side module.

[Device Manager]

This component manages the relationships between ID of each devices and connection information (e.g., URL, socket ID in the case of using HTTP, WebSocket, respectively) and provides an interface to negotiate SDP and to start/stop and sending/receiving multimedia stream data.

In the following, we describe the procedure to create a multimedia IoT application using our framework functionalities. The main players in this scenario are *Field Site Administrator*, *Media Service Developer*, and *IoT Application Developer*. These players may be either the different persons respectively or the same person.

(1) Registration of the gateway

The *Field Site Administrator* edits the configuration file in the gateway module to define the field site ID. When starting up the gateway-side module, a request for registering this gateway is automatically sent to the cloud.

(2) Registration of the devices

The *Field Site Administrator* registers devices to the gateway-side module via using gateway-side APIs by specifying the device ID and device type. Device IDs needs to be identifiable only in the same field site since they are managed by each gateway.

(3) Registration of the media service

The *Media Service Developer* writes the media service description (such as shown in Figure 5) and registers it to the cloud-side module using cloud-side APIs.

(4) Execution of the media service

The *IoT Application Developer* connects his/her application to the specified media service using server-side APIs to operate media services.

6 DETAILS OF FUNCTIONALITIES

In this section, we describe the detailed behaviors of the functionalities of each component above.

Interpretation/execution of media services

The *Media Service Developers* describe their services in JSON format (shown in Figure 5). This example shows a

```
[
  {
      id: 0,
    //type of  device
    type: "DeviceEndpoint",
    //specify the device id to use
    key:"camera001",
    //specify a place to execute media service
    place: "gw",
    //specify the field site id ($0 is variable)
    front_id: "$0",
    out:[0]
  },
  {
      id: 1,
    type: "FaceOverlayFilter",
    //filter's own property (specify image url to overlay)
    img: "http://XXXX/hat.jpg",
    place: "gw",
    in:[0],
    out:[1]
  },
  {
      id: 2,
    type: "RecorderEndpoint",
    place: "cloud",
    in:[1]
  }
]
```

Figure 5: Example of media service description

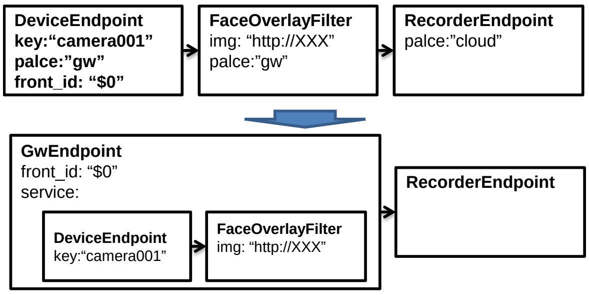


Figure 6: Transformation of media service description

media service where the video stream captured from a specified camera at a specified field site is processed to put a specified image on the face area on the gateway side and sent to cloud side to be recorded. Here, **type** is used to specify the type of filter/endpoint, **place** is used to specify the execution place (i.e., gateway or cloud), **front_id** is used to specify the field site where the media service is applied, and **in/out** is used to specify the relationship between elements.

The **Media Service Interpreter** divides the received media service description and creates a *GwEndpoint* that includes a partial media service description that should be executed on the gateway side (shown in Figure 6). Based on this converted media service description, the **Media Service Executor** initializes filters and endpoints on the media server and connects them.

In addition, the media service description can accept variable definition. For example, in Figure 5, **front_id** (i.e., the ID that specifies where the device is installed) is defined as a variable (**"$0"**) so that it can be set when this media service is executed.

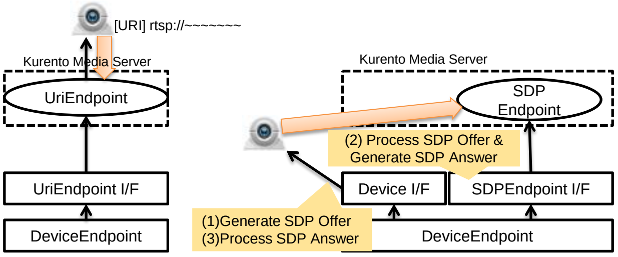


Figure 7: Obscuring the initialization procedure that varies according to device type.

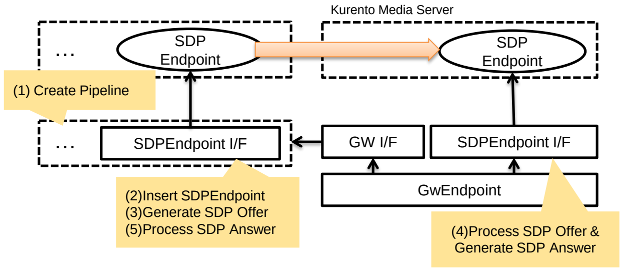


Figure 8: Session establishment between gateway and cloud

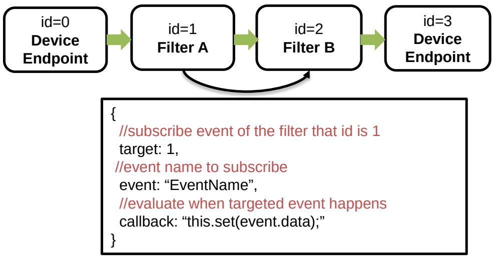
Virtualization of Media Devices

As an endpoint of multimedia stream data via a network, Kurento Media Server has three different types of endpoints: *RTSPEndpoint*, *WebRTCEndpoint* and *RTPEndpoint*. When using a camera that supports RTSP, it is required to simply specify the resource URL to establish a media session between the device and a media server, while it is necessary to manually negotiate SDP when using a camera that supports WebRTC or RTP. Moreover, since each device may provide its own interface to operate (start, stop), developers need to take care of how to establish a session and how to operate devices that vary according to device type.

Therefore, the framework provides a set of classes, each of which implements required procedures to establish a session according to the corresponding device type (Figure 7). *Field Site Administrators* are required to specify the device type when registering a new device. Owing to this, *Media Service Developers* do not need to be concerned about such differences in devices. A media session is automatically established when executing the media service.

Cooperation between Gateways and Cloud

To execute a media service distributedly on gateways and a cloud, it is necessary to establish a media session between divided partial media services. Therefore, the framework automatically inserts an *SDPEndpoint* (i.e., either *RTPEndpoint* or *WebRTCEndpoint*) at the end of the gateway-side me*dia service* description. The framework also creates an *SDPEndpoint* on the cloud-side media server and establishes a media session between gateway-side and cloud-side *SDPEndpoints* (as shown in Figure 8). Thereafter,



```
{
  //subscribe event of the filter that id is 1
  target: 1,
 //event name to subscribe
  event: "EventName",
  //evaluate when targeted event happens
  callback: "this.set(event.data);"
}
```

Figure 9: Event description between elements

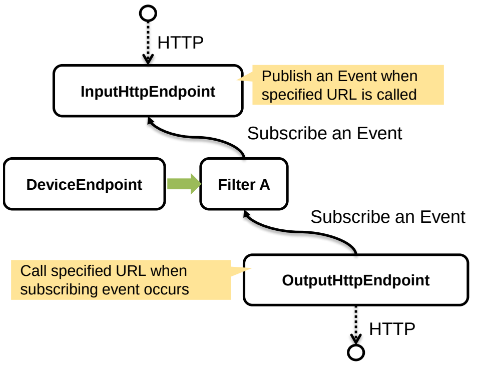


Figure 10: Input/output endpoint to external systems

when the cloud-side module receives the request to operate a *media service*, it propagates this request to the corresponding gateway-side module.

Management of Events

The framework enables developers to describe event subscription between two filters in a media service description. As shown in Figure 9, three attributes are required to define an event subscription: **target** (to specify which filter or endpoint publishes the event), **event** (to specify which type of event to subscribe), and **callback** (to describe the callback function that is evaluated when the specified event occurs). In the callback function, variable **this** is bound as the subscriber element itself.

Additionally, as shown in Figure 10, the framework provides two endpoints in order to cooperate with external systems: *InputHttpEndpoint* publishes the event when a specified URL is called by an external system and *OutputHttpEndpoint* subscribes inner events and calls the external URL (specified in advance) when a specified event occurs. By using these endpoints, developers can easily create a media service that can process media stream data according to environmental changes or can store time series of metadata extracted from media stream data into external databases.

Graphical Editor for Media Service Description

Developers are able to create media service by following JSON format as shown in Figure 5 without coding complicated logic. Furthermore, we implemented a web-based graphical flow editor for easily creating media service descriptions (Figure 11-13). In the following, we describe how to use this client.

Figure 11 shows an example of the screen for creating and editing the media service, which is implemented using a SVG-based JavaScript library, JointJS [10]. When a user selects an item from the left-side list, a new node appears on the center area. The user can make a link from an input port of a node to an output port of another node by dragging and dropping. When a selected item requires some properties (e.g., image URL path for *FaceOverlayFilter*), input forms corresponding to each property appear on the right side of the screen. Event subscription can be defined in this area.

Figure 12 shows the screen for executing specified media service. The user can select which field sites to apply the specified media service to. Figure 13 shows a list of executed media services and the user can operate (i.e., start, stop, pause, release) each media service.

7 PROOF-OF-CONCEPT APPLICATIONS

In order to verify the effectiveness of our framework, we created three proof-of-concept IoT applications. In this section, we explain these IoT applications and discuss the features of each application.

[Prototype 1] Supporting lectures in the classroom

In the lectures at universities, teachers often use the projector to present their documents on the screen display. In such lectures, a teacher may use a stick or laser pointer to specify the focus area of the screen display. However, students may not clearly see the specified area in a large classroom. Therefore, we implemented an application that supports such lectures. We implemented it by connecting *TrapezoidCorrectorFilter* (which transforms a trapezoid-shaped area to square), *FingerDetectorFilter* (which detects the coordinates of fingertips and raises an event), and **ScalerFilter** (which expands the area around a specified point) as shown in Figure 14.

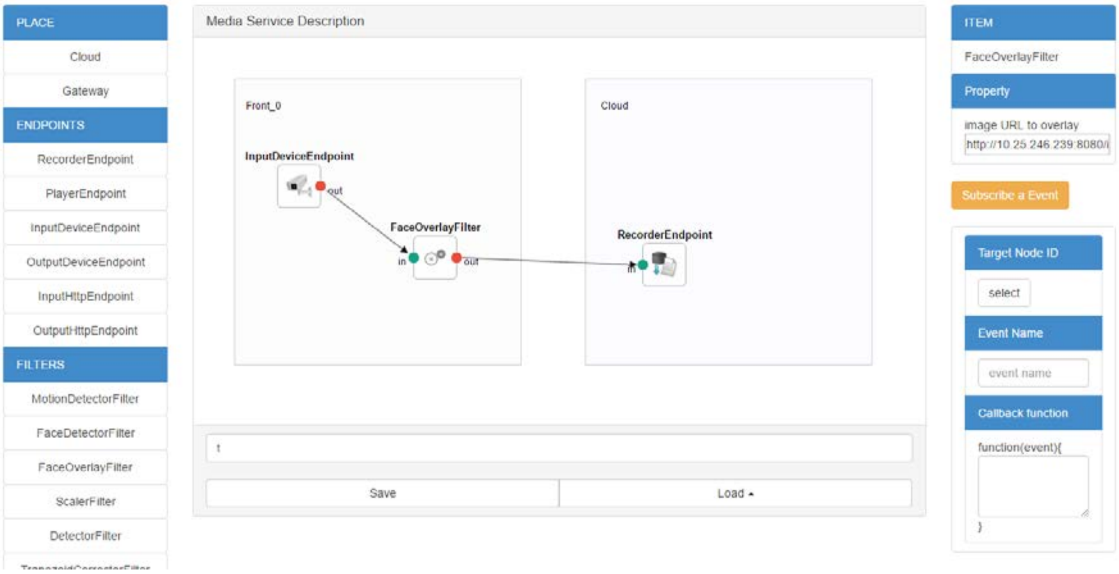
There are three devices registered to the framework in the classroom: a camera (which captures video stream data including screen display area for detecting fingertips), a screen capturer (which captures video stream data from the teacher’s PC screen), and a display screen (which displays the video stream process by *ScalerFilter*). Using this combination, the area of the screen the teacher points is scaled so that students can look at the focused area clearly. Since this media service is executed on the gateway-side, immediate response (i.e., followability of finger motion) can be expected compared with executing on the cloud-side.

[Prototype 2] Monitoring suspicious person

Suspicious person monitoring services using networked cameras are now widely used in various areas. However, when operating such monitoring services on the cloud, it consumes a large amount of network bandwidth and storage. Therefore, we created an IoT application that transfers a video stream data to the cloud while detecting a moving object and records it on the cloud.

We realized this by connecting *MotionDetectorFilter* (which detects moving objects using background

[10] JointJs: http://www.jointjs.com/

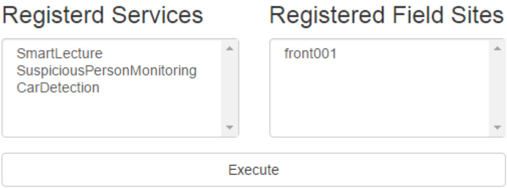


t

image URL to overlay: http://10.25.246.239:8080/i

function(event){
}

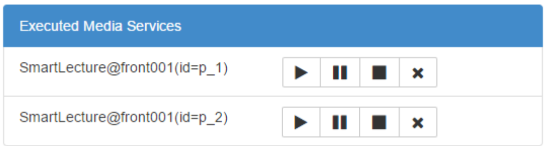
Figure 11: Media service description screen



SmartLecture
SuspiciousPersonMonitoring
CarDetection

front001

Figure 12: Media service execution screen



SmartLecture@front001(id=p_1)

SmartLecture@front001(id=p_2)

Figure 13: List of executed media services

subtraction) and *SwitchFilter* (which can be switched to drop or pass-through received buffer) as shown in Figure 15. Since video stream data is transferred to the cloud only when moving objects are detected on the gateway side, network bandwidth and cloud storage can be saved.

[Prototype 3] Preventing workers from heatstroke

In summer, outdoor manual laborers are exposed to a risk of heatstroke due to both high temperatures and high-humidity. To prevent this, there is a rule on restricting continuous work according to the heat index called WBGT[11]; however, it is difficult for the field overseer to know the WBGT of the corresponding field site and the health conditions of all workers at all time. Therefore, we implemented a monitoring application which records video stream data that captures a specified field site when WBGT is above a threshold and reports to the field overseer when a worker stops moving.

Here, we adopted an existing IoT platform that can detect events according to the changes in time series data. The WBGT value, calculated from temperature and humidity using sensors installed at the field, is continuously registered to the IoT platform. Whenever the WBGT value goes above a specified threshold, the IoT platform calls the web API defined by *InputHttpEndpoint*. This cooperation makes it possible to control media service (e.g., start recording video stream data, start detecting moving objects from video stream data) according to the changes in sensory data (e.g., WBGT). At the same time, our framework notifies the result of moving object detection to the IoT platform using *OutputHttpEndpoint*, and the IoT platform can send warnings to the overseer and workers according to the result. *InputHttpEndpoint* and *OutputHttpEndpoint* make it easy to create multimedia IoT applications that cooperate with existing IoT platforms.

8 EVALUATION

We evaluated the effectiveness of our framework based on the above proof-of-concept prototypes from the perspectives as follows:

8.1 Complexity

To process single multimedia stream data cooperatively using multiple media servers, developers are required to establish a media session using RTP or WebRTC in addition to implementing originally required media processing. In addition, developers are required to take care of which endpoint to use and how to establish a media session, which varies with the device type installed at the field site. Our framework spares developers from such complexity, so IoT application developers can be dedicated to connecting devices at field sites with media processing modules.

[11] National Weather Service Weather Forecast Office: http://www.srh.noaa.gov/tsa/?n=wbgt

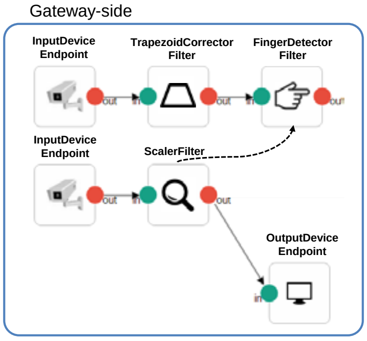


Figure 14: **[Prototype 1]** Media service description for lecture support

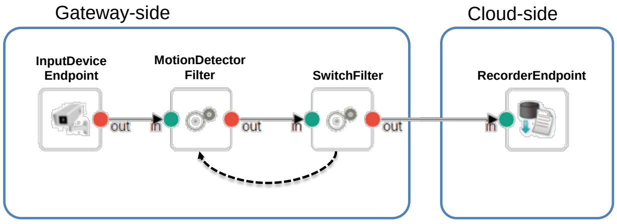


Figure 14: **[Prototype 2]** Media service description for suspicious person monitoring

8.2 Productivity

Table 3 shows the comparison of the number of program lines between cases using our framework and not. These numbers are counted without brackets. In the case of media service used by Prototype 2**,** our framework works effectively since developers are not required to write logic to establish the media session between the gateway and the cloud.

On the other hand, considering the case of Prototype 3, i.e., the case where media processing pipeline topology changes according to an event, we defined a single static media service description that includes *SwitchFilter* in the case of using our framework, while we implemented this switching as IoT application-side logic in the case not using our framework. As a result, the number of lines not using our frameworks is fewer in spite of writing logic to establish a session with the gateway and the device. Therefore, the current media service description format is not suitable for describing a dynamically changing pipeline. We need to consider an effective way to describe event-driven media services as a future work.

8.3 Ease of trial and error

It is important to repeatedly use trial and error for creating IoT applications. However, in multimedia application, logic to handle multimedia input and output are tightly-coupled with multimedia processing logic itself. As a result, when application developers modify their application, they are required to rewrite programming logic itself. For example, in order to change the behavior of Prototype 1 to put a circle on the fingertips, developers are required only to replace ScalerFilter with another (i.e., a filter to put marker) and do not need to deploy the application again by using our framework.

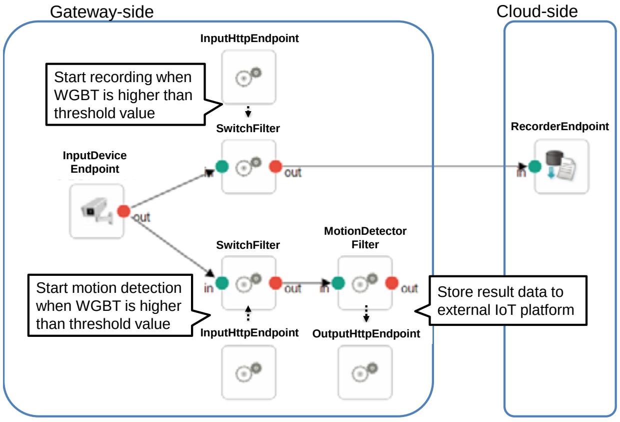


Figure 16: **[Prototype 3]** Media service description for heatstroke prevention

Table 3: Comparison of the number of program lines

	Used	Not used
Prototype 1	38	50
Prototype 2	26	42
Prototype 3	50	48

8.4 Division of labor

Considering the case of creating IoT applications such as **Prototype 1** using only Kurento Media Server, the application may not be adaptable to other classrooms, since the installed device type may differ from classroom to classroom. On the other hand, by using our framework, an IoT application once created can be adapted for any other classroom only if each device is registered in the same name owing to our framework's device virtualization mechanism. This makes it possible to create IoT applications independently from device installation.

In addition, not only cameras, microphones and screens, but also any software modules implemented to meet the specification of a virtualized device interface can be registered as a device. Developers can equally treat both cameras and screen capturer modules on our framework.

8.5 Cooperation with external systems

Our framework can deal with not only the case where a media stream data captured from a device is processed, recorded and transferred to another device, but also the case where an external system affects multimedia stream data using *InputHttpEndpoint* and notifies the external system when an event occurs using *OutputHttpEndpoint*. In the case of Prototype 3, we adopted an existing IoT platform, but we are not limited to such IoT platforms. For example, cooperating with the existing complex event processing system enables more advanced event detection by using both sensory stream data stream and multimedia stream data.

In other words, our framework can be used as an extension to make it easy to deal with media stream data on an existing IoT platform rather than a substitute. Currently, our framework supports cooperation using only HTTP requests, but we plan to implement endpoint modules for protocols other than HTTP (e.g., MQTT, WebSocket) to make it easier to create IoT applications that deal with both multimedia and sensory data streams.

9 SUMMARY

We designed and implemented a framework that makes it easy to deal with multimedia data such as audio and video generated from devices installed at field sites. The features of our framework are as follows:

- Virtualization of multimedia input/output devices
- Distributed execution of media service between gateway and a cloud
- Simple media service description using a graphical flow editor.

In this paper, we presented the three proof-of-concept applications we created and discussed the effectiveness of our framework from the perspective of complexity, productivity, and ease of trial and error.

As future work, we need to improve the media service description format and create endpoints other than HTTP. We plan to offer our framework to workshops and hackathons to verify the effectiveness of our framework from both qualitative and quantitative perspectives.

REFERENCES

[1] Akpinar, K., Hua, K. A., and Li, K., “ThingStore: A Platform of Internet-of-Things Application Development and Deployment,” ACM International Conference on Distributed Event-Based Systems (ACM DEBS 2015), pp.162-173, 2015.

[2] Alam, S., Chowdhury, M. M., and Noll, J., “SenaaS: An Event-Driven Sensor Virtualization Approach for Internet of Things Cloud,” IEEE International Conference on Networked Embedded Systems for Enterprise Applications (IEEE NESEA 2010), pp. 1-6, 2010.

[3] Blackstock M., and Lea, R., “Toward a Distributed Data Flow Platform for the Web of Things (Distributed Node-RED),” International Workshop on Web of Things (WoT 2014), pp.34-39, 2014.

[4] Fernandez, L., Diaz, M. P., Mejias, R. B., and Lopez, F. J., "Kurento: A Media Server Technology for Convergent WWW/Mobile Real-Time Multimedia Communications Supporting WebRTC," IEEE International Symposium and Workshops on World of Wireless, Mobile and Multimedia Networks (IEEE WoWMoM 2013), pp. 1-6, 2013.

[5] Munjin, D. and Morin, J. H., "Toward Internet of Things Application Markets." Proc. IEEE International Conference on Green Computing and Communication (IEEE GreenCom, 2012), pp.156-162, 2012.

TCP Simulation-based Evaluation Method for Network Capacity Planning

Ryuji Matsunaga, Shohei Mitsuya, Masaki Suzuki, Takeshi Kitahara,and Shigehiro Ano

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, Japan
ry-matsunaga@kddilabs.jp

Abstract – Recently, due to traffic growth including mobile and fixed service traffic, it is a high priority goal for communication network operators to provide high quality network service at low cost. However, if they provide high quality network services, they require more capital expenditure. To ensure quality of network service, it is significant to prevent traffic overload of the link from happening in the commercial network. So it is necessary to find an appropriate threshold of the link and keep traffic below the threshold level and to expand the link capacity before the load exceeds the threshold. On the other hand, traffic should be accommodated into the network as much as possible from the perspective of the capital expenditure. Thus finding an appropriate threshold is the main effort for network capacity planning so that the threshold should be set high as much as possible. The existing method, packet by packet queuing simulation is an effective approach for finding appropriate threshold. However, it is not sufficient to evaluate the actual network because in the existing method TCP flow's behavior which accounts for much of the actual traffic is not taken into account. Therefore, we propose a practical testing method using TCP simulation to evaluate the maximum traffic load satisfying the target level of quality. Specifically, in the proposed method, the TCP flow's retransmissions and RTTs factor into the calculation. We also evaluate this method using actual traffic data-set in an operator's network and had a positive result that the higher threshold could be set than the result of the existing method.

Keywords: capacity planning, traffic monitoring, queuing simulation, TCP simulation

1 INTRODUCTION

Communication network operators should be responsible for providing high quality network service with end users at law cost. However, recently due to huge traffic growth including mobile and fixed service, it is necessary to expand the network massively in order to provide high quality service. Which means that there exists trade-off between quality and cost. To ensure quality of network service, it is important to prevent traffic overload of the link from happening in the commercial network. So it is necessary to find an appropriate threshold of the link and keep traffic below the threshold level and to expand the link capacity before the load exceeds the threshold. Where the threshold is the proportion of the available link bandwidth to the max link bandwidth. For instance, if the threshold of the 100Gbps link is 50%, 50Gbps is the available bandwidth as the operator's policy. On the other hand, traffic should be accommodated into the network as much as possible from the perspective of the capital expenditure. Thus finding an appropriate threshold is the main effort for network capacity planning so that the threshold should be set high as much as possible. To solve this problem, it should be noted that the threshold value depends largely on traffic characteristics on the link. When the bursty traffic go through in the link, instantaneous peak load would be much higher than visible load which can be observed by a common network management system. In such a case, the threshold value should be set to low in order to avoid packet loss. Therefore detailed evaluation of each link is necessary. A lot of past work addressed self-similar and long-range dependent traffic in terms of statistical nature [e.g. 1], mathematical modeling [e.g. 2], and performance analysis [e.g. 3]. However, network topology or architecture, network usage, and commonly used applications on network are quite diverse recently, thus there are no generalized traffic models which can be applied to individual network operating and/or planning task. From the practical perspective, specific and concrete steps which are widely applicable to actual operations are important. The existing method of capacity planning to evaluate the maximum traffic load satisfying the target level of quality is an effective method. However the existing method, packet by packet queuing simulation is not sufficient to evaluate the actual network. Because in the existing method TCP flow's behavior which accounts for much of the actual traffic is not taken into account. Considering these background conditions, we propose a more practical method to evaluate the maximum traffic load satisfying the target level of quality. Specifically, in the proposed method, the TCP flow's retransmissions and RTTs factor into the calculation. We also evaluate this method using actual traffic data-set in an operator's network. In our proposed method, we had a positive result than the result of existing method. Which means that we had a higher threshold of the link capacity than ever.

2 RELATED WORKS

In many past works or studies, self-similar and log-range dependent traffic in terms of statistical nature, mathematical modeling, and performance analysis are addressed mainly. In our study, the existing method [4] have been used for many years to evaluate actual network and to find an appropriate threshold of the link capacity. For years some network operators apply this method and conduct capacity planning. However, recently due to huge traffic growth, network operators demand to expand the network and the links (e.g. 100Gbps or 40Gbps) and accommodate more traffic into the links than ever. Because from the viewpoint of the capital expenditure, setting of the high threshold of the link capacity is important challenge for them. The existing method is a packet by packet queuing simulation, discrete event-driven simulation of single server queue. A goal of this simulation is to solve the least required link capacity under the given condition that is targeting the packet loss ratio and buffer size of the nodes that exist at either end of that link. The targeting packet loss ratio is tolerated loss ratio of the communication network operators. The notations of R, B, X, m and p are defined as below.

R: output rate from the queue
B: buffer size of the node
X: input traffic to the queue
m: input mean rate to the queue
p: target packet loss ratio

Input traffic data includes packet size and arrival time of each packet. The data is calculated from the packet capture data. In the simulation, an event will be processed each time the packet #n (n = 1,…,N) arrives at the queue. Note that X consists of N packets. And the notation of t_n, l_n and $Q[n]$ are defined as below.

t_n: the time when the packet #n arrives
l_n: the length of the packet #n
$Q[n]$: the queue length at the time t_n

Fig.1 represents the relationship among the notations.Fig.2 shows an example of queuing process. In this example, packet #n enters into the queue but packet #(n+1) is dropped.

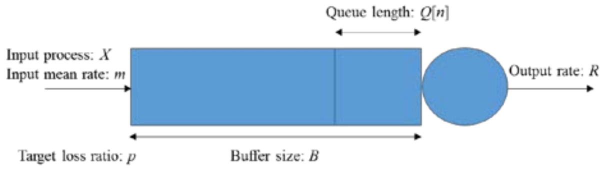


Fig.1 Notation of the simulation

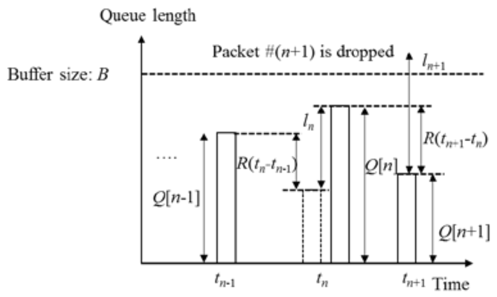


Fig.2 Example of the simulation process

In this existing method, the simulation is processed packet by packet and even if packet overflow is occurred and the packet is dropped, the retransmission is not occurred. Which means that TCP (Transmission Control Protocol) flow's behavior is not taken into account. So, it is not sufficient to evaluate actual network to find appropriate threshold of the link capacity. Because TCP flow traffic accounts for much of the actual traffic going through the network. In our data-set which is observed at one point of the actual network accommodates the internet service, TCP traffic accounts for about 70% in all traffic and the rest of them are almost UDP (User Datagram Protocol) traffic. To find an appropriate threshold of the link capacity suited to the actual network from the viewpoint of practical capacity planning, we should consider the TCP flow's behavior that is packet retransmission and RTT (Round Trip Time) and so on. By doing that we could find an appropriate threshold that is maybe higher than that resulted of the existing method. Because TCP has a network congestion avoidance algorithm and we could use network resource more effectively. For example, if a single packet is lost in a stream, TCP sender retransmits the packet. And another thing, the congestion avoidance algorithm is worked. If there is no congestion in the network and packet loss does not occurs, TCP sender assumes that there is enough resource in the network. So to use high bandwidth network more efficiently, a larger TCP window size may be used. On the other hand, when there is congestion in the network and packet loss occurs TCP sender uses a smaller TCP window size to avoid network congestion. Furthermore TCP sender uses a larger TCP window size when the measured RTT is lower than the calculated RTT. Other way round if the measured RTT exceeds the calculated RTT, TCP sender assume that the network is congested and a smaller TCP window size may be used.

3 PROPOSED METHOD

In order to address the challenges we propose a TCP-based evaluation method for network capacity planning. The point of this proposed method is using TCP flow's behavior simulation. We use NS-3 [5] to simulate the TCP flow's behavior. NS-3 is a discrete event network simulator. It is a free software targeted mainly for research and education. And it is publicly available for research, development and use. Fig.3 shows the overview of the proposed method. The proposed method, TCP-based evaluation method for network capacity planning is consists of 3 steps: (1) capture of the actual traffic in the network, (2) simulation using NS-3 of IPv4/v6 TCP and UDP and (3) output of RTT and throughput of the flows. Before we conduct the proposed evaluation method, we should capture the traffic data in the actual network which we would like to do capacity planning. We should capture the traffic data through the interface which may be a bottleneck in the network. Because the threshold of the link which is calculated in a bottleneck may be lower. To avoid quality degradation of network, it is important to set the lowest threshold in the network. We should make effort to calculate the highest threshold on that condition. The packets which is captured in the network consist of TCP,UDP/IPv4 and TCP,UDP/IPv6 packets. In the simulation step, we input the traffic data-set (pcap file) which is captured in the network. And we could input some parameters as below.

L: link bandwidth
B: buffer size of the node
p: target packet loss ratio

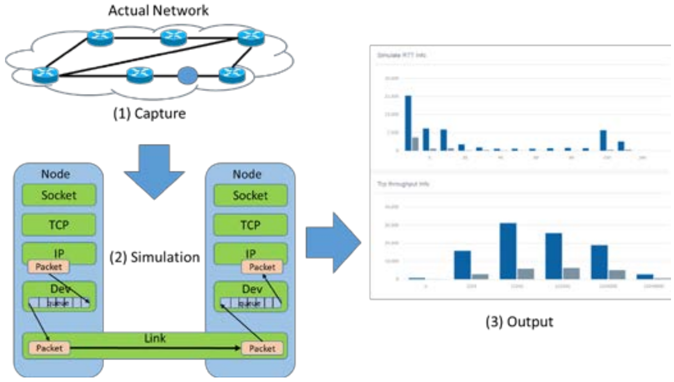


Fig.3 Proposed method

On the NS-3-based simulator we implemented TCP simulation (IPv4/IPv6) and UDP simulation (IPv4/IPv6). Moreover we implemented calculation feature of TCP RTT. As fig.4 depicts, we define the time from detection of the SYN packet to detection of the corresponding SYN ACK packet as the RTT of TCP sender. And another thing, we define the time from detection of the SYN ACK packet to detection of the corresponding ACK packet as the RTT of TCP receiver. The calculated RTT value is memorized per each TCP flow. In the case of UDP packets, it is impossible to calculate the RTT with only UDP packets. Because UDP uses connectionless transmission model and has no handshaking dialogues like as TCP. Therefore if there is TCP connection associated to the UDP, the RTT of TCP connection is calculated. When there is no TCP connection associated to the UDP flow, the RTT of the flow is set the average of the RTTs of all TCP flows.

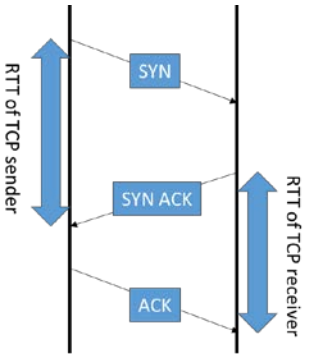


Fig.4 RTT of TCP sender and receiver

The TCP throughput of the flow is defined as below. We define the time from the arrival of SYN or SYN-ACK packet to the arrival of FIN or RST packet as the duration of the flow. The total count of the receive packets during the duration of the flow. The TCP throughput is obtained by dividing the number of bytes of the received packets by the duration time.

Packet loss behavior on the device’s queue goes as follows. Fig.5 depicts that the simulator has a queue intended to transmit. When the packet is sent to the device it put the packet into the queue. Next if the device could transmit the packet, it takes the packet from the queue and send to the link. When it put the packet into the queue and the overflow of the queue is occurred, the packet is lost. However the packet is retransmitted by the TCP retransmission behavior.

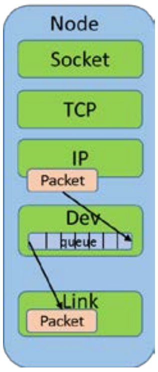


Fig.5 Transmission of the packet

Fig.6 depicts the packet drop by error model. When the device receives the packet from the link, it drops the packet by error model prepared on NS-3. NS-3 has error models as Table 1 shows.

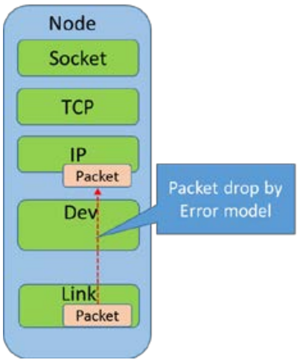


Fig.6 Packet drop by error model

Table 1: Error models

Items	Description
RateErrorModel	Packet loss by error rate
BurstErrorModel	Bursty packet loss by error rate
ListErrorModel	Specific packet loss by label
ReceiveListErrorModel	Packet loss at the receiver by arrival turn

Next, we describe about the link bandwidth (output rate) between the nodes. As Fig. 7 depicts, it simulate the link bandwidth by putting the delay according to the transmit rate when the packet go through from the sender to the receiver. And in the simulation we set the value of the link bandwidth to find an appropriate link bandwidth that is the threshold of the link. As we discuss later, by the simulation we have the throughput and RTT information per flow. Thus we could find a more appropriate threshold based on careful study of the customer quality of experience.

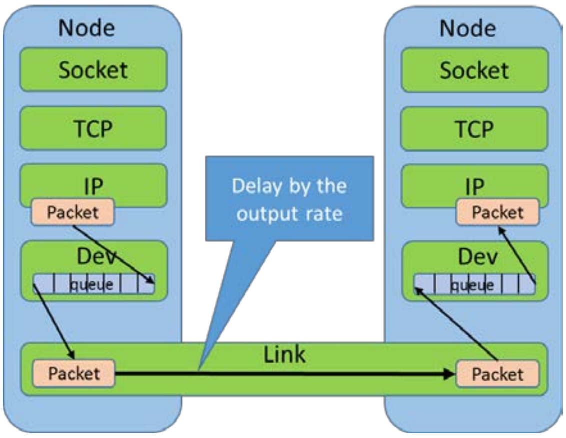


Fig.7 Simulation of the link bandwidth

Finally, we discuss about the output of the proposed method. After the simulation we have the statistical information of the simulation, throughput and RTT per flow. The statistical information of the simulation includes the items as Table 2 shows. And the statistical information of the throughput includes the throughput per TCP and UDP flow as Table 3 shows. The throughput is based on the user data length differently from the statistical information of the simulation and the unit is Mbps. As Table 4 shows, the statistical information of RTT includes the distribution per IPv4 and IPv6.

Table 2: The statistical information of the simulation

Items	Description
Benchmark:	Time taken to analyze 1 sec data
Pcap:	Total throughput of the data
Pcap:Tx:	Tx throughput of the data
Pcap:Rx:	Rx throughput of the data
Pcap:ParseFailed:	The throughput of the control packet
NS3:Rotuer:Tx:	Tx throughput of the Router
NS3:Router:Rx:	Rx throughput of the Router
NS3:Router:IPv4:Tx|Rx:	IPv4 Tx|Rx throughput of the Router
NS3:Router:IPv6:Tx|Rx:	IPv6 Tx|Rx throughput of the Router
NS3:Router:TCP:Tx|Rx:	TCP Tx|Rx throughput of the Router
NS3:Router:UDP:Tx|Rx:	UDP Tx|Rx throughput of the Router

Table 3: The statistical information of throughput

Items	Description
tcp-throughput	TCP throughput from establishment to closing of the session
(time)-tcp.csv	TCP throughput per 1 sec
(time)-udp.csv	UDP throughput per 1 sec

Table 4: The statistical information of RTT

Items	Description
rtt-(IPv4|IPv6).csv	The distribution of RTT per IPv4 and IPv6
(time)-rtt-(IPv4|IPv6).csv	The distribution of RTT per 1 sec

4 EVALUATION

We evaluated the proposed method using actual traffic data-set. The data-set is measured in the 10Gbps-based metro network accommodating the consumer service in Japan. In the proposed method, we focus on the RTT and throughput information because these factors are important to improve the customer quality of experience. Firstly Fig.8 and Fig.9 depict the result of the simulation, RTT and throughput. In the simulation the link bandwidth is set so that the threshold of the link will be 50%. That is to say that there is no congestion and enough resource in the network. In this case the result of the simulation is served as useful reference to find an appropriate threshold.

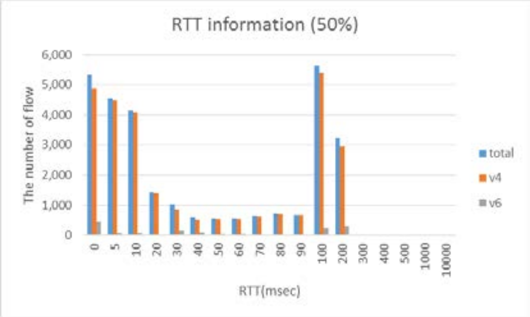


Fig.8 RTT distribution of the threshold 50%

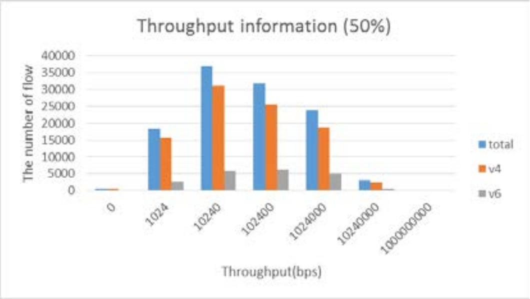


Fig.9 Throughput information of threshold 50%

Next, Fig.10 and Fig.11 show the result of the simulation that the link bandwidth is set so that the threshold of the link will be 85%. In our experience and our study, the actual network operated based on like this threshold of the link may be stable from the network resource perspective.

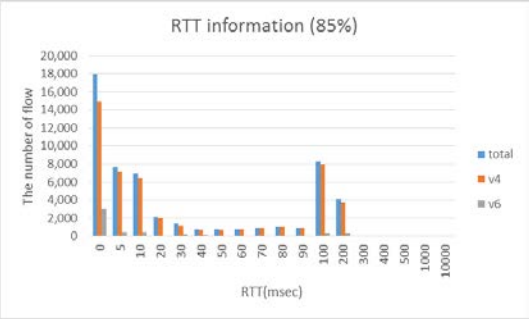


Fig.10 RTT information of threshold 85%

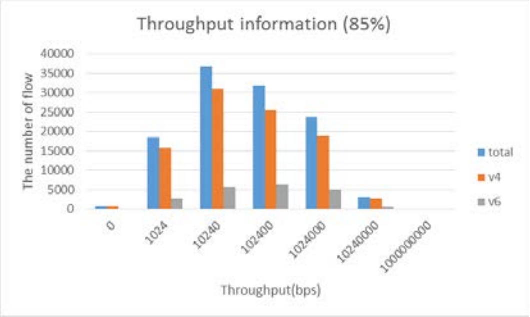


Fig.11 Throughput information of threshold 85%

Fig.12 and Fig.13 depict the result of the simulation that the link bandwidth is set so that the threshold of the link will be 90%. Maybe the operators may set the threshold of the link to around 90%. Actually, by the existing method using the same data-set, the threshold of the link is calculated as around 90% with targeting packet loss ratio to 10^(-5) . In the result, what is important is that there is little difference between the throughput distribution of threshold 50% and that of threshold 90%.

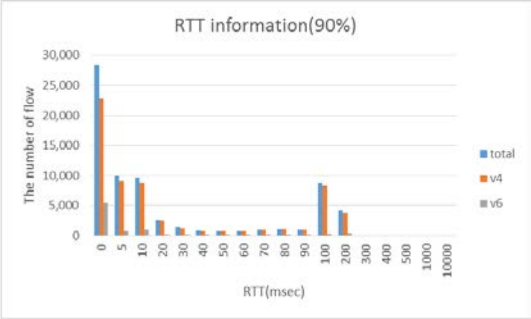


Fig.12 RTT information of threshold 90%

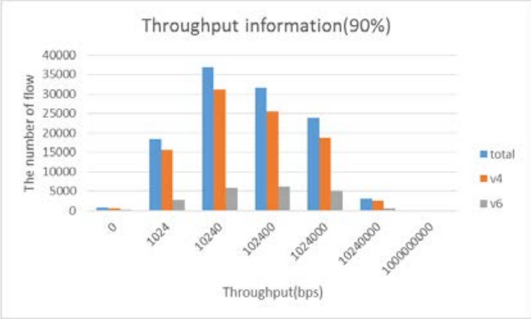


Fig.13 Throughput information of threshold 90%

Finally Fig.14 and Fig.15 show the result of the simulation that the link bandwidth is set so that the threshold of the link will be 95%. When the operational threshold of the link is around 95%, it is beneficial from the viewpoint of the capital expenditure. Additionally the distribution of RTT and throughput in this case are little difference from previous described results.

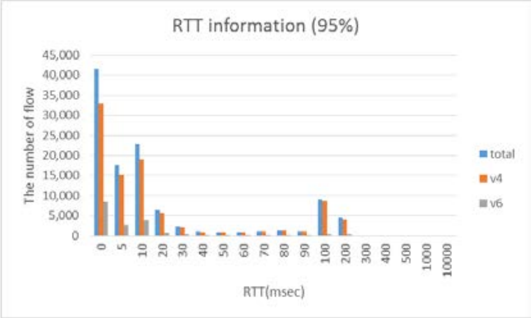


Fig.14 RTT information of threshold 95%

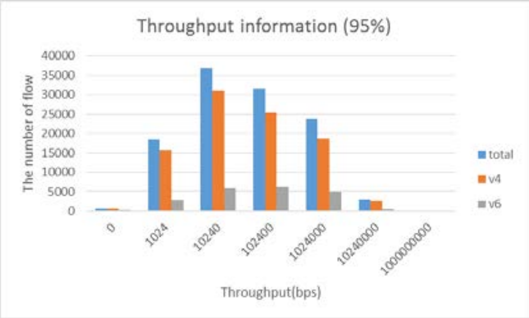


Fig.15 Throughput information of threshold 95%

However, by the existing method, packet by packet simulation with the same data-set, we had the correlation packet loss ratio and the threshold of the link as Fig.16 depicts. If the network operator’s targeting packet loss ratio is 10^(-5), they decide the threshold is 90% which is satisfying the network quality target. Because on the other hand, in the existing method when the threshold is 95% the packet loss ratio is not satisfying the target packet loss ratio. However as we mentioned above, if the network operator could set higher threshold, it is very beneficial from the viewpoint of capital expenditure. Thus based on the proposed method result, the network operators could decide to set the threshold value 95%. Because considering the TCP flow’s behavior, the TCP throughput or RTT more strongly affected the customer quality of experience than the simple packet losses.

However it is not easy to decide to the threshold of the link because there is not purely theoretical methods to calculate the appropriate threshold of the link from many factors. We consider that for network operators using the existing method and the proposed method concurrently is an effective. Specifically after we got the threshold by the existing method, we simulate by the proposed method using the link bandwidth so that the threshold of the link will be some percent higher than the previous threshold. In the result if the throughput and the RTT of the higher threshold are little differently from the result of the simulation with the previous threshold, we could propose the higher threshold for the network operators.

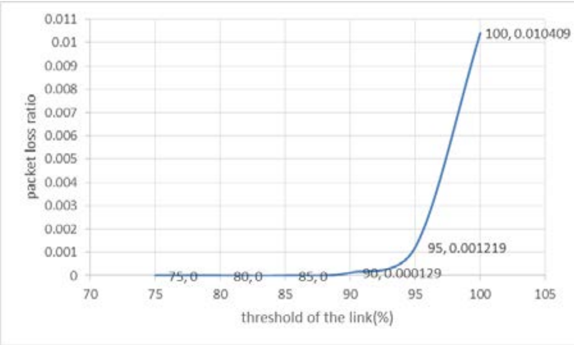


Fig.16 correlation of packet loss ratio and threshold of the link by existing method

5 CONCLUSION

In order to enable communication network operators to find an appropriate threshold of the network links so that the network operators provide stable and high quality network service, this paper proposed a TCP-based evaluation method of capacity planning.

We evaluated the proposed method using actual traffic data-set which is observed in the actual metro network.

We confirm the practicality of the proposed method. By this method we introduce the process of finding an appropriate threshold of the link from the viewpoint of the customer quality of experience. That is to say the proposed method is taken account into TCP flow behavior.

Since the proposed method is quite specific and practical, the network operators could easily apply this method to their work to find an appropriate threshold of the network link. Additionally we believe that by this proposed method they could improve the quality of their capacity planning work and it could contribute to the reduction in capital expenditure.

6 REFERENCES

[1] Leland, W.E., et al.: On the self-similar nature of Ethernet traffic (extended version). In:IEEE/ACM Trans. On Networking, Vol.2, No. 1.(1994)

[2] Taqqu, M.S., et al.: Proof of a fundamental result in self-similar traffic modeling. In: ACM SIGCOMM Computer Communication Review, Vol.27, No.2.(1997).

[3] Erramilli, A., et al.: Experimental Queuing Analysis with Long-Range Dependent Packet Traffic. In:IEEE/ACM Trans. On Networking, Vol. 4, No. 2.(1996)

[4] Takeshi,K et al.: A practical Evaluation Method of Network Traffic Load for Capacity Planning, In: Testing Software and Systems, 27th IFIP WG 6.1 International Conference, ICTSS 2015, Sharjah and Dubai, United Arab Emirates, November 23-25, 2015, Proceedings, pp263-268(2015)

[5] ns-3 project : https://www.nsnam.org/

A Proposal of Malicious URLs Detection based on Features Generated by Exploit Kits

Yuma Sato[†], Yoshitaka Nakamura[‡], Hiroshi Inamura[‡] and Osamu Takahashi[‡]

[†]Graduate School of Systems Information Science, Future University Hakodate, Japan
[‡]School of Systems Information Science, Future University Hakodate, Japan
{g2115016, y-nakamr, inamura, osamu}@fun.ac.jp

Abstract - With the spread of Web access, cyber attacks are increasing. Drive-by Download attack is a kind of cyber attacks which may happen when visiting a website. Drive-by Download attacks redirect Web users to malicious Web pages. Drive-by Download attacks force Web users to download malware by exploiting the vulnerabilities of Web browsers or plug-ins when these users visit malicious Web pages. Attackers use heavily Exploit Kits to build Web sites for Drive-by Download attacks. Some characteristics such as string length and the number of special symbol, depending on the types of Exploit Kit are seen in the URLs of malicious Web pages used for these attacks. In addition, domain name of malicious Web sites tends to be short-lived to avoid blacklisting. Therefore, it is difficult to detect these attacks by using blacklists of URLs. However, the characteristic of the path and query of URLs does not change if an attacker does not change Exploit Kit to use. Therefore we can detect an attack from these characteristic of the path and query of URLs even if the attacker changed the domain name of the Web sites to use for attacks. These characteristics are extracted by decision tree learning. In this paper, we propose a novel malicious URLs detection method of Drive-by Download attacks by using features of Path and Query components of URLs used in Exploit Kits.

Keywords: Drive-by Download Attacks, Web Security, Malware, URLs, Exploit Kits

1 INTRODUCTION

The threats of cyber attacks are increasing with the spread of using the Web. The Drive-by Download attack is one of the cyber attacks through the Web. And it is increasingly sophisticated and becomes the large threat in late years. Drive-by Download attack forces Web users to download malware unconsciously. Figure 1 shows number of detected cases reported by IBM TOKYO SOC Report[1]. It shows that 2,740 of Drive-by Download attacks are detected in the first half year period of 2015. Furthermore, 800 attacks are detected in all half year periods from 2013.

Drive-by Download attackers compromise legitimate Web sites and embed malicious contents[2]. Attackers guide Web users from legitimate Web sites to the malicious Web sites and transmit the malware to PC of users. Generally these attacks exploit software of Web users and transmit the malware. It is difficult to detect this type of attack because malware is downloaded without any user noticing. Some kind of script code is used in Drive-by Download attacks. Script codes are almost always obfuscated. In addition, the detection by IPS(Intrusion Prevention System) becomes difficult because the obfuscated code does not appear in the attack patterns memorized in IPS. In this way, Drive-by Download attacks are being sophisticated and complex during recent years.

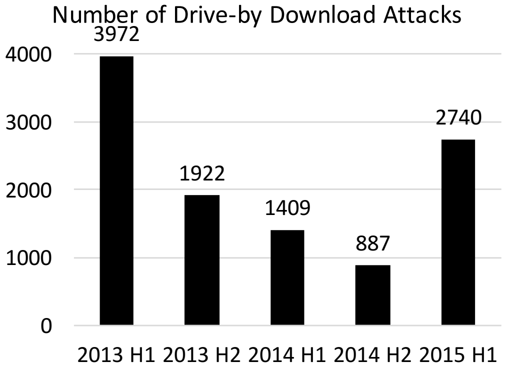


Figure 1: Number of Drive-by Download Attacks

Figure 2 shows typical flow of a Drive-by Download attacks. Firstly, attackers tamper with legitimate Web pages in order to redirect Web users to intermediate sites which guides Web access to the malicious sites. Secondly, Web users visit the compromised Web pages. Web users are redirected to multiple intermediate sites by compromised redirection. Usually, there are multiple redirections to make attack detection difficult. After multiple redirections, Web users are redirected to the exploit sites and the malware download sites. The exploit site attacks the vulnerabilities of operating system, Web browser, and plug-ins of Web browser. Finally, Web accesses of users are redirected to the malware download sites and distribution sites transmit malware to PC of Web users.

Recently, attackers often use Exploit Kits at the time of Drive-by Download attack[3]. An Exploit Kit is the packaged tool kit consisting of some exploit codes which can exploit various type of vulnerabilities. Exploit codes attacking newly discovered vulnerabilities are added to Exploit Kit continuously. Exploit Kits can be managed by GUI, and the user without the technical knowledge can make effective attacks easily.

In the Drive-by Download attack, a fingerprinting technology distinguishing the environmental information of Web users is used. Because many exploit codes are included in Exploit Kit, the attacker can use code fitted to the environment of each Web users. Therefore if there is even one vulnerability in the environment of the Web users, the attack aimed at the vulnerability succeeds, and malware is downloaded to the user's terminal.

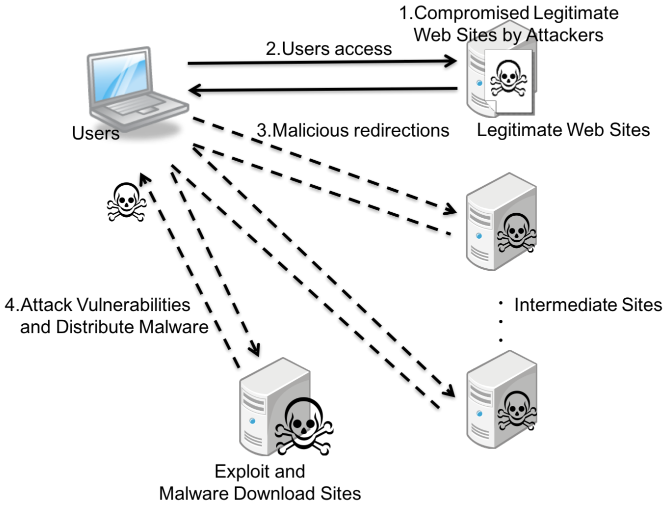

1.Compromised Legitimate Web Sites by Attackers
2.Users access
Users
3.Malicious redirections
Legitimate Web Sites
Intermediate Sites
4.Attack Vulnerabilities and Distribute Malware
Exploit and Malware Download Sites

Figure 2: Typical Drive-by Download Attacks Flow

There are some types of Exploit Kits, and each Kits have features in generated URL. Web user can detect the pattern of the Drive-by Download attack by using these features and can prevent downloading of malware.

In this paper, we propose a detection method of malicious URLs of the Drive-by Download attacks based on features of URLs generated by Exploit Kits. As a contribution, even if the domain name of the website changed in a short term, our approach can detect malicious Web sites. Our approach does not need the blacklist management and can detect malicious Web sites at lower cost. And the proposed method shows possibility of the detection using URL's path and query information.

2 RELATED WORK

There are some approaches to detect Drive-by Download attacks.

Some approaches are blacklist type method. Google Safe Browsing is a typical example of such approaches. Generally, the blacklist is generated by 3 steps. First step is to trace links included in the Web page using Web crawler. Second step is to extract URLs considered to be malicious based on the features of HTML codes and URLs. As the last step, analyzers actually access Web pages with URLs considered to be malicious and evaluate the codes of HTML and JavaScript of Web pages and convert into scores. However, construction and maintenance of the blacklist requires large cost because domain name of malicious Web pages change frequently. To solve this problem, Invernizzi proposed low-cost construction method of blacklist by efficient Web crawling [4].

However, domain name of malicious Web sites tend to be short-lived to avoid blacklisting. Reference[5] shows that 40% of malicious URLs are changed within one month. And Ref.[6] describes that 80% of domain names of the malicious site are not used in six months. If attackers change domain name of malicious site, blacklisting can't detect attacks such as Drive-by Download attacks because changed domains is not listed. It is difficult to perfectly detect by blacklists. As a result, the blacklist-type detection becomes more difficult because the update of the blacklist does not catch up with the changes of domain names if attackers frequently change the domain names of the malicious sites.

Some methods are proposed to prevent Drive-by Download attack without blacklists. One method of Ref.[7] is detecting Drive-by Download attack by analyzing attack script codes and extracting features of frequently appearing characters used in codes. Because algorithm used in attack code varies according to the kinds of Exploit Kit, this feature can be used for the detection. It can detect attacks by these differences. Reference[8] introduce the method to detect malicious Web sites used for attacks by monitoring Web communication log, and the method to detect attacks by analyzing HTTP header information.

3 PROPOSED METHOD

3.1 Basic Concept

There are some features in the Web pages generated using Exploit Kit in Drive-by Download attacks. At first there is feature that URLs generated by Exploit Kits are longer than those of legitimate Web page. And different features appear in these URLs depending on Exploit Kits used for genetation. It is known that these URLs can be detected by regular expressions.

In this paper, we propose a method to detect URLs of malicious Web pages based on features of URLs generated by such Exploit Kits. Our approach focuses on URLs path and query information as features. If the Exploit Kit using for generation is the same, the features of the path and query of URLs do not have any change even if attackers change domain name of malicious sites frequently. These features can be used for the detection of malicious sites. Therefore we use features of URL paths and queries for the detection of malicious sites in the Drive-by Download attack. Our method vectorizes the features of paths and queries of URLs used in malicious URLs, and constructs the decision tree using these vectors. In this way, we can detect malicious URLs of Drive-by Download attacks.

3.2 URL's path and query

In this paper, "URL's path and query" is defined as the string which combined path of URL with query of URL using character "?".

For example, in the URL of "http://www.example.com/dir/file.html?key=value", part of "dir/file.html" becomes the path of this URL, and part of "key=value" becomes the query of this URL. In other words this URL's path and query becomes "dir/file.html?key=value".

3.3 Binary decision tree

Binary decision tree is classifier that classify input data into premade classes. It has leaf nodes, root nodes, and internal nodes. Leaf nodes represent the premade multiple classes. Internal nodes and root nodes except leaf nodes represent the test for input data. Binary decision tree can classify which class the input data applied to by repeating tests for the input data from the root node to a leaf node.

Weka (Waikato Environment for Knowledge Analysis) is open source software for machine learning, which was made by machine learning group at the University of Waikato[9]. Weka has many functions such as data preprocessing and data visualization in machine learning.

Our method uses J48 classifier which is an implementation of C45 algorithm developed by Quinlan[10] for the generation of binary decision tree.

3.4 Detection method of malicious URLs

3.4.1 Detection process

The proposed method uses URL's path and query to detect malicious URL. As preparations, the method needs to collect the information of legitimate URLs and malicious URLs beforehand. Firstly the method extracts URLs that occurred by communication from the communication data which accessed to the legitimate Web pages, and from attack communication data of Exploit Kit. Secondly, URL's path and query is extracted from these legitimate and malicious URLs. Thirdly, these URL's path and query are vectorized using multiple features of malicious Web sites like the next subsection. Binary decision tree for classifications to detect malicious URLs is made by these vectors and C4.5 algorithm. This decision tree classifies whether the URL was generated by Exploit Kits. By using this result the proposed method can determine whether the URL is malicious or the legitimate.

3.4.2 Vectorization of URL's path and query

The extracted URL's path and query is converted into a vector using item (1) to item (9) of Table 1.

In Ref.[11], L. Xu et al. describe that the average length of the legitimate URL is 18.23, and the average length of the malicious URL is 25.11. They also describe that long and random character string tends to be used in malicious URL. From this result, we use the length of the URL's path and query as a component of the vectors (item (1)).
Reference[11] also describe that the average number of special symbols of the legitimate URL is 3.36 and the average number of special symbols of the malicious URL is 2.93. From this result, we use the number of special symbols as a component of the vectors (item (2)).

In Ref.[12], J. Ma et al. describe that the longest and average path lengths of URLs are available as malicious detection factor. Therefore, path length is used as a component of the vectors (item (3)). And they also describe that the number of digits included in URLs is available as malicious detection factor. Therefore, number of digits is used as a component of the vectors (item (3)).

Because the number of alphabets and keys in query are thought to increase in malicious URLs by features of item (1), these parameters are used as component of the vectors (item(6), (7)).

Because many redirections are used in the Drive-by Download attack, redirection URL may be included in query. Therefore we use the information whether the query includes character string "http" or whether the query includes IP address as components of the vectors (item(8), (9)).

Table 1: Vector Components

Item	Vector Components	Values
(1)	Length of URL's path and query	Integer
(2)	Number of special symbols	Integer
(3)	Path length	Integer
(4)	Number of digits	Integer
(5)	Query length	Integer
(6)	Number of alphabets	Integer
(7)	Number of keys in query	Integer
(8)	Inclusion of string "http"	0 or 1
(9)	Inclusion of IP addresses	0 or 1

4 EXPERIMENTAL EVALUATION

4.1 Overview of experimentation

This experiment intends for only a URL query path with disregard to the operation of the Web browser by the user. We extracted only URL's path and query in a text file and use the text file as input data.

In this experiment, we make the binary decision tree which classifies legitimate URL and malicious URL according to the kind of Exploit Kit. The construction of the decision tree uses C4.5 algorithm. Specifically, we use J48 implemented in Weka. This experiment evaluate the proposed method with 10-fold cross validation by using constructed decision tree for input data.

4.2 Experimental data

URLs of malicious and legitimate communication data are used as experimental data. About the malicious data, we use the PCAP type format data including in Malware - Traffic - Analysis.net from 2013 through 2015[13]. These data include malicious and legitimate communication data, and downloading communication data of malware. The number of the URL's path and query included in the experimental data is 7,212. We also extract the set of Exploit Kits from the same PCAP data. These Exploit Kits are classified using file name. PCAP data is named each Exploit Kits used. The types and number of Exploit Kits included in the experimental data are as follows.

We use DMOZ as legitimate communication data[14] because data have many URLs included many path and query. DMOZ is the largest Web directory that constructed and maintained by a global community of volunteer editors. In this experiment, we extract 300 of URLs of Web pages indexed in DMOZ at random. We randomly extract 2,368 of URL's paths and queries as legitimate data from the URL's path and query of the request URL occurred at the time of access to these URLs.

Table 2: Types and Number of Exploit Kits

Types	Number
Angler Exploit Kit	1,947
Blackhole Exploit Kit	39
Cool Exploit Kit	18
Dotkachef Exploit Kit	54
Fiesta Exploit Kit	1,071
Flashpack Exploit Kit	225
Goon Exploit Kit	225
Hello Exploit Kit	9
KaiXin Exploit Kit	18
Magnitude Exploit Kit	1,128
Neutrino Exploit Kit	408
Nuclear Exploit Kit	1,314
Rig Exploit Kit	387
Styx Exploit Kit	135
Sweet Orange Exploit Kit	234

4.3 Evaluation process

In this experiment, we evaluate the detection of the proposed method using five evaluation items such as true positive rate ($TPrate$), false negative rate ($FNrate$), true negative rate ($TNrate$), a false positive rate ($FPrate$), and accuracy (ACC). We evaluate every URL's paths and queries included in a request to any Web page. In the following formulas, "URL's P & Q" denotes to "URL's path and query".

When the proposed method correctly classifies the URL's path and query of the malicious URL as malicious, we define it as true positive. It is true positive if malicious URLs are classified as any Exploit Kits. The true positive rate is a ratio of true positive number in all malicious URL's paths and queries. This rate is expressed as formula (1).

$$TP\ rate\ = \frac{\#\ of\ malicious\ URL's\ P\&Q\ classified\ as\ malicious}{\#\ of\ malicious\ URL's\ P\&Q} \tag{1}$$

False negative is that malicious URL's paths and queries incorrectly classified as legitimate. The false negative rate is a ratio of false negative number in all malicious URL's paths and queries. This rate is expressed as formula (2).

$$FN\ rate\ = \frac{\#\ of\ malicious\ URL's\ P\&Q\ classified\ as\ legitimate}{\#\ of\ malicious\ URL's\ P\&Q} \tag{2}$$

True negative is legitimate is classified as legitimate. True negative is that legitimate URL's paths and queries correctly classified as legitimate. The true negative rate is a ratio of true negative number in all legitimate URL's paths and queries. This rate is expressed as formula (3).

$$TN\ rate\ = \frac{\#\ of\ legitimate\ URL's\ P\&Q\ classified\ as\ legitimate}{\#\ of\ legitimate\ URL's\ P\&Q} \tag{3}$$

False positive is that legitimate URL's paths and queries incorrectly classified as malicious. The false positive rate is a ratio of false positive number in all legitimate URL's paths and queries. This rate is expressed as formula (4).

$$FP\ rate\ = \frac{\#\ of\ legitimate\ URL's\ P\&Q\ classified\ as\ malicious}{\#\ of\ legitimate\ URL's\ P\&Q} \tag{4}$$

The accuracy is defined as a ratio of true positive and true negative number in all URL's paths and queries. This rate is expressed as formula (5).

$$ACC\ = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

5 RESULT AND DISCUSSION

5.1 Result

Table 3 shows the result of classification using the proposed method by each evaluation item. Table 4 shows the classification result by Exploit Kits. From this result, the classification of the Exploit Kits with much used number achieves high detection precision. On the other hand, the classification of the Exploit Kits with a little used number tends to have a low detection precision.

Table 3: Result of Classifying

Evaluation Items	Value
True Positive (TP)	89.02%
False Negative (FN)	10.98%
True Negative (TN)	81.67%
False Positive (FP)	18.33%
Accuracy	87.20%

5.2 Discussion

Our proposed method has a strong point to be able to detect the URL of malicious Web page, even if the domain name of the malicious Web site is changed, because our method focuses on URL's path and query occured in Exploit Kits.

From the result of experiment, the more number of URLs are generated by the Exploit Kits, the higher detection precision is achieved. In order to increase the detection precision of malicious URLs, it is necessary to collect much more access data made by Exploit Kits.

Our approach only vectorize from path and query in URLs. For instance, when URLs path and query is "/", vector of legitimate URLs may be the same as vector of malicious. If it comes to that, legitimate URLs are classified as malicious and vice versa. For this problem, in order to block exploit attacks and downloading malware precisely, we need to make improvement adding some other components such as the number of redirections. For the future, we are going to enhance the capability to detect by classify in detail URLs generate by Exploit Kits.

Table 4: Result of Classifying by Exlopoit Kits

			Number of Classifying															
			A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Exploit Kits	A	Angler	1572	0	0	1	9	3	5	0	2	77	9	100	0	3	0	166
	B	Blackhole	0	25	0	0	0	0	3	0	0	0	0	0	0	0	0	11
	C	Cool	0	0	5	0	0	2	0	0	0	0	0	0	0	0	0	11
	D	DotKachef	3	0	0	36	0	0	3	0	0	0	0	0	0	0	0	12
	E	Fiesta	10	0	0	0	889	2	12	0	0	34	2	37	6	0	0	79
	F	Flashpack	1	0	2	0	3	139	3	0	0	1	15	5	0	0	2	54
	G	Goon	12	0	0	1	9	1	105	0	1	12	2	26	11	0	3	42
	H	Hello	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	3
	I	KaiXin	3	0	0	0	0	0	0	0	3	0	0	1	0	0	0	11
	J	Magnitude	86	0	0	0	27	1	15	0	0	859	6	54	6	2	0	72
	K	Neutrino	25	0	0	0	4	11	3	0	0	11	237	57	0	0	2	58
	L	Nuclear	73	0	0	0	28	2	8	0	0	55	30	1013	5	0	2	98
	M	Rig	5	0	0	1	12	0	9	0	0	7	1	17	304	3	0	28
	N	Styx	5	0	0	0	2	0	0	0	0	6	0	1	2	70	0	49
	O	Sweet Orange	5	0	0	0	3	0	3	0	0	1	0	4	2	0	118	98
Legitimate	P	Legitimate	97	3	4	5	53	26	19	1	0	61	37	82	13	13	20	1934

6 CONCLUSION

In this paper, we proposed malicious URLs detection method based on features generated by Exploit Kits by using URLs path and query. This method builds binary decision tree from practical communications, and classifies malicious or legitimate URLs with paths and queries. From the experimental evaluation, the method achieved 89.02% of true positive rate and 81.67% of true negative rate.

As a future work, we will classify Exploit Kits in detail and need propose a detection method in extreme precision with string pattern or regular expressions.

REFERENCES

[1] IBM, The First Half of 2015 Tokyo SOC information analysis Report, Available: https://www-304.ibm.com/connections/blogs/tokyo-soc/resource/PDF/tokyo_soc_report2015_h1.pdf?lang=ja.

[2] T. Matsunaka, A. Kubota, and T. Kasama, An Approach to Detect Drive-By Download by Observing the Web Page Transition Behaviors, Proceedings of the 9th Asia Joint Conference on Information Security (ASIA JCIS2014), pp. 19–25 (2014).

[3] Trend Micro Incorporated., 3Q 2015 Security Roundup. Available: http://www.trendmicro.co.jp/cloud-content/jp/pdfs/security-intelligence/threat-report/pdf-sr2015q3-20151119.pdf?cm_sp=threat-_-sr2015q2-_-lp-btn.

[4] L. Invernizzi, and P. M. Comparetti, EvilSeed: A Guided Approach to Finding Malicious Web Pages, Proceedings of the 2012 IEEE Symposium on Security and Privacy, pp. 428–442 (2012).

[5] M. Akiyama, T. Yagi, and M. Itoh, Searching structural neighborhood of malicious URLs to improve blacklisting, Proceedings of the IEEE/IPSJ 11th International Symposium on Applications and the Internet(SAINT2011), pp. 1–10 (2011).

[6] M. Akiyama, T.Yagi, and T. Hariu, Measuring Lifetime of Malicious Website Based on Redirection from Compromised Websites, The Special Interest Group Technical Reports of IPSJ, Vol. 2014-SPT-8, No. 10, pp. 1–6 (2014).

[7] M. Cherukuri, S. Mukkamala, and D. Shin, Similarity Analysis of Shellcodes in Drive-by Download Attack Kits, Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing(CollaborateCom2012), pp. 687–694 (2012).

[8] T. Matsunaka, A. Kubota and T. Kasama, An Approach to Detect Drive-by Download by Observing the Web Page Transition Behaviors, Proceedings of the 9th Asia Joint Conference on Information Security (AsiaJ-CIS2014), pp. 19–25 (2014).

[9] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Available: http://www.cs.waikato.ac.nz/ml/weka/index.html.

[10] J. R. Quinlan. C4.5: Programs for Machine Leaning. Morgan Kaufmann, (1993).

[11] L. Xu, Z. Zhan, S. Xu, and K. Ye, Cross-layer detection of malicious websites Proceedings of the third ACM conference on Data and application security and privacy (CODASPY'13), pp. 141-152, (2013).

[12] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious urls, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD'09), pp. 1245-1254, (2009).

[13] Malware-Traffic-Analysis.net, http://www.malware-traffic-analysis.net/.

[14] DMOZ - the Open Directory Project. https://www.dmoz.org/.

A Fast EDA-based Bayesian Network Learning Algorithm Using GPU Computation

Takashi Mori[†*], Yuma Yamanaka[†], Takatoshi Fujiki[†], and Takuya Yoshihiro[‡**]
[†]Graduate School of Systems Engineering, Wakayama University, Japan
[‡]Faculty of Systems Engineering, Shizuoka University, Japan
[*] s171053@sys.wakayama-u.ac.jp
[**] tac@sys.wakayama-u.ac.jp

Abstract - In this paper, we present a fast EDA (Estimation of Distribution Algorithm)-based algorithm to learn Bayesian Network structure accelerated by GPU computation. Bayesian Network is a graphical model that expresses causal relationship among events, which is useful in decision making in various practical scenes. A number of algorithms to learn Bayesian Network structure from data have been proposed so far, but since the problem to learn Bayesian Network structure is proved to be NP-hard, it takes considerable time to learn sub-optimal structures. However, recently, EDA-based genetic algorithms are reported to be an efficient approach for this problem. In this paper, we propose a method to extend an EDA-based structure-learning algorithm to achieve far faster execution by using GPU acceleration. Through evaluation, we show that the proposed algorithm runs about 14-times faster than the original one executed on CPU.

Keywords: Bayesian Networks, GPU, EDA, PBIL

1 INTRODUCTION

Bayesian Networks (BNs) are well-known graphical models used to analyze causal relationship among events. There are so many practical fields in which Bayesian Networks are effectively utilized, such as bioinformatics research, medical analyses and diagnosis, computer security, system diagnosis and monitoring, etc. Recently, we are surrounded by so much data coming from the Internet, sensors embedded to the environment, or various information systems, the importance of Bayesian Networks as analytic tools is continuously growing larger.

Considerable amount of researches have been dedicated to learn good BNs efficiently in the literature. It is well known that BN structure learning can be formulated as an optimization problem that optimizes a model score defined as an information criterion. However, because it is proved to be NP-hard [1], the constraint of *variable order* to reduce the search space is introduced in the early years by a heuristic method called K2 [2]. The *variable order* is a constraint on events $n_1, n_2, \ldots, n_k$ where n_i can be a parent of n_j only if $n_i \prec n_j$. However, there are many cases in practice in which order constraint is not applicable.

To solve the optimization problem without order constraint, many studies tried to find sub-optimal BN models. Recently, Algorithms based on genetic algorithms (GAs) are well-studied as shown in the survey article [3]. Among them, we in this paper focus on a kind of GA-based algorithm called EDA (Estimation of Distribution Algorithms) in which the distribution of model scores on the graph space is estimated in order to find better-score Bayesian Network models efficiently. specially, we treat a EDA-based algorithm called PBIL (Probability-Based Incremental Learning)[4], which is reported to be the best to learn BNs among EDAs [5]. The problem here is that the above algorithms including PBIL-based one take significant time to learn BNs. To cope with the large data available today, acceleration of those algorithms to run within shorter time is important.

In this paper, we present a method to accelerate a PBIL-based BN-learning algorithm to run much faster using GPU computation. In evaluation, we achieved about 14-times faster running speed than the original CPU-executed algorithm using consumer-class GPU hardware. From above, we conclude that the algorithm proposed in this paper computes good solutions within a short time by utilizing a GPU.

This paper is organized as follows. In Sec.2, we introduce the optimization problem to learn BN structures and give the specific description of PBIL-based BN-learning algorithm. After we concisely explain the GPU architecture related to our work in Sec.3, we describe the proposed method to accelerate the PBIL-based algorithm using GPU computation in Sec.4. In Sec.5 we evaluate the running speed of the proposed method, we finally conclude the work in Sec.6.

2 PRELIMINARIES

2.1 Problem Formulation

A Bayesian Network model is a graphical model that represents the causal relationship among events. A Bayesian Network model has a structure represented by a directed graph where events are denoted by nodes while causal relationships are denoted by directed edges. In many cases (including this work), each node takes multinomial discrete values, and conditional probabilities among them are expressed by a model. See Fig. 1 for a concise example. Nodes n_1, n_2, and n_3 represent distinct events, where they take 1 if the corresponding events occur, and take 0 if the events do not occur (in this case we show a binomial case for conciseness). Edges $n_1 \rightarrow n_3$ and $n_2 \rightarrow n_3$ represent causal relationships, which mean that the probability of occurrence for each n_3 value depends on the values of n_1 and n_2. If edge $n_1 \rightarrow n_3$ exists, we call that n_1 is a parent of n_3 and n_3 is a child of n_1. Because nodes n_1 and n_2 do not have their parents, they have own prior probabilities $P(n_1)$ and $P(n_2)$. On the other hand, because node n_3 has two parents n_1 and n_2, it has a conditional probability $P(n_3|n_1, n_2)$. In this example, the probability that n_3 occurs is 0.890 under the assumption that both n_1 and n_2 occur. Note that, from this model, Bayesian inference is possible: if n_3 is known, then the posterior probability of n_1 and n_2 can be determined, which enables us to infer more accurately the occurrence of events.

The Bayesian Networks model can be learned from the data obtained through the observation of events. Let $N = \{n_i\}$, $(1 \leq i \leq |N|)$ be a set of events, and $O = \{o_j\}$, $(1 \leq j \leq |O|)$ be a set of observations, where $|N|$ is the number of events and $|O|$ that of observations. Let $o_j = (x_{j1}, x_{j2}, \ldots, x_{j|N|})$ be j-th observation, which is a set of observed values x_{ji} on event n_i for all $i (1 \leq i \leq |N|)$. We try to learn a good Bayesian Network model m from the given set of observations. Note that, good Bayesian Network model m is the one that creates data sets similar to the original observation O. As an model score (i.e., evaluation criterion) to measure the level of fitting between m and O, several information criteria such as AIC (Akaike's Information Criterion) [9] are used. Formally, the problem of learning Bayesian Networks that we consider in this paper is defined as follows:

Problem 1: From the given set of observations O, find a Bayesian Network model m that has the lowest model score.

2.2 PBIL

In PBIL, an individual creature m is defined as a vector $m = \{e_1, e_2, \ldots, e_L\}$, where $e_i (1 \leq i \leq L)$ is the i-th element that takes a value 0 or 1, and L is the number of elements that consist of an individual. Let $P = \{p_1, p_2, \ldots, p_L\}$ be a probability vector where $p_i (1 \leq i \leq L)$ represents the probability to be $e_i = 1$. The algorithm of PBIL is described as follows:

(1) As initialization, we let $p_i = 0.5$ for all $i = 1, 2, \ldots, L$.

(2) Generate a set M that consists of $|M|$ individuals according to probability vector P, i.e., element e_i of each individual is determined by the corresponding probability p_i.

(3) Compute the score for each individual $m \in M$.

(4) Select a set of individuals M^{Topk} whose members have evaluation scores within top k in M, and update the probability vector according to M^{Topk}. Specifically, the formula applied to every p_i to update the probability vector is shown as follows.

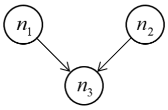
$$p_i^{\mathrm{new}} = ratio(i) \times \alpha + p_i \times (1 - \alpha), \qquad (1)$$

where p_i^{new} is the updated value of the new probability vector (p_i is replaced with p_i^{new} in the next generation), $ratio(i)$ is the function that represents the ratio of individuals in M^{Topk} that include edge i (i.e., $e_i = 1$), and α is the parameter called learning ratio.

(5) Repeat steps (2)-(4) until P converges.

By merging top-k individuals, PBIL evolves the probability vector such that the good individuals are more likely to be generated. Different from other genetic algorithms, PBIL does not include "crossover" between individuals. Instead, it evolves the probability vector as a "parent" of the generated individuals.

n_1	$P(n_1)$
0	0.79
1	0.21



n_2	$P(n_2)$
0	0.88
1	0.12

n_1	n_2	$P(n_3 \mid n_1, n_2)$	
		0	1
0	0	0.999	0.001
0	1	0.710	0.290
1	0	0.360	0.640
1	1	0.110	0.890

Figure 1: An Example of Bayesian Network Models

P		Parent Node					
		n_1	n_2	…	n_i	…	$n_{\lvert N\rvert}$
Child Node	n_1	0.0	0.5	…	p_{i1}	…	0.5
	n_2	0.5	0.0	…	p_{i2}	…	0.5
	⋮	⋮	⋮	⋱	⋮	…	⋮
	n_j	p_{1j}	p_{2j}	…	p_{ij}	…	p_{Nj}
	⋮	⋮	⋮	⋮	⋮	⋱	⋮
	$n_{\lvert N\rvert}$	0.5	0.5	…	p_{iN}	…	0.0

Figure 2: A Probability Vector.

2.3 PBIL-based Bayesian Networks Learning

In this section, we describe a PBIL-based algorithm that learns BN models. Because our problem (i.e. Problem 1) to learn BN models is a little different from the general description of PBIL shown in the previous section, a little adjustment is required. In our problem, individual creatures correspond to each BN model. Namely, with the set of events N, an individual model is represented as $m = \{e_{11}, e_{12}, \ldots, e_{1|N|}, e_{21}, e_{22}, \ldots, e_{|N|1}, e_{|N|2}, \ldots, e_{|N||N|}\}$ where e_{ij} corresponds to the edge from an event n_i to n_j, i.e., if $e_{ij} = 1$, the edge from n_i to n_j exists in m, and if $e_{ij} = 0$ it does not exist. Similarly, we have the probability vector P to generate individual models as $P = \{p_{11}, p_{12}, \ldots, p_{1|N|}, p_{21}, p_{22}, \ldots, p_{|N|1}, p_{|N|2}, \ldots, p_{|N||N|}\}$ where p_{ij} is the probability that the edge from n_i to n_j exists. A probability vector can be regarded as a table as illustrated in Fig. 2. Note that, because BNs do not allow self-edges, p_{ij} is always 0 if $i = j$. The process of the BN-learning algorithm is basically obtained from the steps of the general PBIL, as described in the following (See also Fig. 3 that illustrate these steps).

(1) Initialize the probability vector P as $p_{ij} = 0$ if $i = j$, and $p_{ij} = 0.5$ otherwise, for each $i, j (1 \leq i, j \leq |N|)$.

(2) Generate M as a set of $|M|$ individual models according to P.

(3) Compute the evaluation scores for all individual models $m \in M$.

(4) Select a set of individuals M^{Topk} whose members have

top-k evaluation values in M, and update the probability vector according to the formula (1).

(5) Repeat steps (2)-(4) until P converges.

Same as the genearal PBIL, the BN-learning algorithm evolves the probability vector so that we can generate better individual models. However, there is a constraint specific to BNs, that is, a BN model is not allowed to have cycles in it. To consider this constraint in the algorithm, step 2 is detailed as follows:

(2a) Consider every pair of events (i, j) where $1 \leq i, j \leq |N|$ and $i \neq j$, create a random order of them.

(2b) For each pair (i, j) in the order created in step (2a), determine the value e_{ij} according to P; every time e_{ij} is determined, if e_{ij} is determined as 1, we check whether this edge from n_i to n_j creates a cycle with all the edges determined to exist so far. If it creates a cycle, let e_{ij} be 0.

(2c) Repeat steps (2a) and (2b) until all the pairs in the order are processed.

These steps enable us to learn good BN models within the framework of PBIL.

2.4 PBIL-RS

Note that PBIL introduced above does not include mutation operators. Therefore, naturally, it easily converges to a local minimum solution. To avoid converging to the local minimum solution and to continuously improve the solution after that, several mutation operators have been proposed such as Bitwise Mutation (BM) [6], Transpose Mutation (TM) [5], and Probability Mutation (PM) [7]. PBIL-RS (PBIL-Repeated Search) [8] is also a method to avoid converging to local minimum solution, which, when it detects convergence, spreads the search area again. PBIL-RS is shown to find better solutions compared to the mutation-based methods such as BM, TM, and PM by repeating spreading and converging [8].

3 GPU ARCHITECTURE

3.1 GPU Structure

GPU (Graphics Processing Unit) is an arrayed processor that is originally developed to accelerate graphical computing, which currently is used to accelerate general scientific computation. A GPU structure is illustrated in Fig. 4. A GPU has a hierarchical structure where it consists of multiple (tens to hundreds of) Streaming Multi-processors (SMs) and a SM further includes multiple (tens to hundreds of) Streaming Processors (SPs). Since each SP in a SM concurrently executes a fragment of program code, a GPU executes a number of fragmented codes in parallel, which potentially results in significant performance.

To exchange data between a CPU and a GPU, a memory called *global memory* is prepared in the GPU that can be accessed by both the CPU and the GPU. Also, to accelerate parallel computation, each SM has a small high-speed memory called *shared memory* that can be accessed by all SPs in the same SM. A thread runs in each SP, and the threads in the same SM runs the same bytecode in parallel with different values of variables. Thus, memory accesses of threads are expected to occur simultaneously. To optimize the efficiency of the parallel access is the key issue to design algorithms for GPU. Note that, if the number of threads to execute exceeds the number of SPs in a SM, a single SP executes multiple threads in turn until all of them are executed.

3.2 Coalesce Access to Global Memory

To have as much performance gain as possible from GPU, one of the most basic techniques is to consider the efficient access to the global memory called *coalesce access*, which is illustrated in Fig. 5. Coalesce access is a synchronized parallel access technique in which threads in a SM simultaneously access the successive addresses in the global memory to achieve high-throughput memory access. When the access is scheduled completely to the successive addresses, the SM read/write the memory block in a single action that completes the access of all SPs. To utilize coalesce accesses is an important technique in designing GPU algorithms.

4 ACCELERATING PBIL WITH GPU

4.1 Overview

The method we propose in this paper extends PBIL and its family algorithms (such as PBIL-RS and the mutation extensions) to run in significantly shorter time by means of parallel computation of GPU. We re-designed Step (3) of PBIL described in Sec.3 for GPU execution to compute the model scores for a collection of models M. Because, in PBIL, hundreds of models are to be computed in a single generation to estimate a distribution of model scores, introducing parallel computation in Step (3) is significantly effective.

Specifically, we detailed the step (3) in the following.

(3a) Transporting data from CPU to the global memory.

(3b) For each model m, we compute the evaluation score by executing the following substeps (3b-1) and (3b-2).

(3b-1) Counting the occurrences in the observation set that match each value pattern.

(3b-2) Computing evaluation scores from the counts.

(3c) Transporting the computed scores back to CPU.

As written in Step (3b), we first count the number of occurrences in the observation data O that match each value pattern of events. Here, value patterns are defined on each event as a set of values taken by the event and its parent events. For definition, see Fig. 1 again. The value patterns on event n_3 is the combination of values of n_3 and its parents n_1 and n_2. Since those three variables take binomial values (i.e., 0 or 1), we have 8 value patterns such as $(n_1, n_2, n_3) = (0, 0, 0), (0, 0, 1), \ldots, (1, 1, 1)$. For conciseness, we denote it by $n_1n_2n_3 = \{000, 001, \ldots, 111\}$. In general, the set of value patterns for event n_i is denoted by

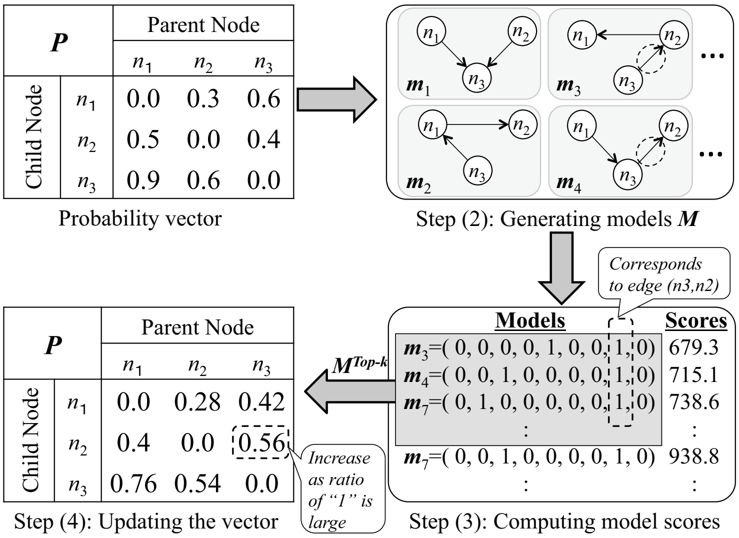

P
Parent Node
n_1 n_2 n_3
Child Node
n_1 0.0 0.3 0.6
n_2 0.5 0.0 0.4
n_3 0.9 0.6 0.0
Probability vector
m_1 m_2 m_3 m_4
...
Step (2): Generating models M
Corresponds to edge (n3,n2)
Models
Scores
m_3=(0, 0, 0, 0, 1, 0, 0, 1, 0) 679.3
m_4=(0, 0, 1, 0, 0, 0, 0, 1, 0) 715.1
m_7=(0, 1, 0, 0, 0, 0, 0, 1, 0) 738.6
m_7=(0, 0, 1, 0, 0, 0, 0, 1, 0) 938.8
Step (3): Computing model scores
$M^{Top\text{-}k}$
P
Parent Node
n_1 n_2 n_3
Child Node
n_1 0.0 0.28 0.42
n_2 0.4 0.0 0.56
n_3 0.76 0.54 0.0
Increase as ratio of "1" is large
Step (4): Updating the vector

Figure 3: Overview of PBIL

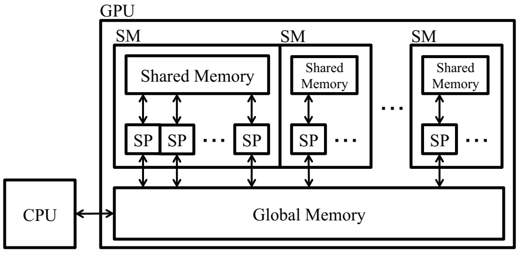


Figure 4: GPU Architecture

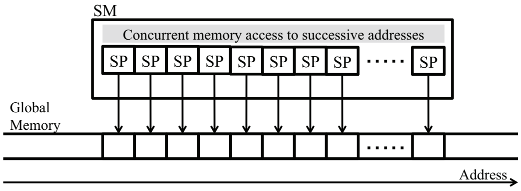

SM
Concurrent memory access to successive addresses
SP SP SP SP SP SP SP SP ····· SP
Global Memory
·····
Address

Figure 5: Coalesce Access

$V_i = p_1p_2\dots p_rn_i = \{0..00, 0..01, \cdots, 1..11\}$ where the parents of n_i in model m is $\mathrm{pa}(n_i) = \{p_1, p_2, \dots, p_r\}$ and r is the number of parents of n_i.

Then, our task in step (3b) is to compute the number of occurrences $\mathcal{N}_{iv}$ in O that takes value pattern v, for every event $i \in N$ and value pattern $v \in V_i$. Although there are several information criterion such as AIC, BIC and MDL that are used as model scores in learning Bayesian Network structure, they are all computed from the counts of value patterns, as shown in Sec. 4.3.

The basic strategy for counting value patterns is to assign a SM to a single model m, and to use all SPs in the SM in parallel to count all value patterns for all events in m. By assigning a model m to a single SM, we compute the model score of m in the SM. Each SM in a GPU processes models one by one in parallel to compute model scores for all models in M. In the following subsections, we describe the algorithm to process m within a SM to show how to make efficient manipulation of data, especially to gain from the coalesce access of global memory.

4.2 Data Structure

In Step (3a), we transport the data required to compute model scores to the global memory. We declare three arrays that represent the following sets, respectively, in the global memory.

(i) The observation set O.

(ii) The model set M.

(iii) An array to record the computed model scores.

The pseudo code to define these data items is shown in the following.

```
u_int8_t observation[N][O];
boolean model[M][N][N];
float modelScore[M];
```

Here, variables `M,N,O` represents $|M|, |N|$, and $|O|$, respectively. We represent the observation set O as a two-dimensional array `observation` where the 1st dimension is events and the 2nd observations. We represent The model set M as a three-dimensional array `model` where 1st dimension is used for the index of models and 2nd and 3rd dimensions are used to describe each model. Each model is expressed as an adjacency matrix such that `model[m][i][j]` is true if there is a directed link from event i to j in m. The array `modelScore` is used to retain the value computed by the algorithm.

We use the Shared Memory to place the counters that retain the counts of every value patterns for all events in a model, as follows.

```
u_int16_t counter[Vi];
```

Here, `Vi` denotes the number of value patterns on event $i \in N$. Note that `Vi` is determined depending on the number of parents of i in the model m, and the number of their multinomial values. Thus, this array may exceed the capacity of the shared memory. (Consider that, if p_j may take a value from $w(p_j)$ distinct values, `Vi` $= w(i)w(p_1)w(p_2)\dots w(p_r)$.) If the shared memory can afford to store this array in size, we execute fast counting algorithm that we call *case-1* shown in Sec. 4.4, and otherwise, we use alternative algorithm that we call *case-2* shown in Sec. 4.5.

4.3 Model Scores

Note that we count the number of observations in each case to compute evaluation scores. Although there are several information criterion such as AIC, BIC and MDL, used as model scores in learning Bayesian Network structure, they all are computed from the number of observations in each case. For instance, AIC is computed with the following formula:

$$AIC = -2l(\theta|O) + 2k, \tag{2}$$

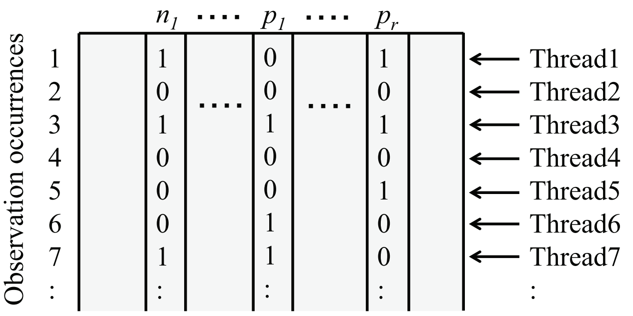


Figure 6: Counting Value Patterns (Case 1)

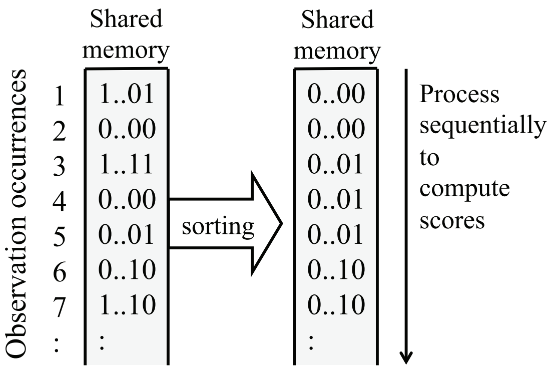


Figure 7: Counting Value Patterns (Case 2)

where θ denotes the parameter set, $l(\theta|O)$ denotes the likelihood of θ under observation O, and the term k represents the number of parameters. Here, we further show that the function $l(\cdot)$ is represented by the following formula

$$l(\theta|O) \propto \sum_{i \in N} \sum_{v \in V_i} (\mathcal{N}_{iv}) \log \theta_{iv}, \tag{3}$$

where i denotes an event, v denotes a value pattern, $\mathcal{N}_{iv}$ denotes the number of occurrences in observations O that match i and v, and θ_{iv} is a parameter computable from $\mathcal{N}_{iv}$. This means that, by counting the number of observations for each value pattern, we can compute the evaluation score of the model m such as AIC. Note that other information criteria used in Bayesian Networks such as BIC, MDL, etc., also can be computed through counting the matching value patterns in the observation set O.

4.4 Computing Model Scores (Case 1)

If the size of the array `counter[Vi]` is within the capacity of shared memory, the procedure described here (i.e., *case-1*) is applied. Before computing the model score of the given model $m \in M$, the SM responsible to this task computes $\mathcal{N}_{iv}$ for all $i \in N$ and $v \in V_i$. In the procedure, the SM proceeds each node i sequentially. Thus, we now fix $i \in N$ and focus on computing $\mathcal{N}_{iv}$ for all $v \in V_i$.

Our strategy to do this is to process occurrences in O in parallel using SPs. Thus, we designed the procedure such that each thread reads a single occurrence of O and increments the corresponding value in `counter`. Namely, we have as large number of threads as the occurrences of O. By the scheduler of GPU that assigns threads to SPs, we can read occurrences from successive addresses of the global memory as shown in Fig. 6 (remember that the addresses of occurrences for $n_i, p_1, p_2, \ldots, p_r$ are successive in array `observation`, respectively), gaining from coalesce access.

As the result of the above procedure executed for all events $i \in N$, we can obtain $\mathcal{N}_{iv}$ for all $i \in N$ and $v \in V_i$.

After computing $\mathcal{N}_{iv}$ for all $i \in N$ and $v \in V_i$, we compute the model score of m from them. This is simply done by computing the value according to formula (2). To compute it in parallel, we assign a SP for each $i \in N$ and sum up the model scores on the shared memory using a technique called *parallel reduction*. We finally store the computed score into the array `modelScore` in the global memory.

4.5 Computing Model Scores (Case 2)

If the size of the array `counter[Vi]` exceeds the capacity of the shared memory, we have to choose a less efficient algorithm. In the Alarm Network used in the evaluation of this paper, we can use the *case-1* algorithm only when the number of parents r is less than 7, where each event takes 4 distinct values and we have about 48KBytes shared memory.

In the *case-2* algorithm, instead of the array `counter[Vi]`, we define the array `patternValue[O]` as follows.

```
u_int16_t patternValue[O];
```

We simply use this array by storing pattern values of each occurrences of O. Retrieval of the pattern values from global memory can be done using coalesce access in the similar way to *case-1*. After retrieving the pattern values, we sort the values as shown in Fig. 7. Note that a GPU-specific sorting algorithm called *bitonic sort* can be used for high-throughput sorting. Then, we trace through the array sequentially in order to count each value pattern and sum up the model score. Although it takes a little longer than *case-1*, we can compute the model score in relatively short time.

5 EVALUATION

We evaluate the proposed method in terms of both running time and quality of the output model.

To clarify the performance of the proposed method to accelerate computational speed, we compare the running time of the proposed algorithm executed on GPU with its base PBIL-based algorithm executed on CPU. Note that their output is the same, only running time is different.

5.1 Computational Time

We compare the running time of the proposed algorithm executed on GPU with its base algorithm PBIL-RS executed on CPU. We implemented both algorithms in C++ language with CUDA library for GPU processing. The execution environment is shown in Table. 1. We used Alarm Network [10] including 37 nodes as the base BN model; we generate an observation set including 1024 occurrences based on Alarm Network and learn BN models using each algorithm.

Table 1: Evaluation Environment

OS		CentOS 5.0
CPU		Intel Core i7 4770k (3.50GHz)
Memory		32GBytes
GPU	Model	nVidia GeForce GTX TITAN Black (0.98GHz)
	# of SM	15
	# of SP per SM	192 (2880 SPs in total)
	Global Memory	6143 MBytes
	Shared Memory	49152 Bytes per SM
	GPU Library	CUDA 6.0
Compiler		g++ 4.1.2

Fig. 8 shows the computational time as generation proceeds. We see that the proposed algorithm that runs on GPU is about 14-times faster than PBIL-RS that runs on CPU only. In this figure, we also show a variant of the proposed algorithm seen as "case-2 only" that always runs *case-2* algorithm instead of *case-1* even if the number of parents is small. This variant takes about 1.6-times longer than the both-algorithm case, which indicates that *case-1* algorithm is considerably faster than *case-2*. To see the difference more precisely, we show the execution ratio of *case-1* and *case-2* in each generation in Fig. 9. Because mostly *case-1* algorithm is executed with 100% ratio, we can estimate that the *case-1* algorithm is about 1.6-times faster than *case-2*.

6 CONCLUSION

We proposed a method to accelerate PBIL-based BN learning algorithms using GPU computation. We make the most of the coalesce access technique of GPU computation to reduce computation time using consumer level hardware. Through evaluation, we confirmed that the proposed method achieves about 14-times faster in running speed than the original PBIL-RS executed on only CPU.

ACKNOWLEDGMENT

This work was partly supported by "the Program for Promo- tion of Stockbreeding" of JRA (Japan Racing Association).

REFERENCES

[1] D.M. Chickering, D. Heckerman, C. Meek, "Large-Sample Learning of Bayesian Networks is NP-Hard," Journal of Machine Learning Research, Vol.5, pp.1287–1330 (2004).

[2] G.F. Cooper, and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data," Machine Learning, Vol.9, pp.309–347 (1992).

[3] P. Larrañaga, H. Karshenas, C. Bielza, R. Santana, "A review on evolutionary algorithms in Bayesian Network learning and inference tasks," Information Sciences, Vol.233, No.1, pp.109-125 (2013).

[4] S. Baluja, "Population-Based Incremental Learning: A method for Integrating Genetic Search Based Function Optimization and Competitive Learning," Technical Report CMU-CS-94-163 (1994).

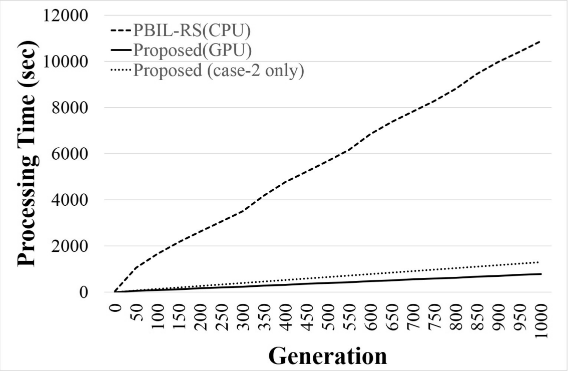


Figure 8: Execution Time (CPU vs. GPU)

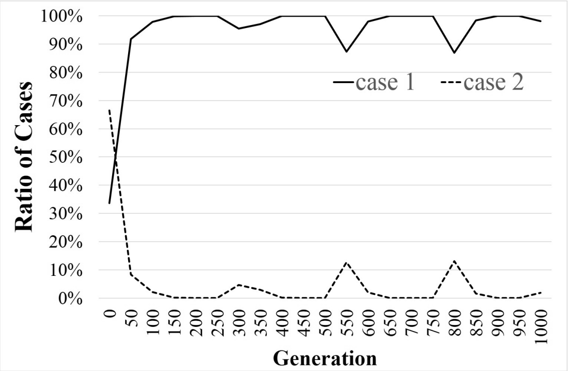


Figure 9: Ratio of Case 1 and 2

[5] D.W. Kim, S. Ko, and B.Y. Kang, "Structure Learning of Bayesian Networks by Estimation of Distribution Algorithms with Transpose Mutation," Journal of Applied Research and Technology, Vol.11, pp.586–596 (2013).

[6] H. Handa, "Estimation of Distribution Algorithms with Mutation," Lecture Notes in Computer Science, Vol.3448, pp.112–121 (2005).

[7] S. Fukuda, Y. Yamanaka, and T. Yoshihiro, "A Probability-based Evolutionary Algorithm with Mutations to Learn Bayesian Networks," International Journal of Artificial Intelligence and Interactive Multimedia, Vol.3, No.1, pp.7–13 (2014).

[8] Yuma Yamanaka, Takatoshi Fujiki, Sho Fukuda, and Takuya Yoshihiro, "PBIL-RS: An Algorithm to Learn Bayesian Networks Based on Probability Vectors," In Proc. IWIN'2015 (2015).

[9] H. Akaike, "Information theory and an extension of the maximum likelihood principle," In Proc. ISIT'73, pp.267-281 (1973).

[10] I.A. Beinlich, H.J. Suermondt, R.M. Chavez, G.F. Cooper, "The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks," In Proc. AIME'89, Vol. 38, pp.247-256 (1989).

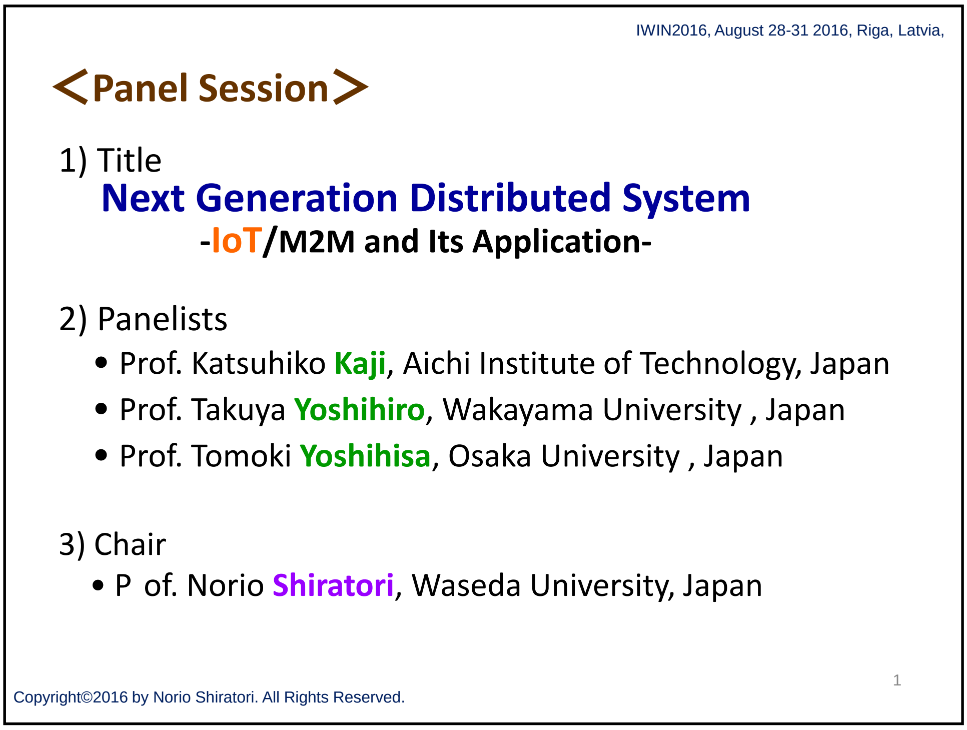
Panel Discussion

(Chair: Norio Shiratori)

(Panelist: Katsuhiko Kaji)

(Panelist: Takuya Yoshihiro)

(Panelist: Tomoki Yoshihisa)



IWIN2016, August 28-31 2016, Riga, Latvia,

＜Panel Session＞

1) Title

Next Generation Distributed System

-IoT/M2M and Its Application-

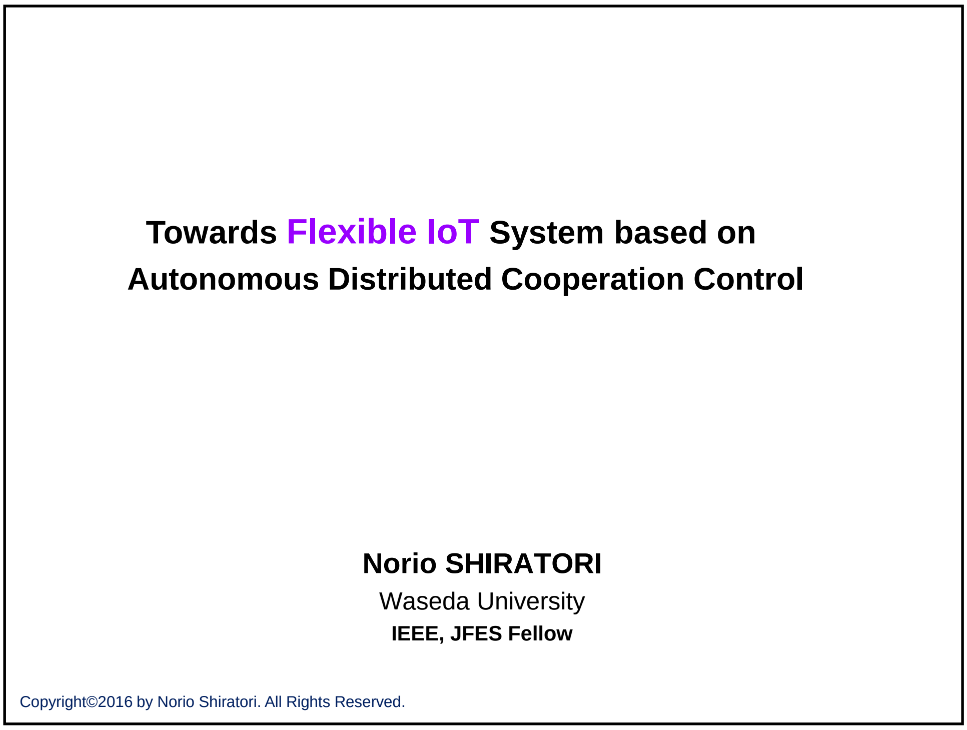
2) Panelists

- Prof. Katsuhiko **Kaji**, Aichi Institute of Technology, Japan
- Prof. Takuya **Yoshihiro**, Wakayama University , Japan
- Prof. Tomoki **Yoshihisa**, Osaka University , Japan

3) Chair

- P of. Norio **Shiratori**, Waseda University, Japan

1



Towards Flexible IoT System based on Autonomous Distributed Cooperation Control

Norio SHIRATORI

Waseda University

IEEE, JFES Fellow

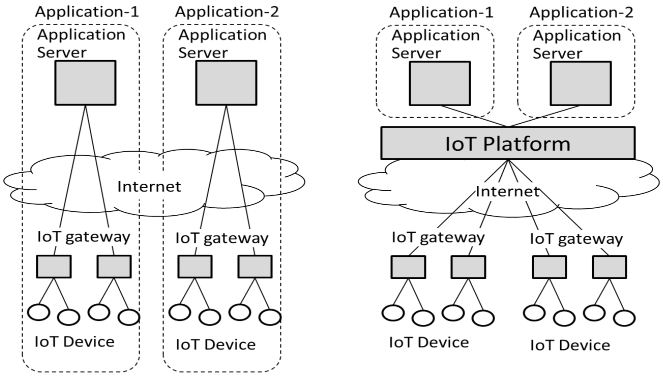
Table of Contents

1. Current Status of **IoT Architecture** and Its Problems
2. Research **Trend** of IoT Architecture
3. **Edge** Computing
4. Towards **Flexible IoT** System based on Autonomous Distributed Cooperation Control

3

1. Current Status of IoT Architecture and Its Problems

1.1 Cloud Centric- Type Systems



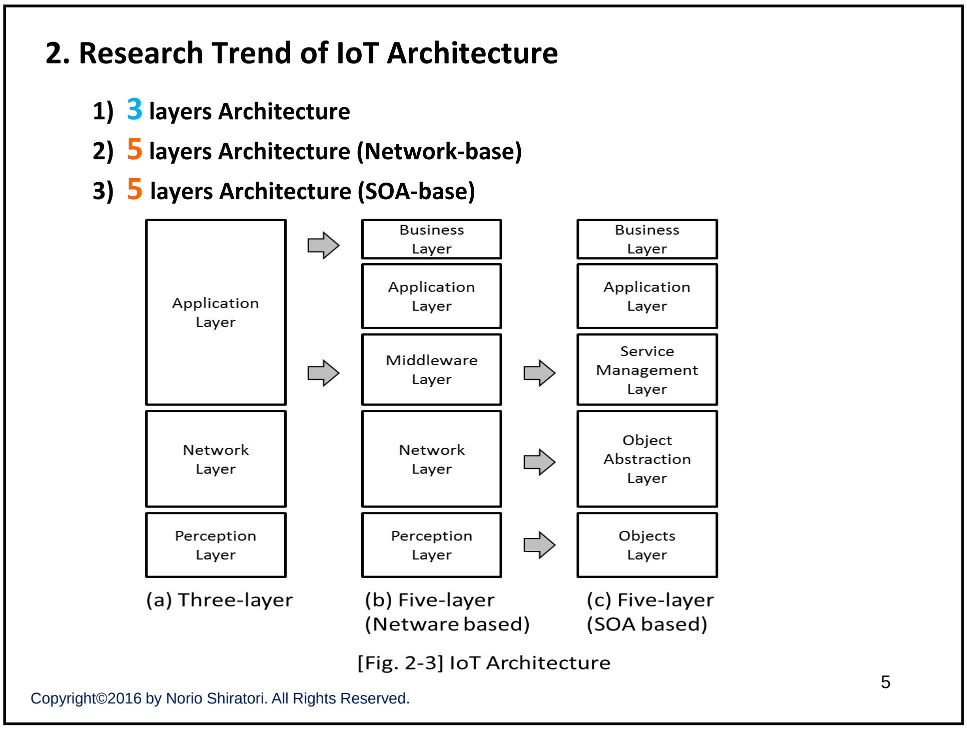
[Fig.2-1] Vertical IoT System

[Fig.2-2] Cloud-centric IoT System with IoT Platform

▪Problems

1) Lack of cloud-side network capacity for large scale IoT systems
2) Decrease of cloud service level by the disconnection btw the platform and gateways
3) Privacy and information leakage problems because all of collected data are stored in the IoT Platform

4

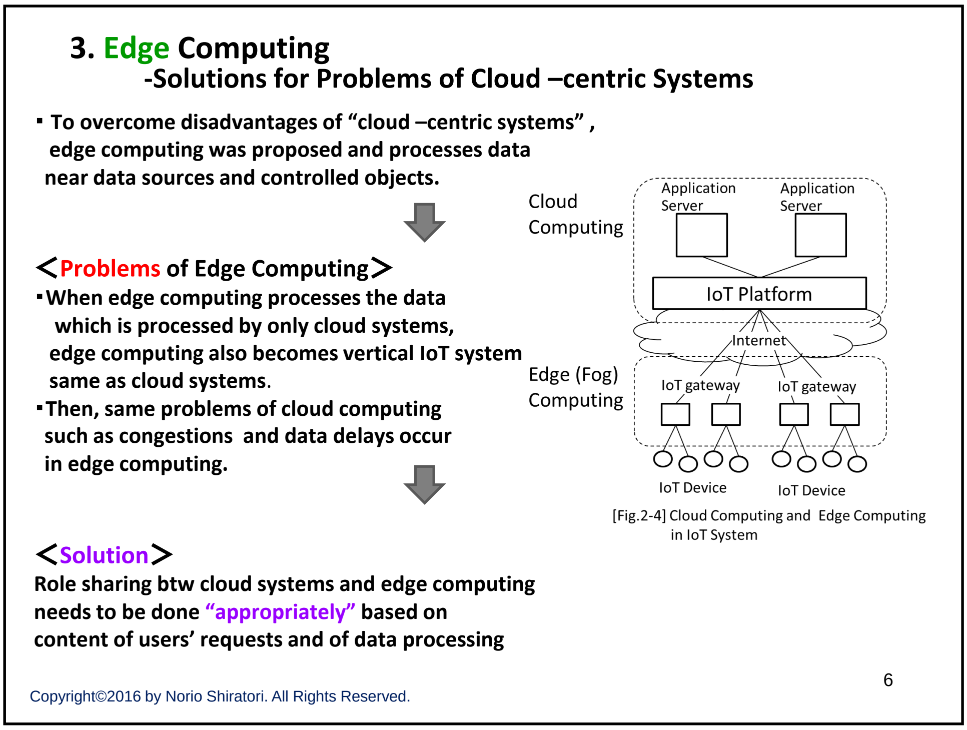


2. Research Trend of IoT Architecture

1) 3 layers Architecture
2) 5 layers Architecture (Network-base)
3) 5 layers Architecture (SOA-base)

[Fig. 2-3] IoT Architecture

5



3. Edge Computing -Solutions for Problems of Cloud –centric Systems

- To overcome disadvantages of "cloud –centric systems" , edge computing was proposed and processes data near data sources and controlled objects.

＜Problems of Edge Computing＞

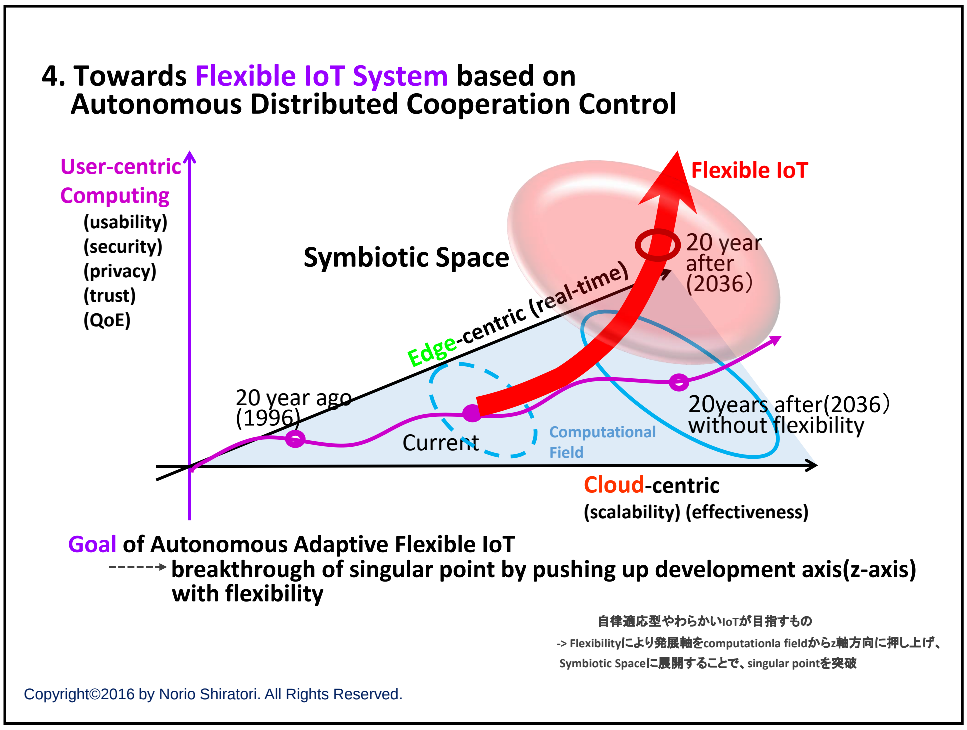
- When edge computing processes the data which is processed by only cloud systems, edge computing also becomes vertical IoT system same as cloud systems.
- Then, same problems of cloud computing such as congestions and data delays occur in edge computing.

＜Solution＞

Role sharing btw cloud systems and edge computing needs to be done "appropriately" based on content of users' requests and of data processing

[Fig.2-4] Cloud Computing and Edge Computing in IoT System

6



4. Towards Flexible IoT System based on Autonomous Distributed Cooperation Control

User-centric Computing
(usability)
(security)
(privacy)
(trust)
(QoE)

Symbiotic Space

Flexible IoT

20 year after (2036）

Edge-centric (real-time)

20 year ago (1996)

Current

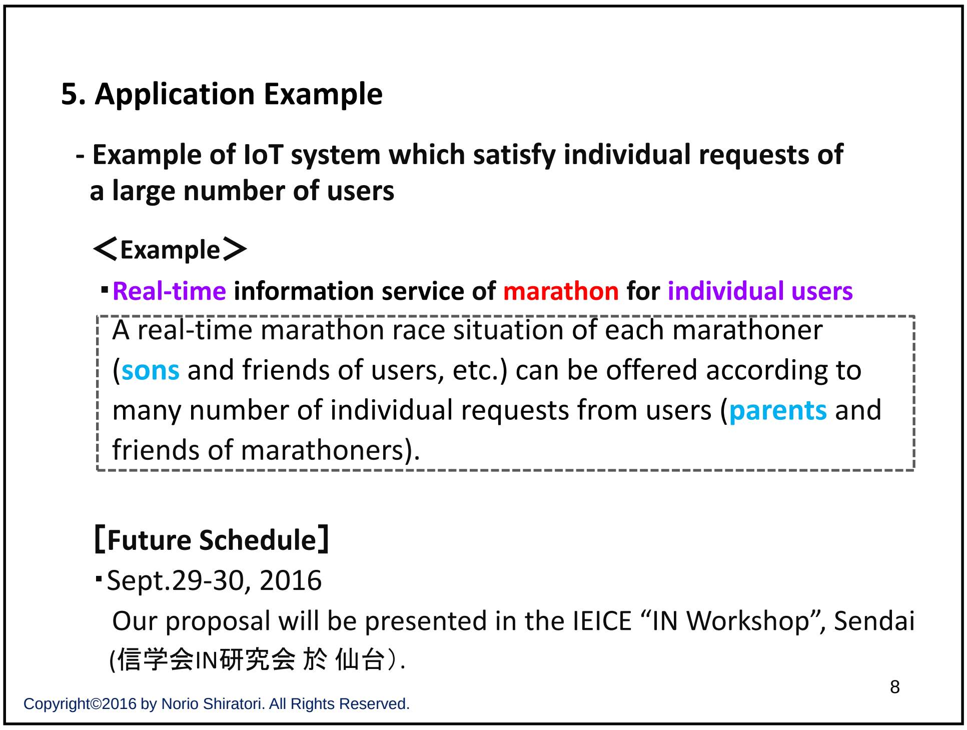
Computational Field

20years after(2036）
without flexibility

Cloud-centric
(scalability) (effectiveness)

Goal of Autonomous Adaptive Flexible IoT
------→ breakthrough of singular point by pushing up development axis(z-axis) with flexibility

自律適応型やわらかいIoTが目指すもの
-> Flexibilityにより発展軸をcomputationla fieldからz軸方向に押し上げ、
Symbiotic Spaceに展開することで、singular pointを突破



5. Application Example

- Example of IoT system which satisfy individual requests of a large number of users

＜Example＞

･Real-time information service of marathon for individual users

A real-time marathon race situation of each marathoner (sons and friends of users, etc.) can be offered according to many number of individual requests from users (parents and friends of marathoners).

［Future Schedule］

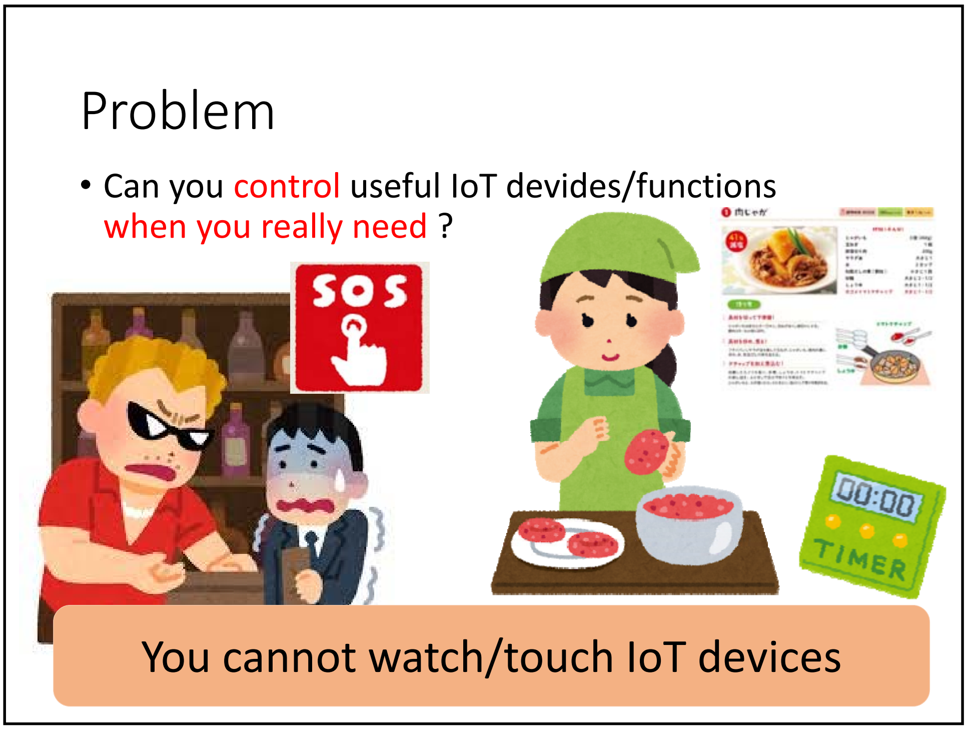
･Sept.29-30, 2016

Our proposal will be presented in the IEICE “IN Workshop”, Sendai (信学会IN研究会 於 仙台）.

8

Let's think about user

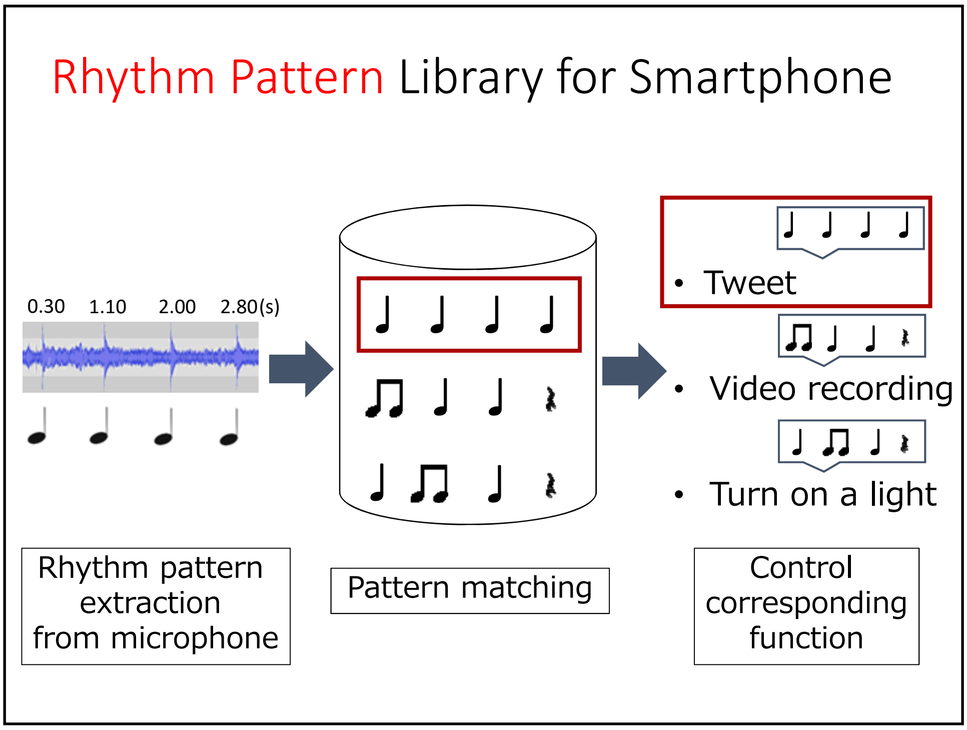
Katsuhiko Kaji
Aichi Institute of Technology



Problem

- Can you control useful IoT devides/functions when you really need ?

You cannot watch/touch IoT devices



Rhythm Pattern Library for Smartphone

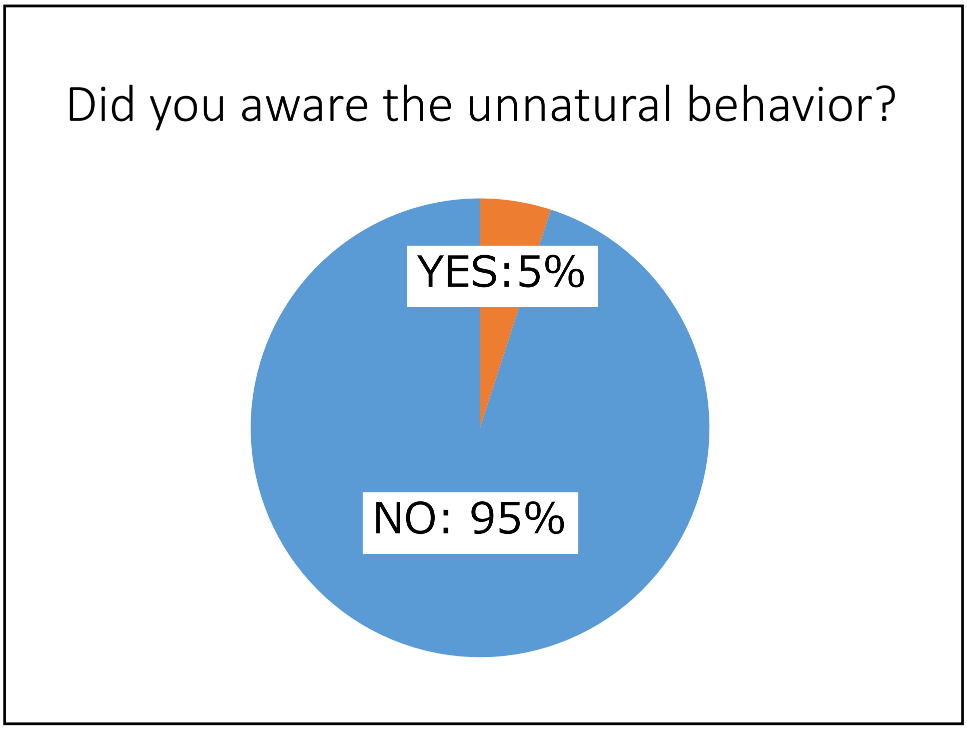
0.30 1.10 2.00 2.80(s)

- Tweet
- Video recording
- Turn on a light

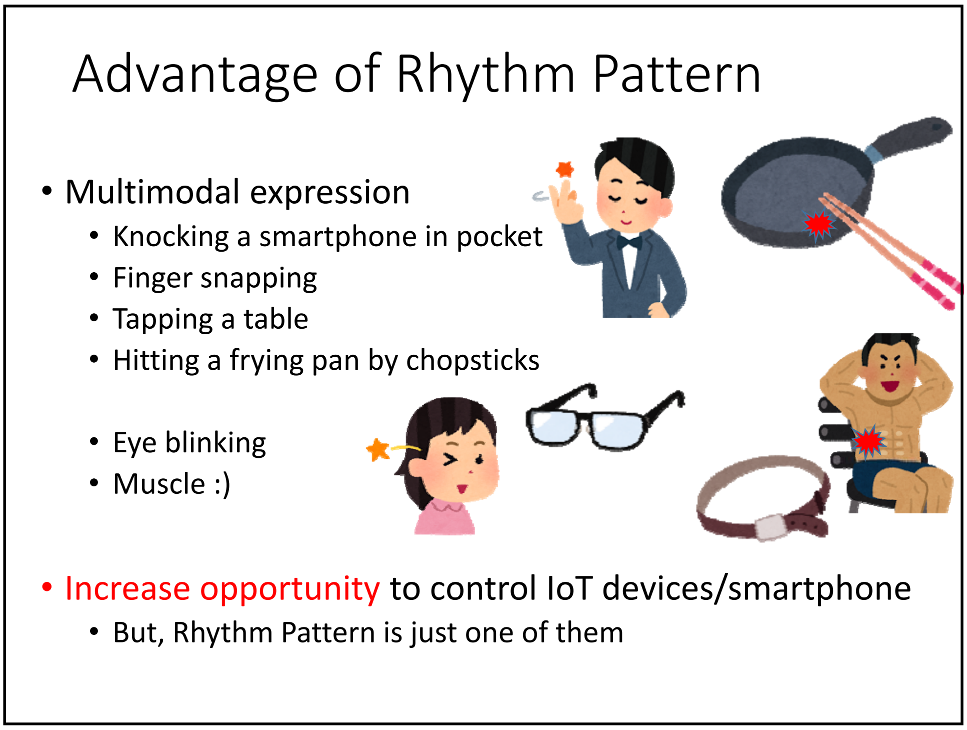
Rhythm pattern extraction from microphone

Pattern matching

Control corresponding function



Did you aware the unnatural behavior?



Advantage of Rhythm Pattern

- Multimodal expression
  - Knocking a smartphone in pocket
  - Finger snapping
  - Tapping a table
  - Hitting a frying pan by chopsticks
  - Eye blinking
  - Muscle :)
- Increase opportunity to control IoT devices/smartphone
  - But, Rhythm Pattern is just one of them

BLE Beacon

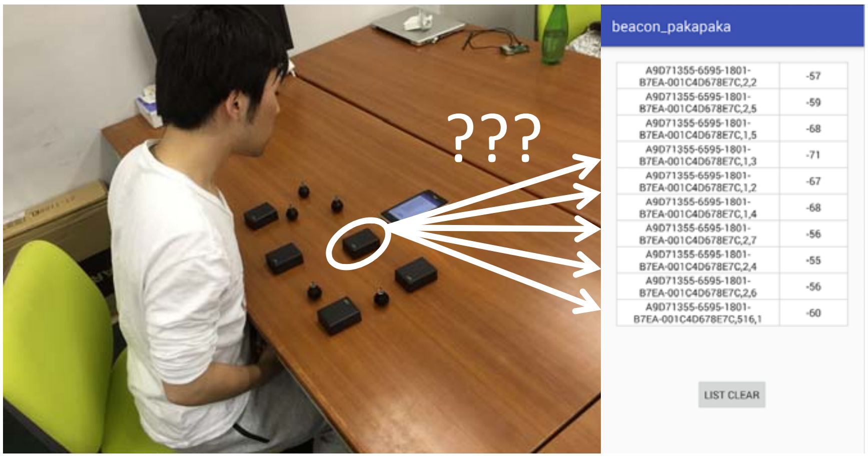
- Easy to detect proximity
- Small, cheap, low consumption, longer operating life
- Coupon service, stamp rally, indoor positioning

Let's Imagine …

- You are an event manager of stamp rally.
- Today, the event finished.
- You should gather the BLE beacons and configure/remove/add
  some of them for next stamp rally.
  (e.g.: ID, signal strength, beacon interval).

Problem

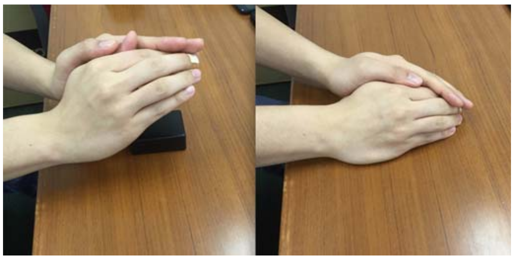
Management terminal



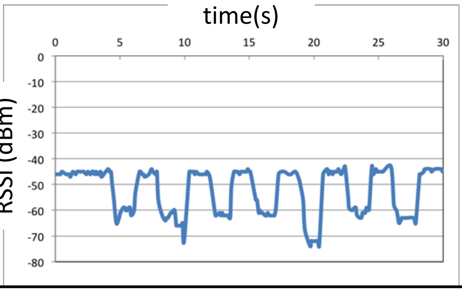
A9D71355-6595-1801-B7EA-001C4D678E7C,2,2	-57
A9D71355-6595-1801-B7EA-001C4D678E7C,2,5	-59
A9D71355-6595-1801-B7EA-001C4D678E7C,1,5	-68
A9D71355-6595-1801-B7EA-001C4D678E7C,1,3	-71
A9D71355-6595-1801-B7EA-001C4D678E7C,1,2	-67
A9D71355-6595-1801-B7EA-001C4D678E7C,1,4	-68
A9D71355-6595-1801-B7EA-001C4D678E7C,2,7	-56
A9D71355-6595-1801-B7EA-001C4D678E7C,2,4	-55
A9D71355-6595-1801-B7EA-001C4D678E7C,2,6	-56
A9D71355-6595-1801-B7EA-001C4D678E7C,516,1	-60

PakaPaka Method for BLE Beacon Identification

PakaPaka

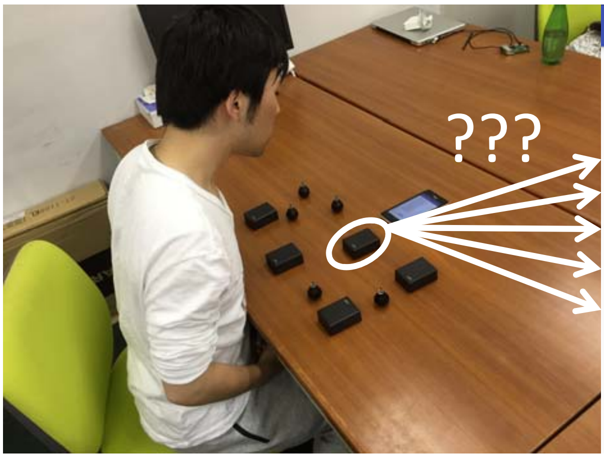


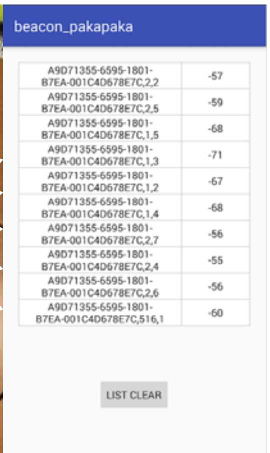
Management terminal



Solved

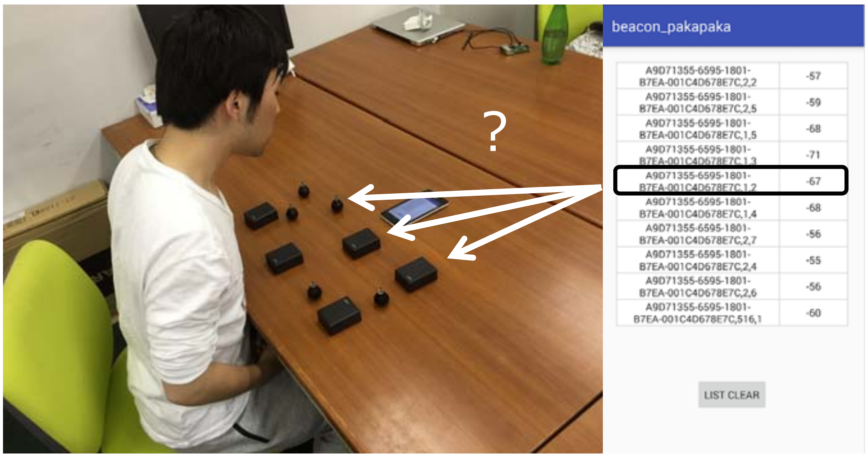
Management terminal



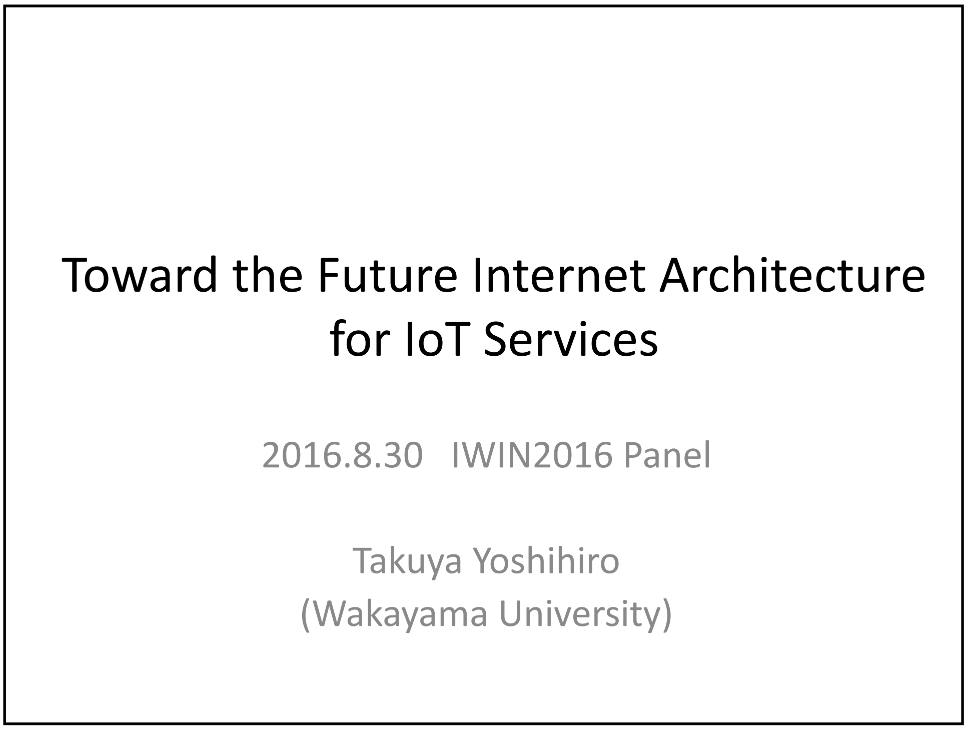


A9D71355-6595-1801-B7EA-001C4D678E7C,2,2	-57
A9D71355-6595-1801-B7EA-001C4D678E7C,2,5	-59
A9D71355-6595-1801-B7EA-001C4D678E7C,1,5	-68
A9D71355-6595-1801-B7EA-001C4D678E7C,1,3	-71
A9D71355-6595-1801-B7EA-001C4D678E7C,1,2	-67
A9D71355-6595-1801-B7EA-001C4D678E7C,1,4	-68
A9D71355-6595-1801-B7EA-001C4D678E7C,2,7	-56
A9D71355-6595-1801-B7EA-001C4D678E7C,2,4	-55
A9D71355-6595-1801-B7EA-001C4D678E7C,2,6	-56
A9D71355-6595-1801-B7EA-001C4D678E7C,516,1	-60

Future work



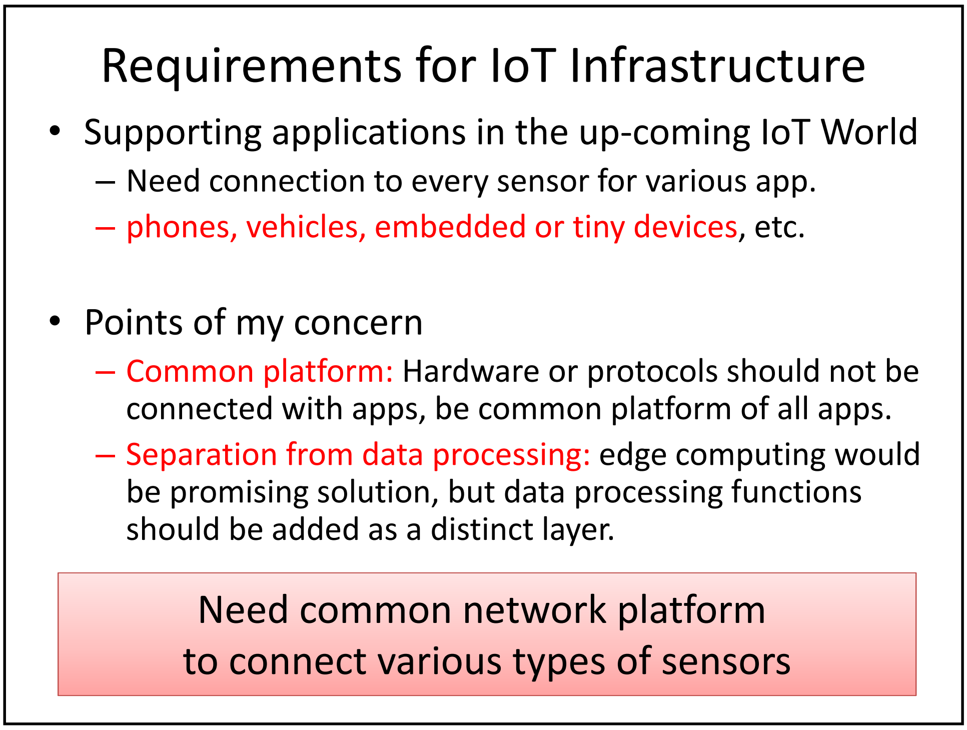
A9D71355-6595-1801-B7EA-001C4D678E7C,2,2	-57
A9D71355-6595-1801-B7EA-001C4D678E7C,2,5	-59
A9D71355-6595-1801-B7EA-001C4D678E7C,1,5	-68
A9D71355-6595-1801-B7EA-001C4D678E7C,1,3	-71
A9D71355-6595-1801-B7EA-001C4D678E7C,1,2	-67
A9D71355-6595-1801-B7EA-001C4D678E7C,1,4	-68
A9D71355-6595-1801-B7EA-001C4D678E7C,2,7	-56
A9D71355-6595-1801-B7EA-001C4D678E7C,2,4	-55
A9D71355-6595-1801-B7EA-001C4D678E7C,2,6	-56
A9D71355-6595-1801-B7EA-001C4D678E7C,516,1	-60



Toward the Future Internet Architecture for IoT Services

2016.8.30　IWIN2016 Panel

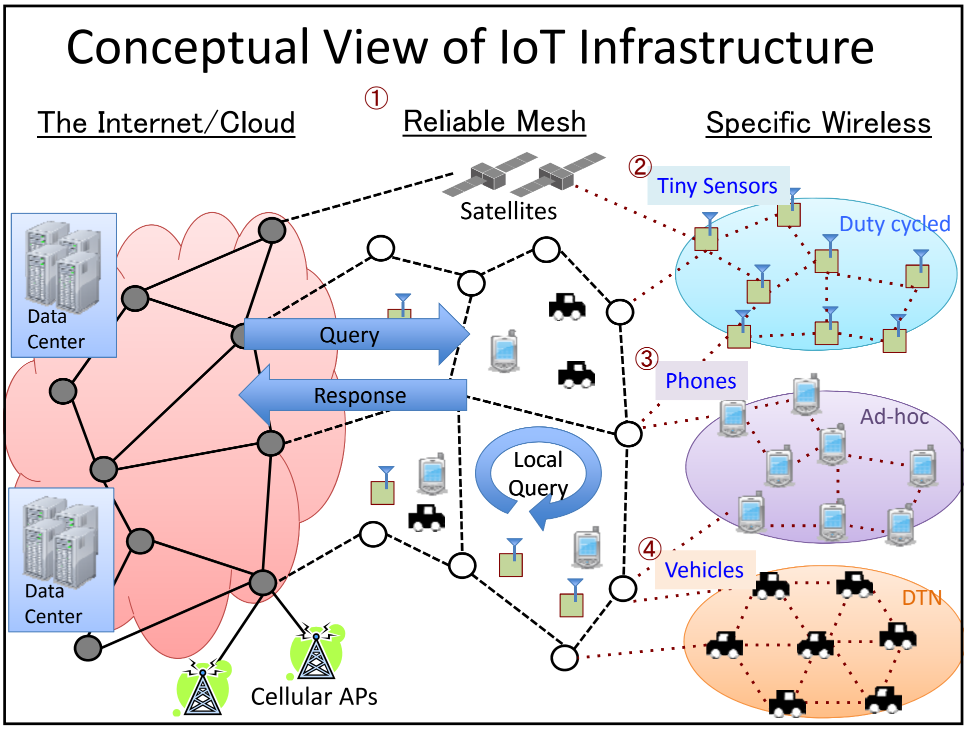
Takuya Yoshihiro
(Wakayama University)



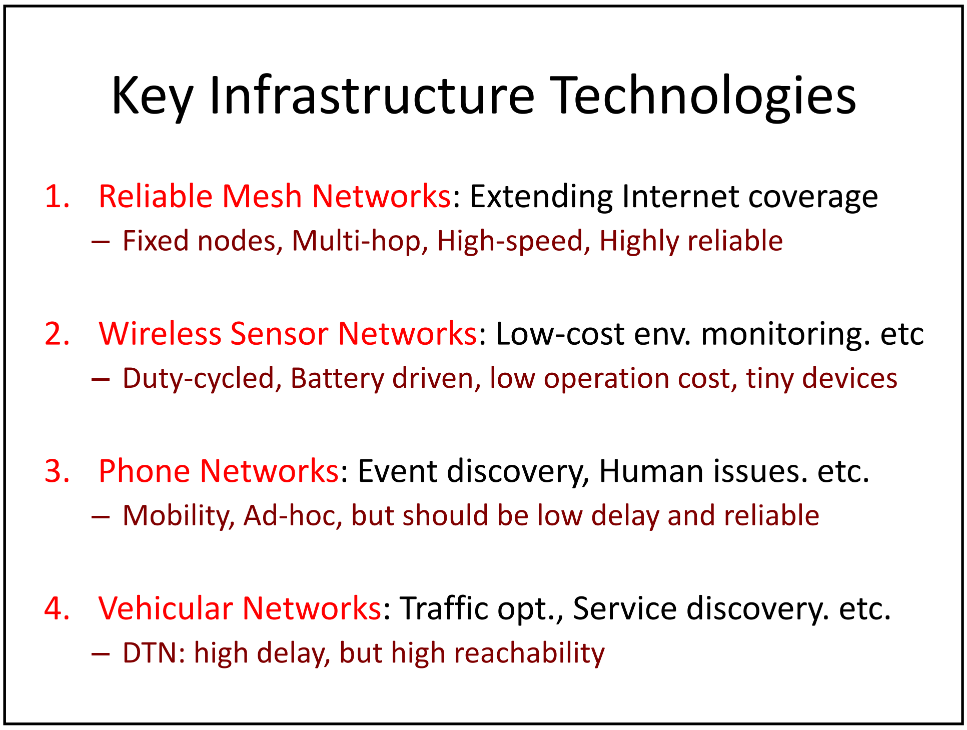
Requirements for IoT Infrastructure

- Supporting applications in the up-coming IoT World
  - Need connection to every sensor for various app.
  - phones, vehicles, embedded or tiny devices, etc.
- Points of my concern
  - Common platform: Hardware or protocols should not be connected with apps, be common platform of all apps.
  - Separation from data processing: edge computing would be promising solution, but data processing functions should be added as a distinct layer.

Need common network platform
to connect various types of sensors

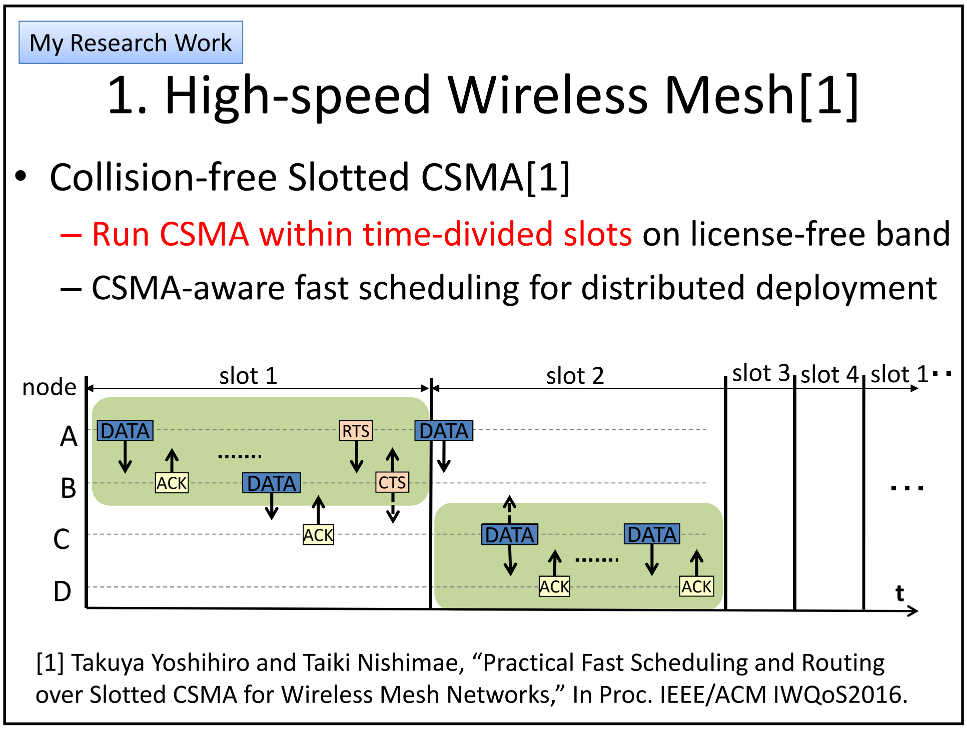


Conceptual View of IoT Infrastructure



Key Infrastructure Technologies

1. Reliable Mesh Networks: Extending Internet coverage
   - Fixed nodes, Multi-hop, High-speed, Highly reliable
2. Wireless Sensor Networks: Low-cost env. monitoring. etc
   - Duty-cycled, Battery driven, low operation cost, tiny devices
3. Phone Networks: Event discovery, Human issues. etc.
   - Mobility, Ad-hoc, but should be low delay and reliable
4. Vehicular Networks: Traffic opt., Service discovery. etc.
   - DTN: high delay, but high reachability

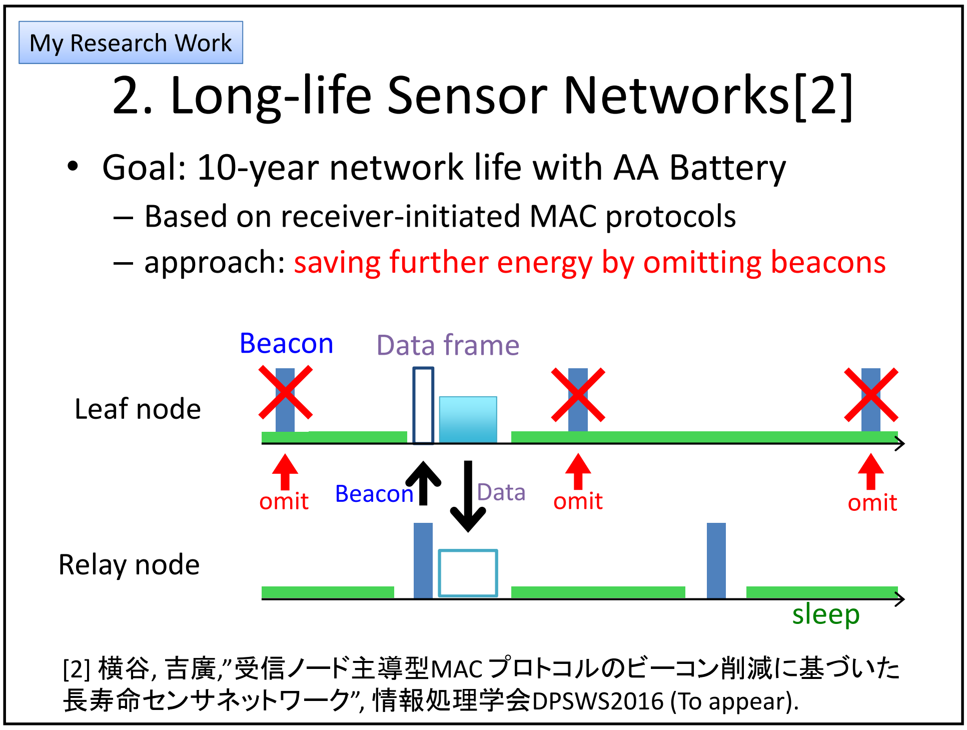


My Research Work

1. High-speed Wireless Mesh[1]

- Collision-free Slotted CSMA[1]
  - Run CSMA within time-divided slots on license-free band
  - CSMA-aware fast scheduling for distributed deployment

[1] Takuya Yoshihiro and Taiki Nishimae, "Practical Fast Scheduling and Routing over Slotted CSMA for Wireless Mesh Networks," In Proc. IEEE/ACM IWQoS2016.

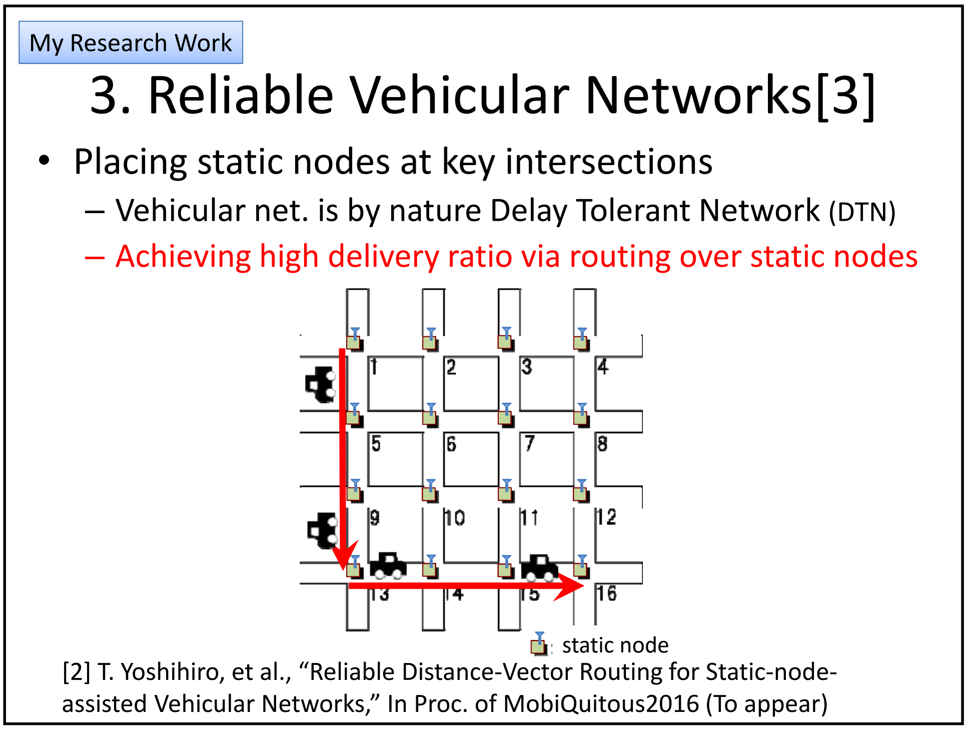


My Research Work

2. Long-life Sensor Networks[2]

- Goal: 10-year network life with AA Battery
  - Based on receiver-initiated MAC protocols
  - approach: saving further energy by omitting beacons

[2] 横谷, 吉廣,"受信ノード主導型MAC プロトコルのビーコン削減に基づいた長寿命センサネットワーク", 情報処理学会DPSWS2016 (To appear).

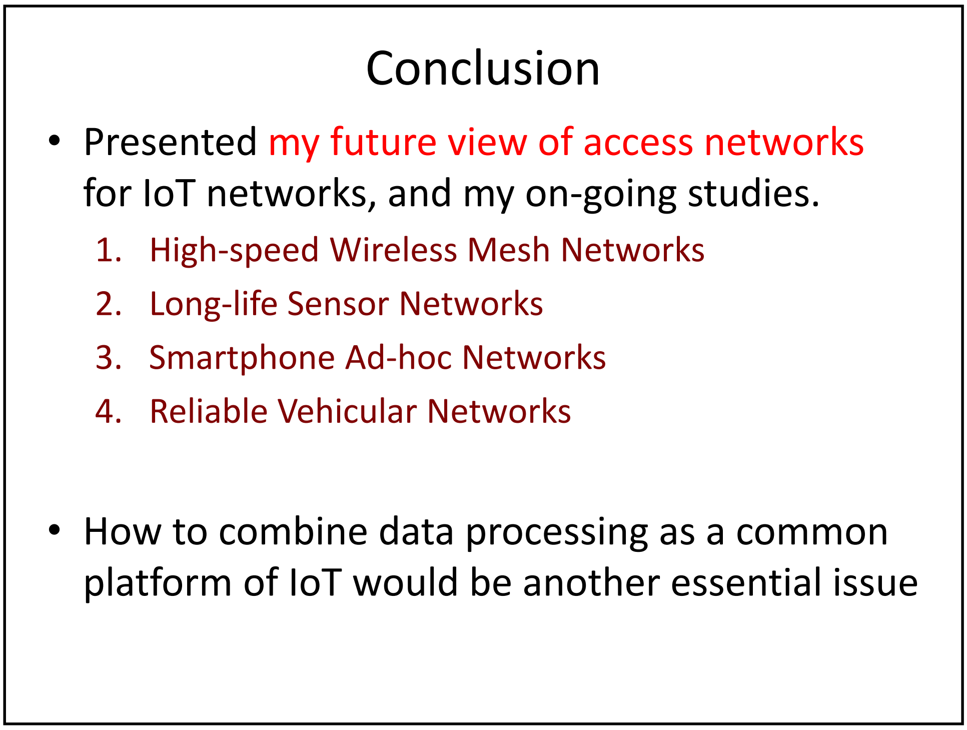


My Research Work

3. Reliable Vehicular Networks[3]

- Placing static nodes at key intersections
  - Vehicular net. is by nature Delay Tolerant Network (DTN)
  - Achieving high delivery ratio via routing over static nodes

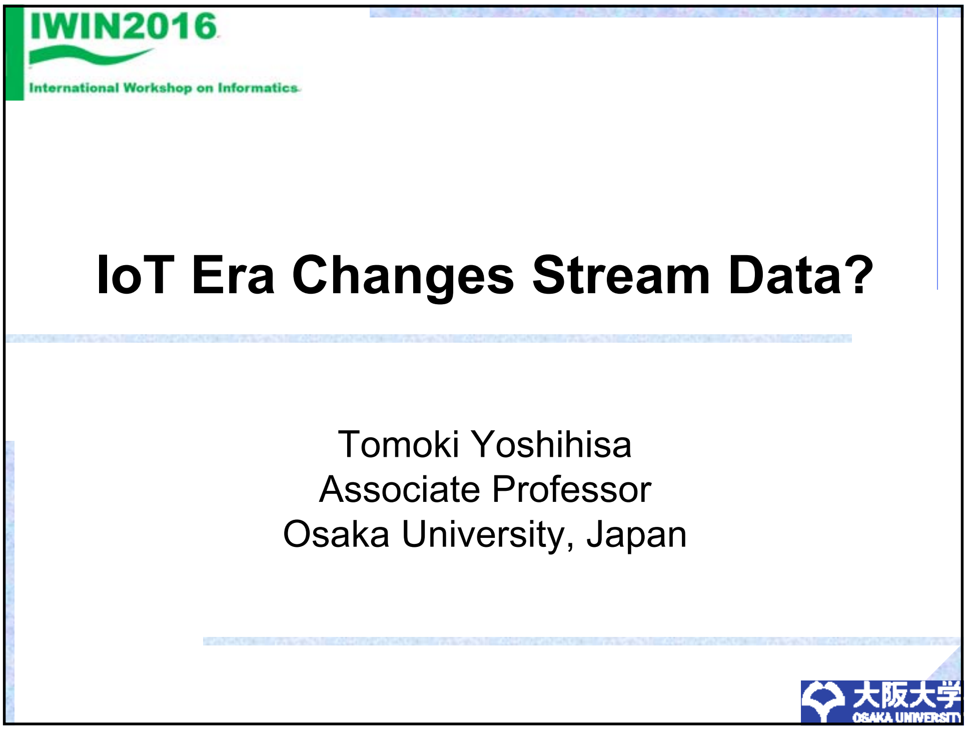
[2] T. Yoshihiro, et al., “Reliable Distance-Vector Routing for Static-node-assisted Vehicular Networks,” In Proc. of MobiQuitous2016 (To appear)



Conclusion

- Presented my future view of access networks for IoT networks, and my on-going studies.
  1. High-speed Wireless Mesh Networks
  2. Long-life Sensor Networks
  3. Smartphone Ad-hoc Networks
  4. Reliable Vehicular Networks

- How to combine data processing as a common platform of IoT would be another essential issue



IWIN2016

International Workshop on Informatics

IoT Era Changes Stream Data?

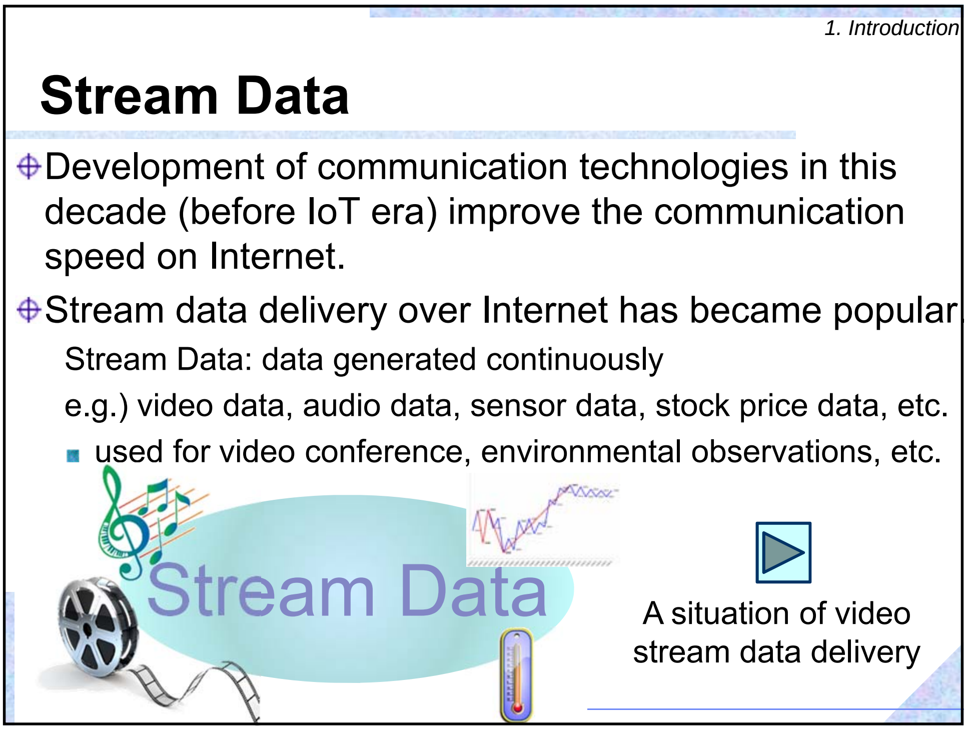
Tomoki Yoshihisa
Associate Professor
Osaka University, Japan

大阪大学
OSAKA UNIVERSITY

1. Introduction

Self-Introduction

- Research field:
  - Stream data (video, sensor) delivery systems, protocols, applications
- Main papers:
  - IEEE Trans. on Broad. [impact factor 2.7], IEEE COMPSAC [acceptance rate 22%], ACM SAC [acceptance rate 24%], IEEE WCNC, Pervasive, ACM SIGGRAPH (Demo)
- Personal history:
  - Assistant Professor at Kyoto University, Japan since 2005.
  - Associate Professor at Osaka University, Japan since 2009.
    (During 2009, Visiting Researcher at University of California, Irvine, USA)
- Awards:
  - ICMU (2 times), IEEE PACRIM golden paper (2007), Osaka University Presidential Awards (2014)
    many domestic conference awards

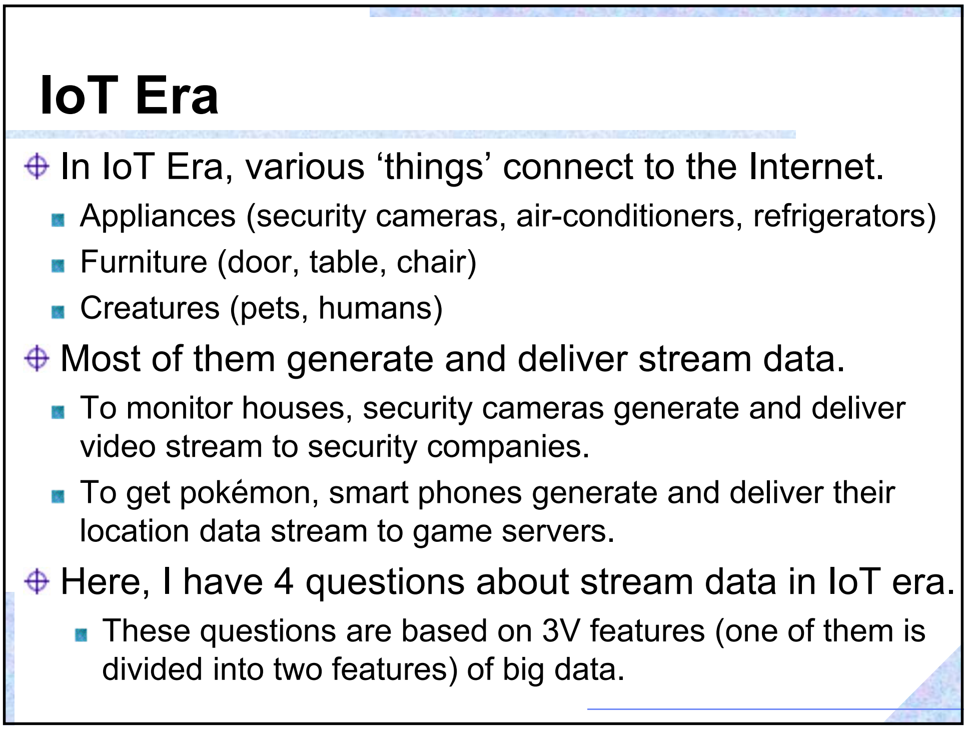


1. Introduction

Stream Data

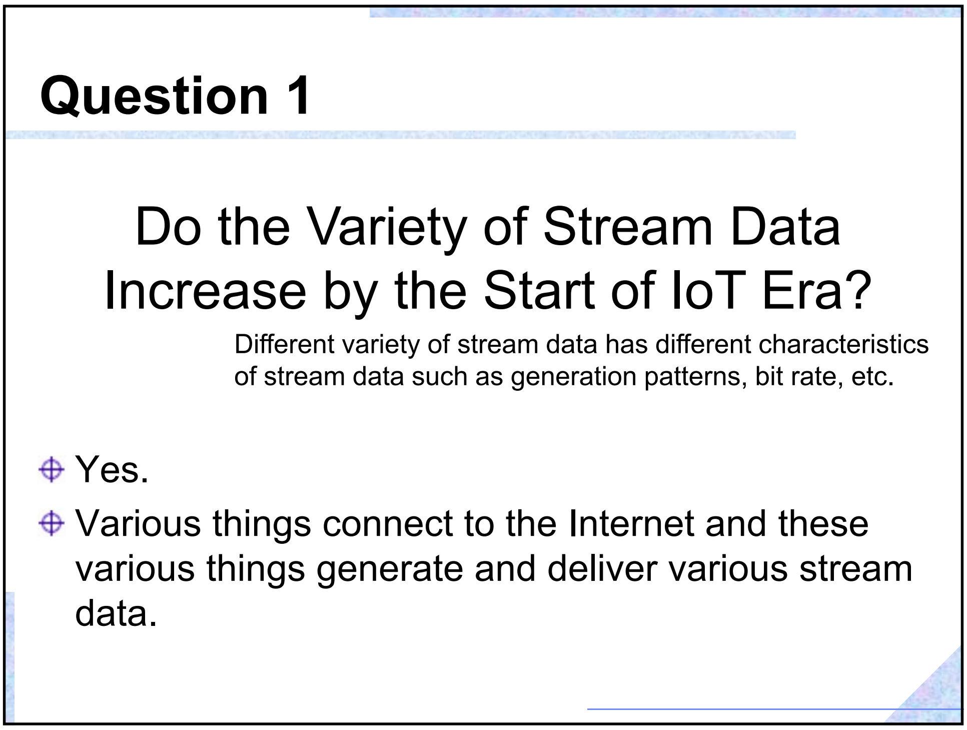
- Development of communication technologies in this decade (before IoT era) improve the communication speed on Internet.
- Stream data delivery over Internet has became popular.
  - Stream Data: data generated continuously
  - e.g.) video data, audio data, sensor data, stock price data, etc.
    - used for video conference, environmental observations, etc.

A situation of video stream data delivery



IoT Era

- In IoT Era, various ‘things’ connect to the Internet.
  - Appliances (security cameras, air-conditioners, refrigerators)
  - Furniture (door, table, chair)
  - Creatures (pets, humans)
- Most of them generate and deliver stream data.
  - To monitor houses, security cameras generate and deliver video stream to security companies.
  - To get pokémon, smart phones generate and deliver their location data stream to game servers.
- Here, I have 4 questions about stream data in IoT era.
  - These questions are based on 3V features (one of them is divided into two features) of big data.

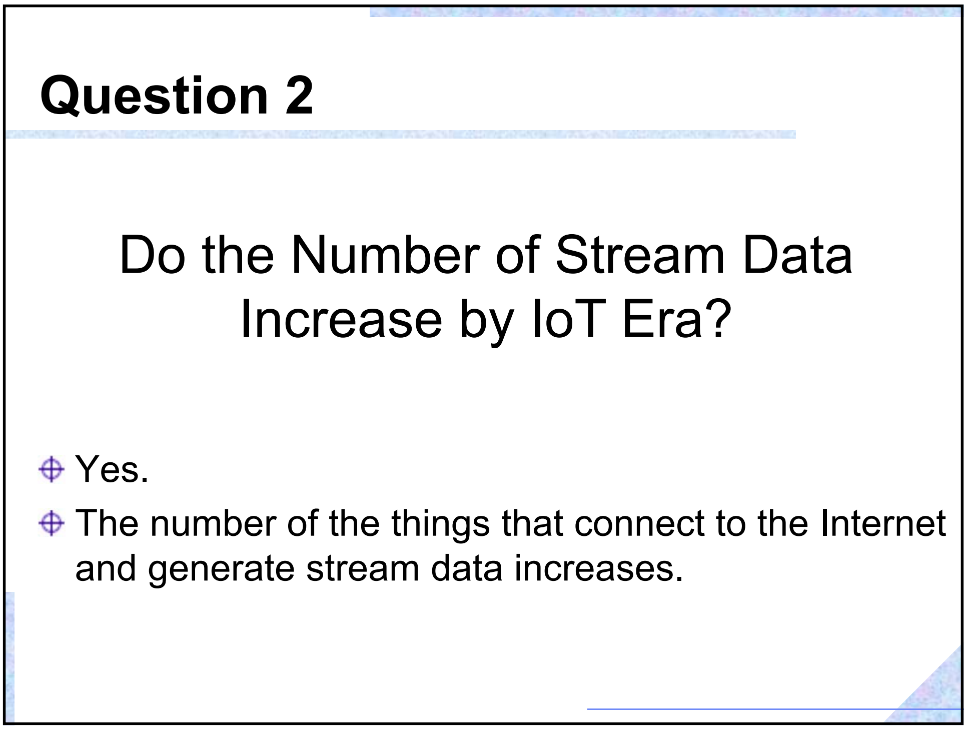


Question 1

Do the Variety of Stream Data Increase by the Start of IoT Era?

Different variety of stream data has different characteristics of stream data such as generation patterns, bit rate, etc.

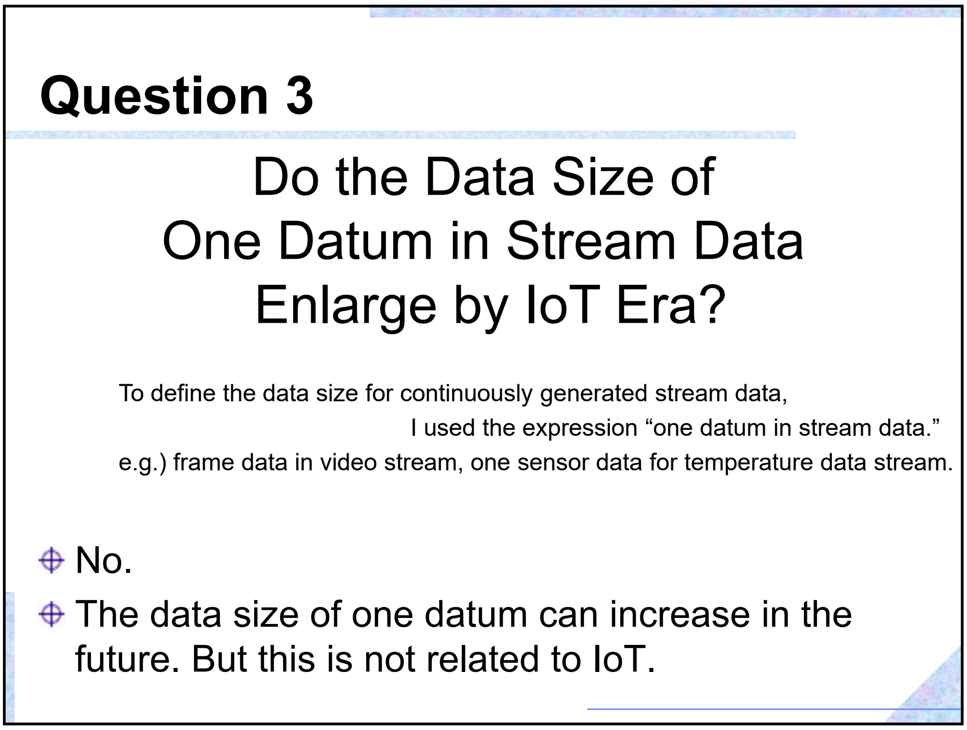
- Yes.
- Various things connect to the Internet and these various things generate and deliver various stream data.



Question 2

Do the Number of Stream Data Increase by IoT Era?

- Yes.
- The number of the things that connect to the Internet and generate stream data increases.

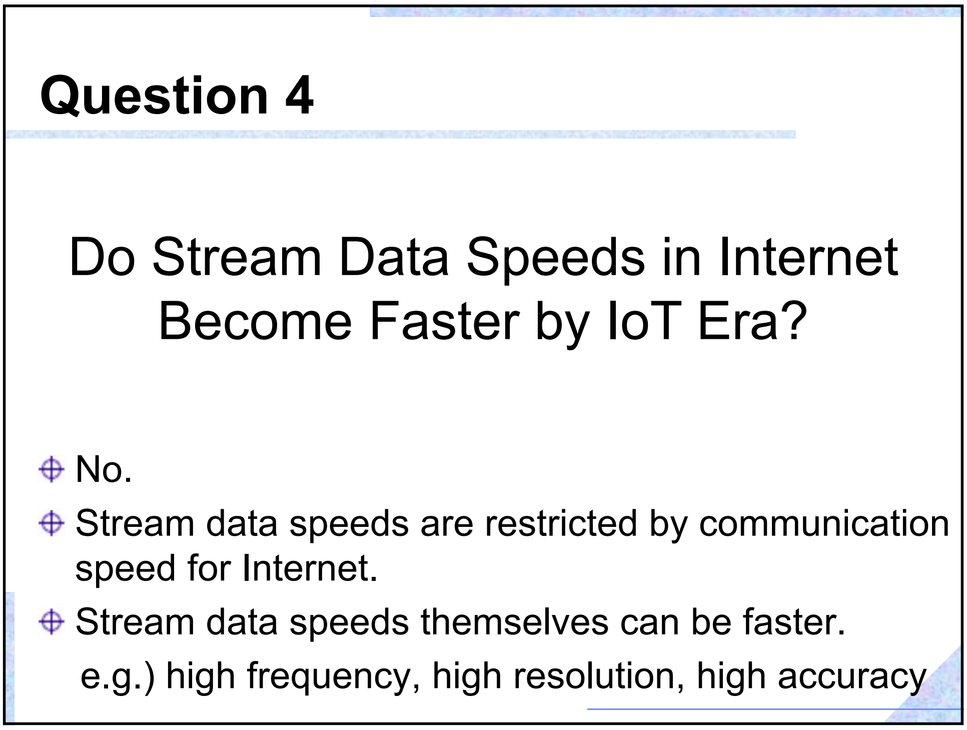


Question 3

Do the Data Size of
One Datum in Stream Data
Enlarge by IoT Era?

To define the data size for continuously generated stream data,
I used the expression "one datum in stream data."
e.g.) frame data in video stream, one sensor data for temperature data stream.

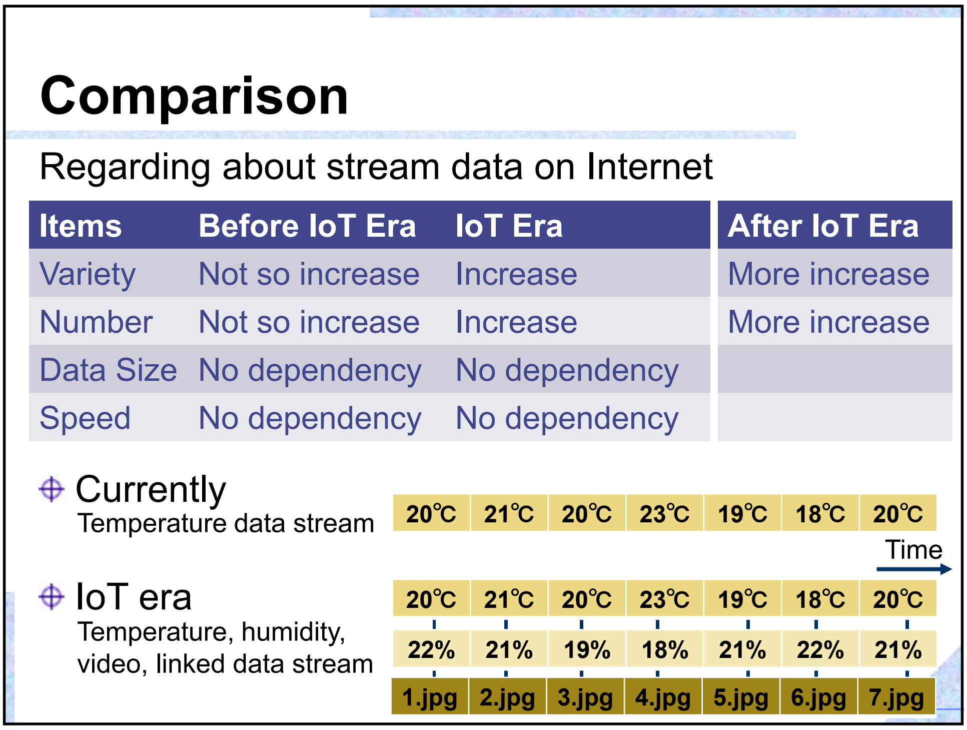
- No.
- The data size of one datum can increase in the future. But this is not related to IoT.



Question 4

Do Stream Data Speeds in Internet
Become Faster by IoT Era?

- No.
- Stream data speeds are restricted by communication speed for Internet.
- Stream data speeds themselves can be faster.
  e.g.) high frequency, high resolution, high accuracy



Comparison

Regarding about stream data on Internet

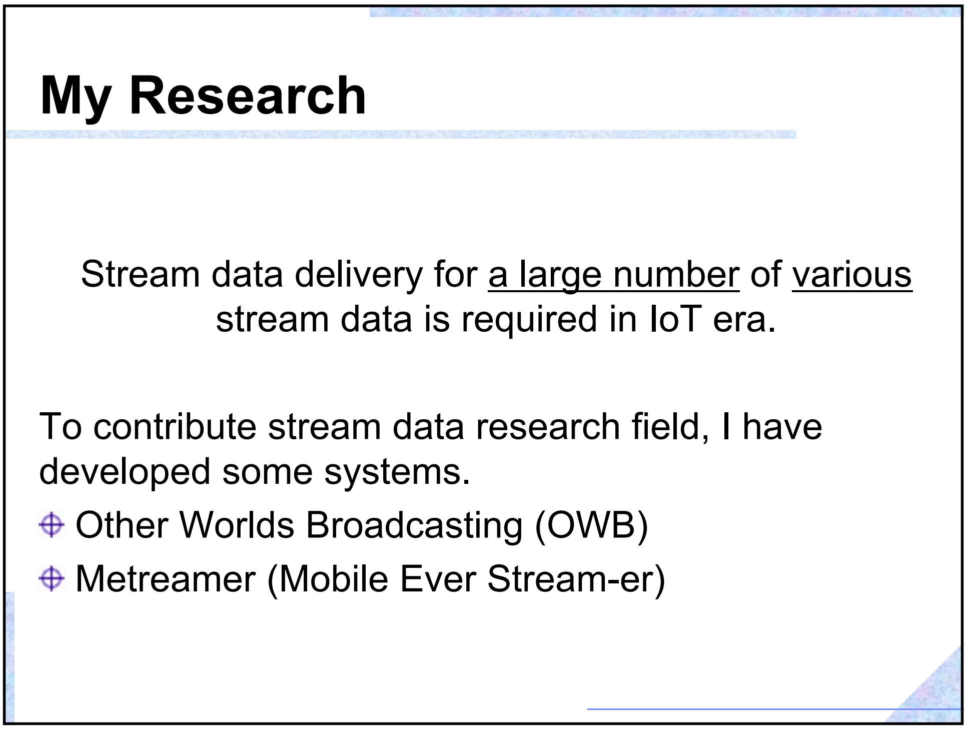
Items	Before IoT Era	IoT Era	After IoT Era
Variety	Not so increase	Increase	More increase
Number	Not so increase	Increase	More increase
Data Size	No dependency	No dependency	
Speed	No dependency	No dependency	

- Currently
  Temperature data stream

20℃	21℃	20℃	23℃	19℃	18℃	20℃

- IoT era
  Temperature, humidity, video, linked data stream

20℃	21℃	20℃	23℃	19℃	18℃	20℃
22%	21%	19%	18%	21%	22%	21%
1.jpg	2.jpg	3.jpg	4.jpg	5.jpg	6.jpg	7.jpg

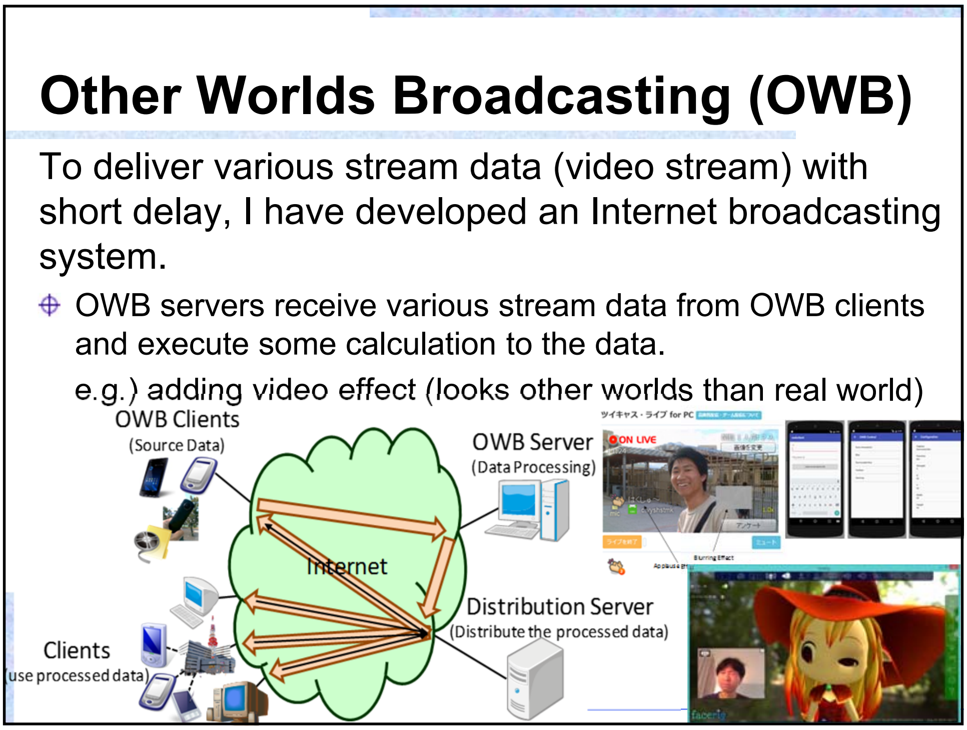


My Research

Stream data delivery for a large number of various stream data is required in IoT era.

To contribute stream data research field, I have developed some systems.

- Other Worlds Broadcasting (OWB)
- Metreamer (Mobile Ever Stream-er)

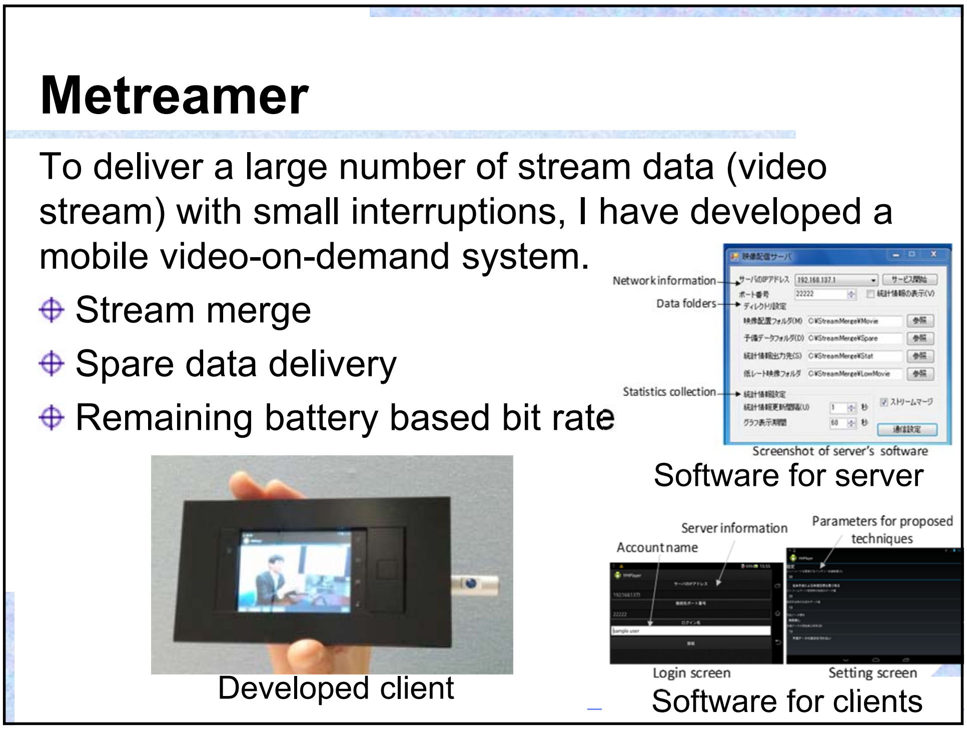


Other Worlds Broadcasting (OWB)

To deliver various stream data (video stream) with short delay, I have developed an Internet broadcasting system.

- OWB servers receive various stream data from OWB clients and execute some calculation to the data.

  e.g.) adding video effect (looks other worlds than real world)



Metreamer

To deliver a large number of stream data (video stream) with small interruptions, I have developed a mobile video-on-demand system.

- Stream merge
- Spare data delivery
- Remaining battery based bit rate

映像配信サーバ

サーバのIPアドレス 192.168.137.1 サービス開始

ポート番号 22222 □ 統計情報の表示(V)

ディレクトリ設定

映像配置フォルダ(M) C:¥StreamMerge¥Movie 参照

予備データフォルダ(D) C:¥StreamMerge¥Spare 参照

統計情報出力先(S) C:¥StreamMerge¥Stat 参照

低レート映像フォルダ C:¥StreamMerge¥LowMovie 参照

統計情報設定

統計情報更新間隔(U) 1 秒

グラフ表示期間 60 秒

☑ ストリームマージ

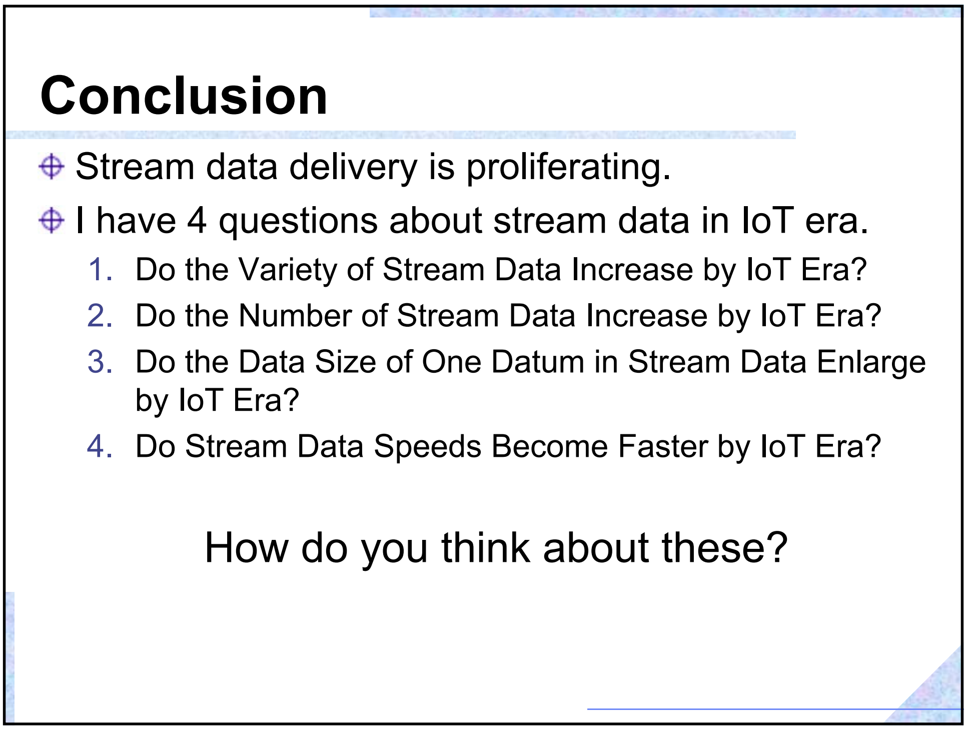
通信設定

Screenshot of server's software

Software for server

Developed client

Software for clients



Conclusion

- Stream data delivery is proliferating.
- I have 4 questions about stream data in IoT era.
  1. Do the Variety of Stream Data Increase by IoT Era?
  2. Do the Number of Stream Data Increase by IoT Era?
  3. Do the Data Size of One Datum in Stream Data Enlarge by IoT Era?
  4. Do Stream Data Speeds Become Faster by IoT Era?

How do you think about these?

Session 6:
Systems and Applications

(Chair: Takuya Yoshihiro)

On Best Time Estimation Method for Phenological Observations Using Geotagged Tweets

Masaki Endo[*,**], Yoshiyuki Shoji[**], Masaharu Hirota[***], Shigeyoshi Ohno[*], and Hiroshi Ishikawa[**]

[*]Division of Core Manufacturing, Polytechnic University, Japan
[**]Graduate School of System Design, Tokyo Metropolitan University, Japan
[***] Department of Information Engineering, National Institute of Technology, Oita College, Japan

Abstract - In recent years, social network services (SNS) such as Twitter have become widely used, attracting great attention for many reasons. An important characteristic of Twitter is its real-time property. Twitter users post huge volumes of Twitter posts (*tweets*) related to daily events in real time. We assume that the tweet contents depend on the region, season, and time of day. Therefore, the possibility exists of obtaining valuable information for tourists from tweets posted during travel. As described in this paper, we propose a method to estimate regional best times for viewing flower blossoms from tweets including flower names. Our proposed method analyzes the number of tweets using a moving average. Additionally, we particularly examine geotagged tweets. Our experiments compare the best time for viewing estimated using our method to the flowering date and the full bloom date of cherry blossoms that the Japan Meteorological Agency has observed and posted. We conducted an experiment using data for the best time for viewing cherry blossoms during 2015 and 2016. Results confirmed that the proposed method can estimate the full bloom period accurately.

Keywords: trend estimation; phenological observation; Twitter

1 INTRODUCTION

In recent years, because of rapid performance improvement and the dissemination of various devices such as smart phones and tablets, diverse and vast data are generated on the web. Particularly, social networking services (SNS) have become prevalent because users can post data and various messages easily. According to the 2014 Communications Usage Trend Survey of the Ministry of Internal Affairs and Communications (MIC) [1], the percentage of Japanese people aged 13–39 years old using SNS is greater than 60%, the figure for people 40–49 years is higher than 50%. Twitter [2], an SNS that provides a micro-blogging service, is used as a real-time communication tool. Numerous tweets have been posted daily by vast numbers of users. Twitter is therefore a useful medium to obtain, from a large amount of information posted by many users, real-time information corresponding to the real world.

Here, we describe the provision of information to tourists using the web. Before SNS were used, local governments, tourism organizations, and travel companies provided regional tourism information using web pages. After SNS became widely used, they also undertook efforts to disseminate more detailed information related to respective tourist spots. The information is useful for tourists, but providing timely and topical travel information entails high costs for the information provider because they must update the information continually. Today, providing reliable information related to local travel is not only strongly demanded by tourists, but also local governments, tourism organizations, and travel companies, which bear high costs of providing the information.

Tourists also want real-time information and local unique seasonal information posted on web sites, according to a survey study of IT tourism and services to attract customers [3] by the Ministry of Economy, Trade and Industry (METI). Current web sites provide similar information in the form of guide books. Nevertheless, the information update frequency is low. Because each local government, tourism association, and travel company provides information about travel destination local unit independently, it is difficult for tourists to collect information for “now” tourist spots.

Therefore, providing current, useful, real-world information for travelers by capturing the change of information in accordance with the season and time zone of the tourism region is important for the travel industry. As described herein, we define "now" as information for tourism and disaster prevention required by travelers during travel, such as a best flower-viewing time and festivals and local heavy rains.

We propose a method to estimate the best time for phenological observations for tourism such as the best time for viewing cherry blossoms and autumn leaves in each region by particularly addressing phenology observations assumed for “now” in the real world. Tourist information for the best time requires a peak period, which means that the best time is not a period after and before falling flowers, but a period to view blooming flowers. Furthermore, the best times differ among regions and locations. Therefore, it is necessary to estimate a best time of phenological observation for each region and location. To estimate the best time for viewing, we must collect much information having real-time properties. For this study, we use Twitter data obtained for many users throughout Japan.

The remainder of the paper is organized as follows. Chapter 2 presents earlier research related to this topic. Chapter 3 describes our proposed method for estimation of best time of phenological observations. Chapter 4 describes experimentally obtained results for our proposed method and a discussion of the results. Chapter 5 summarizes the contributions and future work.

2 RELATED WORK

The amount of digital data is expected to increase greatly in the future because of the spread of SNS. Reports describing studies of the effective use of these large amounts of digital data are numerous. Some studies use microblogs to conduct real-world understanding and prediction by analyzing information transmitted from microblogs. Kleinberg [5] detected a "burst" of keywords signaling a rapid increase in time-series data. Ochiai et al. [6] proposed a disambiguation method for family names that are also used as place names using dynamic characteristic words of topics that vary from period to period, including static characteristic words and locations that are independent of specific seasonal variation according to the location as a target of microblog. Kurata et al. [7] developed a system to detect events in real space using geotagged tweets. This system can grasp what events occur in time and place by the top 10 of frequent word extraction conducted in each time zone. Sakaki et al. [8] proposed a method to detect events such as earthquakes and typhoons based on a study estimating real-time events from Twitter. Consequently, various methods for extracting event and location information are discussed. However, a method used to estimate the start and end of the full bloom period of phenological observations using tweets is controversial.

3 OUR PROPOSED METHOD

This chapter presents a description of a method of analysis for target data collection and our best time estimation to get a guide for phenological change from Twitter in Japan. Our proposal is portrayed in Fig. 1.

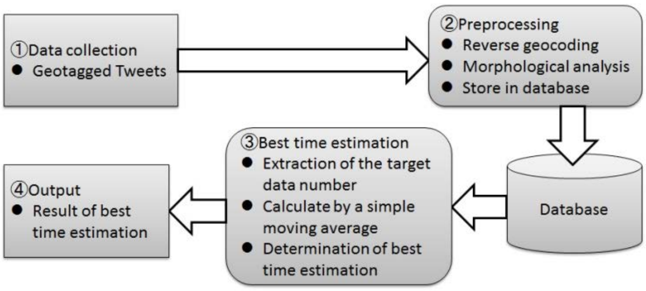

①Data collection
● Geotagged Tweets
②Preprocessing
● Reverse geocoding
● Morphological analysis
● Store in database
Database
③Best time estimation
● Extraction of the target data number
● Calculate by a simple moving average
● Determination of best time estimation
④Output
● Result of best time estimation

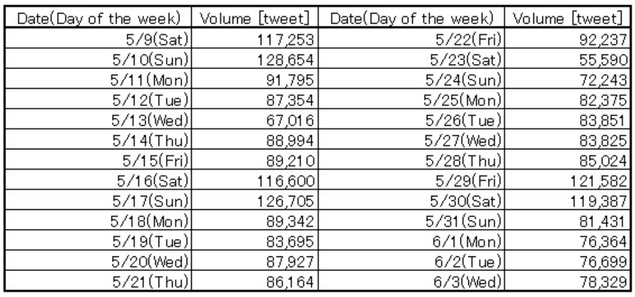
Figure 1: Our proposal summary.

3.1 Data collection

This section presents a description of the Method of (1) data collection shown in Fig. 1. Geotagged tweets sent from Twitter are a collection target. Range geo-tagged tweets include the Japanese archipelago (120.0 ≤ longitude ≤ 154.0 and 20.0 ≤ latitude ≤ 47.0) as the collection target. Collection of these data was done using Streaming API [9], one API provided by Twitter, Inc.

Next, we describe the collected number of data. The percentage of geotagged tweets among tweets originated in Japan, according to the study of Hashimoto et al. [10] as a whole is about 0.18%. Such tweets are very few among all data. However, the collected geo-tagged tweets, shown as an example in Table 1, number about 70,000, even on weekdays. On weekends there are also days on which more than 100,000 such messages are posted. We use about 30 million geo-tagged tweets from 2015/2/17 through 2016/4/30. For each day of collection, the number during the period covered was about 72,000. We calculated the best time for flower viewing, as estimated by processing the following sections using these data.

Table 1: Transition example of geotagged tweets (2015/5/9-6/4)



Date(Day of the week)	Volume [tweet]	Date(Day of the week)	Volume [tweet]
5/9(Sat)	117,253	5/22(Fri)	92,237
5/10(Sun)	128,654	5/23(Sat)	55,590
5/11(Mon)	91,795	5/24(Sun)	72,243
5/12(Tue)	87,354	5/25(Mon)	82,375
5/13(Wed)	67,016	5/26(Tue)	83,851
5/14(Thu)	88,994	5/27(Wed)	83,825
5/15(Fri)	89,210	5/28(Thu)	85,024
5/16(Sat)	116,600	5/29(Fri)	121,582
5/17(Sun)	126,705	5/30(Sat)	119,387
5/18(Mon)	89,342	5/31(Sun)	81,431
5/19(Tue)	83,695	6/1(Mon)	76,364
5/20(Wed)	87,927	6/2(Tue)	76,699
5/21(Thu)	86,164	6/3(Wed)	78,329

3.2 Preprocessing

This section presents a description of the method of (2) preprocessing shown in Fig. 1. Preprocessing includes reverse geocoding and morphological analysis, as well as database storage for data collected through the processing described in Section 3.1.

Reverse geocoding identified prefectures and municipalities by town name from latitude and longitude information of the individually collected tweets. We use a simple reverse geocoding service [11] available from the National Agriculture and Food Research Organization in this process: e.g., (latitude, longitude) = (35.7384446, 139.460910) by reverse geocoding becomes (Tokyo, Kodaira City, Ogawanishi-cho 2-chome).

Morphological analysis divides the collected geo-tagged tweet morphemes. We use the “Mecab” morphological analyzer [12]. By way of example, "桜は美しいです” (in English “Cherry blossoms are beautiful.”)" is divided into "(桜 / noun), (は / particle), (美しい / adjective), (です / auxiliary verb), (。 / symbol)".

Preprocessing performs the necessary data storage for the best time to view, as estimated from results of the processing of the data collection and reverse geocoding and morphological analysis. Data used for this study were the tweet ID, tweet post time, tweet text, morphological analysis result, latitude, and longitude.

3.3 Estimating the best time for viewing

This section presents a description of the method of (3) best time estimation presented in Fig. 1. Our estimation method for the best time to view is one that processes the target number of extracted data and calculates a simple moving average, yielding an inference of the best time to view the flowers. The method defines a word related to the best time for viewing, estimated as the target word. The target word is a word including Chinese characters, hiragana, and katakana, which represents an organism name and seasonal change, as shown in Table 2.

Table 2: Examples of the target word

Items	Target Words	In English
さくら	桜, さくら, サクラ	Cherry blossoms
かえで	楓, かえで, カエデ	Maple
いちょう	銀杏, いちょう, イチョウ	Ginkgo
こうよう	紅葉, 黄葉, こうよう, もみじ, コウヨウ, モミジ	Autumn leaves

Next, we describe the simple moving average calculation, which uses a moving average of the standard of the best time to view judgment. It calculates a simple moving average using aggregate data on a daily basis by the target number of data extraction described above. Fig. 2 presents an overview of the simple moving average of the number of days.

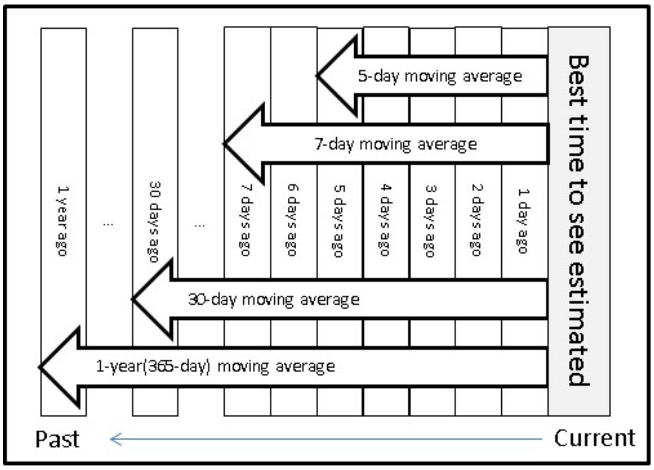


Figure 2: Number of days simple moving average.

We calculate the simple moving average in formula (1) using the number of data going back to the past from the day before the estimated date of the best time for viewing.

$$X(Y) = \frac{P_1 + P_2 + \cdots + P_Y}{Y} \qquad (1)$$

$X(Y)$: Y day moving average

P_n: Number of data of n days ago

Y: Calculation target period

The standard length of time we used for the simple moving average is a 7-day moving average and 1-year moving average. A 7-day moving average has been one week of the criteria of the period in full bloom estimated because a tendency exists for a transition of geotagged tweets, as shown in Table 1 of the above-described increases on weekends than on weekdays. In addition, phenological observations are based on the moving average of a best time for viewing estimated in past years because there are many such “viewing” events every year: cherry blossom viewing, autumn leaf viewing, and even moon viewing.

Next, we describe a simple moving average of the number of days specified for each organism to compare the 7-day moving average and a one-year moving average. In this study, because the best time to view the period varies depending on the specified organism, the individual organism, and the number of days from the biological period.

As an example, we describe cherry blossoms. The Japan Meteorological Agency [13] carries out phenological observations of "Sakura," which yields two output items of the flowering date and the full bloom date observation target. "Sakura of flowering date" [14] is the first day of blooming 5–6 or more wheels of flowers of a specimen tree. "Sakura in full bloom date" is the first day of a state in which about 80% or more of the buds are open in the specimen tree. In addition, "Sakura" is the number of days from general flowering until full bloom: about 5 days. Therefore, "Sakura" in this study uses a 5-day moving average, which is standard.

Next, we describe an estimated judgment of the best time for viewing, which was calculated using the simple moving average (7-day moving average, 1-year moving average, and another biological moving average). It specifies the two conditions as a condition of an estimated decision for the best time for viewing. Condition 1 is the number of data one day before expression. Formula 2 is a simple moving average greater than that of the estimated best time to view date. Condition 2 is the case that follows formulas 3 ((A) / (2)) or more. The short number of days by comparison of the 7-day moving average and another biological moving average is A. A long number of days is B.

$$P_1 \geqq X(365) \qquad (2)$$

$$X(A) \geqq X(B) \qquad (3)$$

Our inference of the best time for viewing was estimated for conditions 1 and 2. Actually, the best time for viewing both holds day to the estimated date for the best time for viewing.

3.4 Output

This section presents a description of the method of (4) output shown in Fig. 1. Output can be visualized using a result of the best time for viewing, as estimated by processing explained in the previous section. This paper presents a visualization that reflects the best time to view inference results in a time-series graph. The graph shows the number of data and the date, respectively, on the vertical axis to the horizontal axis. We are striving to develop useful visualization techniques for travelers.

4 EXPERIMENTS

This chapter presents a description of the experiment to infer the best time to view flowers for the proposed method described in Chapter 3. It shows the dataset to be used for inference of the best time for viewing full blooming flowers in Section 4.1, with similar results shown for 2015 cherry blossom viewing in Section 4.2, and 2016 cherry blossom viewing in Section 4.3.

4.1 Dataset

Datasets used for this experiment were collected using streaming API, as described for data collection in Section 3.1. Data are geo-tagged tweets from Japan during 2015/2/17 – 2016/4/17. The data include about 27 million items. We are using these datasets for experiments to infer the best time for cherry blossom viewing in 2015 (shown in Section 4.2) and in 2016 (shown in Section 4.3).

4.2 Estimation experiment for best time to view cherry blossoms

The estimation experiment to ascertain the best time to view cherry blossoms uses the target word in Table 2: "Sakura". The target word is “cherry blossom,” which is “桜” and “さくら” and “サクラ” in Japanese. The experimental target areas were "Tokyo," "Ishikawa", "Kyoto," and “Hokkaido." For each area, a specimen tree is used for observations by the Japan Meteorological Agency. The cities of “Chiyoda," "Kanazawa," "Kyoto," and "Sapporo" are target areas.

Fig. 3 presents the target area location. Kyoto and Hokkaido are separated by about 1,000 km straight line distance. Kyoto and Tokyo are about 360 km apart. Because of their latitudes, cherry trees flower later in the north in Hokkaido than in Kyoto. Moreover, higher altitudes and consequently cooler temperatures delay flowering even when locations have similar latitudes. Although issues related to altitude were not particularly addressed in this study, they are not expected to affect important results for single sites.

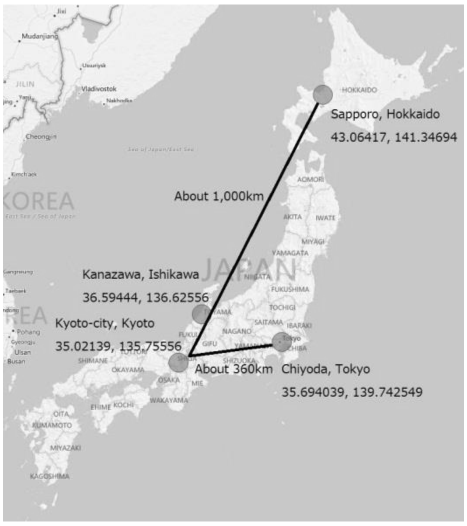


Figure 3: Position of target area.

4.3 Target word results in target areas

Fig. 4 presents experimentally obtained results for the estimated best time for viewing in 2015 using the target word cherry blossoms in the target area of "Tokyo." The dark gray bar in the figure represents the number of tweets. The light gray part represents the period of time it is determined that the best time to view the proposed method. In addition, the solid line shows a 5-day moving average. The dashed line shows a 7-day moving average. The dotted line shows the 1-year moving average. Fig. 5 shows the best time to view estimated experimentally obtained results in 2016 using the target word cherry blossoms in the target area of "Tokyo."

For Tokyo in 2015, as shown in Fig. 4, we obtained the greatest number of data. The greatest number of tweets per day reached about 400. Our proposed method indicates the best time for viewing as 3/23 – 4/3.

Our proposed method shows the best time for viewing as 3/21 – 4/6 in Fig. 5. The estimation for the best time to view in 2016 indicates a longer period than that in 2015, which is consistent with the trend of 2016, with low-temperature days after flowering. Tokyo of 2016, as shown in Fig. 5, also has the largest number of data in the area of the experimental subjects of 2016. More than 1,600 tweets were sent on some days, which is about four times that of 2015. Therefore, the one-year moving average value for the rapid increase in the number of tweets is reduced. For that reason, much noise is included in the estimate of the best time to view.

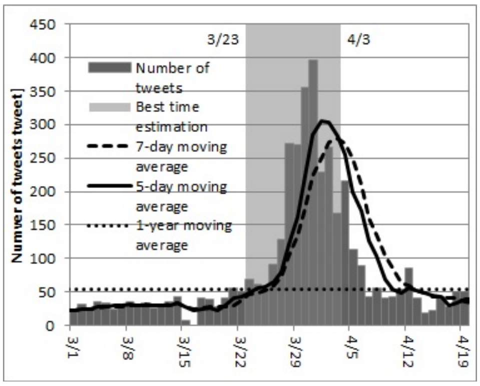


Figure 4: Results of the best time to see, as estimated by Tokyo (2015).

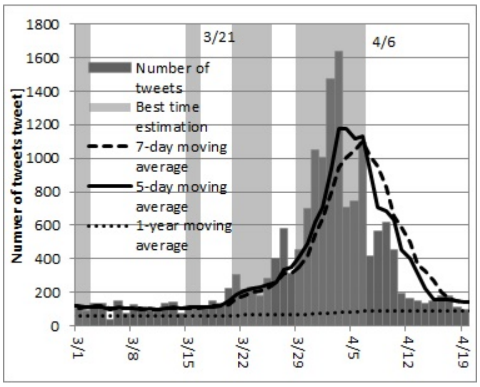


Figure 5: Results of the best time to see, as estimated by Tokyo (2016).

Fig. 6 presents results of 2015 for Ishikawa Prefecture. Results of 2016 for Ishikawa Prefecture are shown in Fig. 7. The greatest numbers of data were, respectively, 15 tweets and 45 tweets. Ishikawa data are far fewer than those of Tokyo. However, 2015 has been the best time to view was estimated as 3/30 – 4/8. In 2016, the best time to view was estimated as 4/1 – 4/6. Noise contents that are unrelated to "Sakura" organisms are also included in tweets. However, before the peak period, tweets abound for budding and flowering cherry. After the peak period, tweets related to cherry blossom leaves are prominent.

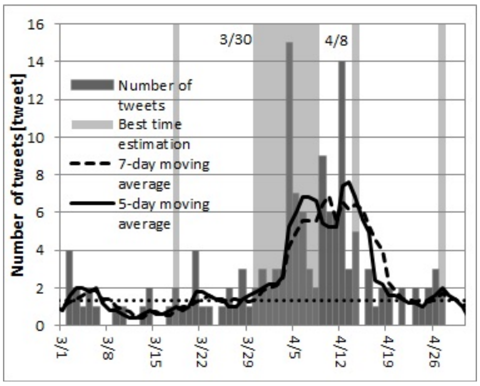


Figure 6: Results of best time to see, as estimated by Ishikawa (2015).

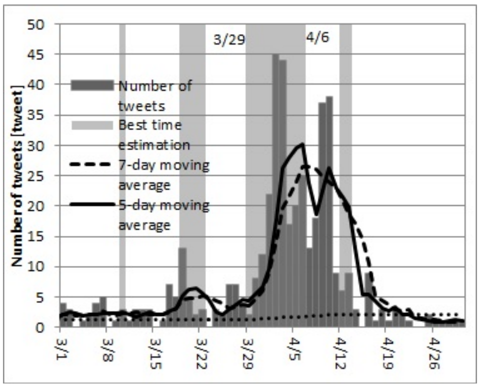


Figure 7: Results of best time to see, as estimated by Ishikawa (2016).

Therefore, the possibility exists of allowing full bloom estimation of the state of more detailed "Sakura" by analyzing the discovered contents. However, these analyses estimate the best time for viewing for each region based only on the target word appearance. No mention is made of the analysis of tweet contents.

4.4 Comparing best time for viewing estimation and observed data

Table 3 presents results of comparison between the best time to view the estimated target area in 2015 and the Japan Meteorological Agency observation data. Dates in the table are the target dates for the estimated best time to view. The thin gray portion of each region is the day determined as the best time for viewing: as an example, Tokyo's best time for viewing in 2015 was 3/23 – 4/3. This result represents the same day and best time to view the estimated thin gray part of the previous section in Fig. 4. Furthermore, the arrow indicates the period of up to "cherry blossoms in full bloom date" from "cherry flowering date" that the Japan Meteorological Agency has observed in each region. As an example, Tokyo observations are based on the specimen tree in Chiyoda. The "Sakura flowering date" of 2015 is 3/23. The "Sakura in full bloom date" of 2015 is 3/29. Specimen trees of the Japan Meteorological Agency of the experimental target area are the following. Ishikawa is Kanazawa. Kyoto is Kyoto City. Hokkaido is Sapporo. Recall and precision using the observed data and best time to view estimated results are calculated for each target area for 2015 from 3/1 – 6/30 using formula (4) and formula (5).

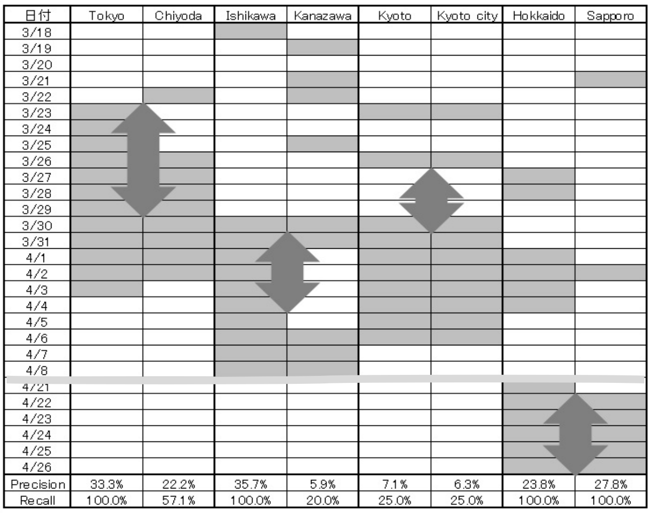
$$Precision = \frac{Number\ of\ days\ to\ match\ the\ observed\ data}{Number\ of\ days\ in\ best\ time\ to\ see\ estimated} \tag{4}$$

$$Recall = \frac{Number\ of\ days\ to\ match\ the\ observed\ data}{Number\ of\ days\ of\ observation\ data} \tag{5}$$

The precision ratio average in Table 3 is about 20%. A low precision ratio does not include the period from full bloom to abscission. The best time for viewing is estimated as 3/30 – 4/3 for Tokyo as determined by the Japan Meteorological Agency as the best time for viewing after full blooming of cherry trees. Therefore, the result presents the possibility of providing the best time to view information that is required to complement tourist observation data of the Japan Meteorological Agency using the proposed method. However, the data are few for areas such as Kanazawa. Therefore, the moving average used to estimate the best time for flower viewing is vulnerable to extreme changes.

Moreover, Hokkaido, Tokyo, Ishikawa, and Kyoto recall is higher than that of municipal districts. These experiments use aggregate data of each whole area against observation data of a sample tree of the Japan Meteorological Agency. Chiyoda and Kanazawa are regions within prefectures. They therefore have a low recall rate because of fewer data. Kyoto and Sapporo show no decrease of recall because many data in the region are city data. Results of this best time estimation should be provided as tourist information in each region for which there is limited information of target areas.

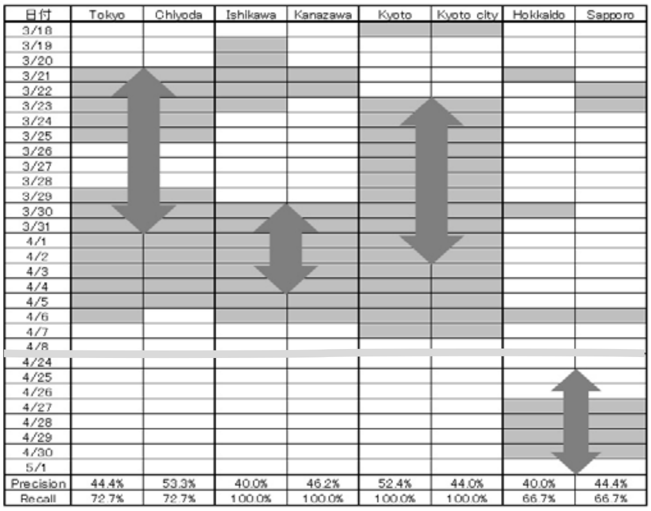
Table 3: Comparison result in target areas of the best time to see the estimated and the observed data (2015)



日付	Tokyo	Chiyoda	Ishikawa	Kanazawa	Kyoto	Kyoto city	Hokkaido	Sapporo
3/18			▨					
3/19				▨				
3/20								
3/21				▨				▨
3/22		▨		▨				
3/23	▨				▨	▨		
3/24	▨							
3/25	▨			▨				
3/26	▨	▨			▨	▨		
3/27	▨	▨					▨	
3/28	▨	▨					▨	
3/29	▨	▨						
3/30	▨	▨	▨	▨	▨	▨		
3/31	▨	▨	▨	▨	▨	▨		
4/1	▨	▨	▨		▨	▨	▨	
4/2	▨	▨	▨		▨	▨	▨	▨
4/3	▨		▨		▨	▨	▨	
4/4			▨		▨	▨	▨	
4/5			▨		▨	▨		
4/6			▨	▨	▨	▨		
4/7			▨	▨				
4/8			▨	▨				
4/21							▨	
4/22							▨	▨
4/23							▨	▨
4/24							▨	▨
4/25							▨	▨
4/26							▨	▨
Precision	33.3%	22.2%	35.7%	5.9%	7.1%	6.3%	23.8%	27.8%
Recall	100.0%	57.1%	100.0%	20.0%	25.0%	25.0%	100.0%	100.0%

Table 4 presents experimentally obtained results for 2016. The notation is the same as that used in Table 3. The experimental period in 2016 was 3/1 – 4/30. Data of 2016 obtained using our proposed method were also confirmed best time estimation for each region. Data confirmed the best time estimation after full bloom observation by the Japan Meteorological Agency. Compared to 2015, 2016 was confirmed to have a long best time period because of low temperatures after flowering. However, precision and recall for some data loss are lower than in 2015.

Table 4: Comparison result in target areas of the best time to see the estimated and the observed data (2016)



日付	Tokyo	Chiyoda	Ishikawa	Kanazawa	Kyoto	Kyoto city	Hokkaido	Sapporo
3/18					▨	▨		
3/19			▨					
3/20			▨					
3/21	▨	▨	▨	▨			▨	
3/22	▨	▨	▨	▨				▨
3/23	▨	▨	▨		▨	▨		▨
3/24	▨	▨			▨	▨		
3/25	▨	▨			▨	▨		
3/26					▨	▨		
3/27					▨	▨		
3/28					▨	▨		
3/29	▨	▨			▨	▨		
3/30	▨	▨	▨	▨	▨	▨	▨	
3/31	▨	▨	▨	▨	▨	▨		
4/1	▨	▨	▨	▨	▨	▨		
4/2	▨	▨	▨	▨	▨	▨		
4/3	▨	▨	▨	▨	▨	▨		
4/4	▨	▨	▨	▨	▨	▨		
4/5	▨	▨	▨	▨	▨	▨		
4/6	▨		▨	▨	▨	▨	▨	▨
4/7					▨	▨		
4/8								
4/24								
4/25								
4/26								
4/27							▨	▨
4/28							▨	▨
4/29							▨	▨
4/30							▨	▨
5/1								
Precision	44.4%	53.3%	40.0%	46.2%	52.4%	44.0%	40.0%	44.4%
Recall	72.7%	72.7%	100.0%	100.0%	100.0%	100.0%	66.7%	66.7%

Therefore, our proposed method is useful for estimating the best time for viewing cherry blossoms in areas where about 10 tweets per day were obtained. However, experimental target areas have fewer data in large areas, where there might be a specimen tree in a prefectural capital used for observations by the Japan Meteorological Agency. Therefore, further verification must be done in other regions.

5 CONCLUSION

This paper proposed the use of Twitter to produce a useful approach for estimating the best time for presenting tourism information related to phenological observations. The proposed method used geotagged tweets with target words and organism names to infer the best times for viewing flowers in Japan. Results of cherry blossom experiments show that seasonal changes in tweets and real-world seasonal changes are relevant to estimates. Therefore, our proposed method presents the possibility of a real-world best time estimated by observation of tweets related to organism names. The granularity of the proposed method differs according to the target word and region. However, best time estimation of each prefecture and city were confirmed for different regions. Results verify the possibility of tourist information presentation with real-time properties for different regions by best time estimation using geotagged tweets. Future studies must verify that the proposed method is applicable to other organisms. Additionally, the method must be expanded to a system by which travelers can obtain event information and disaster information related to travel destinations in real time.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 16K00157, 16K16158, and Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas “Research on social big data.”

REFERENCES

[1] Ministry of Internal Affairs and Communications, 2014 Communications Usage Trend Survey Results, URL ⟨http://www.soumu.go.jp/johotsusintokei/statistics/data/150717_1.pdf⟩ (2015) (in Japanese).

[2] Twitter, URL ⟨https://Twitter.com/⟩ (2014).

[3] Ministry of Economy, Trade and Industry, study of landing type IT tourism and attract customers service, URL ⟨http://www.meti.go.jp/report/downloadfiles/g70629a01j.pdf⟩ (2007) (in Japanese).

[4] Manabu Okumura, Microblog Mining, IEICE Technical Report NLC2011 −59 (2012) (in Japanese).

[5] J. Kleinberg, Bursty and hierarchical structure in stream, In Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1-25 (2002).

[6] K. Ochiai, and D. Torii, Toponym Disambiguation Method for Microblogs Using Time-varying Location-related Words, IPSJ TOD, Vol.7, No.2, pp.51-60 (2014) (in Japanese).

[7] A. Kurata, K. Uehara, and J. Murai, Auto situation detecting system using Twitter, The 75th National Convention of JPSJ, 1V-1, pp.97-98 (2013) (in Japanese).

[8] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, WWW 2010, pp.851-860 (2010).

[9] Twitter Developers, Twitter Developer official site, URL ⟨https://dev.twitter.com/⟩ (2014).

[10] Y. Hashimoto, M. Oka, Statistics of Geo-Tagged Tweets in Urban Areas(<Special Issue>Synthesis and Analysis of Massive Data Flow), JSAI, Vol.27, No.4, pp.424-431 (2012) (in Japanese).

[11] National Agriculture and Food Research Organization, simple reverse geocoding service, URL (http://www.finds.jp/wsdocs/rgeocode/index.html.ja) (2014).

[12] MeCab, Yet Another Part-of-Speech and Morphological Analyzer, URL (http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html) (2012).

[13] Japan Meteorological Agency, Disaster prevention information XML format providing information page, URL ⟨http://xml.kishou.go.jp/⟩ (2011).

[14] Japan Meteorological Agency, observation of Sakura, URL(http://www.data.jma.go.jp/sakura/data/sakura2012.pdf) (2015).

Estimating Pedestrian Location Based on Environmental Knowledge

Keisuke Isomura[†], Yoshino Niwa*, Shogo Shimizu*, Kosuke Yotsuya*, Haruka Iwase*,Yuki Aso*,
Tadanori Mizuno*, Katsuhiko Kaji*

[†]Graduate School of Business Administration and Computer Science, Aichi Institute of Technology, Japan
* Faculty of Information Science, Aichi Institute of Technology, Japan
{b16705bb,kaji}@aitech.ac.jp

Abstract - A method for accurately estimating 3D pedestrian location from sensing data using a smartphone equipped with an accelerometer, gyrometer, and air pressure sensor is proposed. We participated in the PDR Challenge of Ubicomp/ISWC 2015 for the most accurate 3D Pedestrian Dead Reckoning (PDR) algorithm. In this contest, accuracy was calculated from the difference between estimated position and correct position in one-second intervals. Position estimation required number of steps, step length, walking direction, and change of height via stairs. All corridors were linear and all corners were orthogonal in the venue where the PDR Challenge was held, so we assumed linear pedestrian trajectories and left/right turns made at right angles. Additionally, we set step length when walking stairs to a small fixed value matching the generally short length of individual steps. Height estimation took into account the actual height difference between floors to avoid unrealistic values. Our team came in third out of five teams with an average error of 12.96 m. We improved our algorithm after the PDR Challenge reducing average error to 7.94 m.

Keywords: Pedestrian dead reckoning, Smartphone

1 INTRODUCTION

Improved accuracy in location estimation technologies is driving the practical implementation of various types of location information services such as navigation and check-in services. Location estimation techniques using wireless LAN and the Global Positioning System (GPS) are frequently used outdoors. However, poor line-of-sight conditions with satellites and multipath error can prevent GPS-based positioning from being sufficiently accurate, which makes it particularly difficult to use GPS indoors. One powerful technique for estimating a person's location indoors is Pedestrian Dead Reckoning (PDR) [1]–[3]. In contrast to direct detection of one's current position, "Dead Reckoning" is the process of obtaining that position by detecting the amount of movement from an initial position. Elemental technologies making up PDR include those for estimating number of steps, step length, walking direction, and movement between the floors of a building. Past implementations of PDR made use of dedicated sensors affixed to the user's waist, leg, head, etc. However, recent improvements in mobile terminal technologies have driven the spread of smartphones equipped with a variety of sensors. This development enables PDR to be implemented without having to carry around dedicated sensors.

In this paper, we propose a method for achieving high-accuracy PDR by introducing knowledge that can be applied to ordinary indoor walking. Specifically, we assume that corridors are linear, corners are orthogonal, movement between floors is infrequent, and step length when ascending/descending stairs is shorter than usual. In ordinary PDR systems, the accuracy of gyrometers is particularly poor, and there are cases in which even straight-ahead walking is mistaken as turning. Consequently, introducing the knowledge that corridors are linear means that even a situation in which a gyrometer has an offset value will be treated as linear walking.

We already proposed a pedestrian trajectory estimation method based on environmental knowledge [4]. However, the method is offline algorithm, so that it cannot apply realtime indoor positioning. Main contribution of this paper is to expand the method as an online method that is available for realtime indoor positioning.

We participated in the PDR Algorithm Category of the UbiComp/ISWC 2015 PDR Challenge [5] with our proposed method competing with other participants for the most accurate algorithm. In this contest, a subject walked along five prearranged routes while carrying a smartphone equipped with several types of sensors. The accuracy of a PDR algorithm was taken to be the result of calculating the error between the estimated trajectory while walking and the correct trajectory determined from the prearranged route in one-second intervals and taking the average of those values. The contest was held in a building with a simple shape, meaning an environment in which we could apply the building-related knowledge assumed by our method. Specifically, the building featured only orthogonal corners and only one staircase for moving between floors. Using the proposed method, we came in third out of five teams in the PDR Challenge.

2 RELATED WORK

Positioning techniques can be broadly divided into two types: absolute positioning and relative positioning. An absolute positioning technique determines the absolute position of the targeted person using an infrastructure installed outside that person. Positioning by GPS, by an indoor messaging system (IMES), or by WiFi access points can be classified as absolute positioning [6]–[8]. WiFi access points are considered to be advantageous in that commercial Wi-Fi facilities are already installed in great number thereby lowering the hurdle associated with the installation of equipment. In contrast, locations in which no IMES or Wi-Fi access points exist suffer from the costs incurred in deploying and maintaining an infrastructure. Furthermore, in Wi-Fi, the accuracy of positioning can deteriorate owing to the existence of many noise sources and the diffraction, reflection, and absorption of radio signals. In GPS, meanwhile, the presence of structures and obstacles can hamper the reception of radio signals making indoor positioning difficult.

A relative positioning technique, on the other hand, estimates current position by having the user carry sensors whose data can be used to determine the amount of movement from immediately preceding positions [1]–[3]. This technique uses no established infrastructure and therefore has the advantage of not being limited to certain locations. Relative positioning corresponds to navigation techniques that determine one's own position by estimating acceleration during movement from a gyrometer and accelerometer (inertial navigation methods) and calculating velocity and distance by integral calculations. This technology is used in car navigation systems for navigating inside tunnels where GPS signals cannot be received, and it is also used in ship and rocket navigation systems. However, the outside environment can easily affect inertial navigation methods, so their application is limited to objects with simple movement. The PDR method can also be classified as a relative positioning technique. A key characteristic of PDR is that error tends to accumulate and increase as the positioning interval becomes longer. This error can be corrected by such methods as radio-frequency identification (RFID) and map matching. The RFID method can effectively correct location information by using RFID readers to detect a tag with a specific ID and transmitting the installation location or detection time of a detecting reader to the user corresponding to that ID [9]. Low accuracy in estimating step length is also connected to an increase in PDR error.

One technique for resolving the above issues is to calculate step length when walking outdoors by using distance and duration of movement obtained by GPS and the number of steps at that time and to then use that step length when walking indoors. However, step length when walking indoors can differ from that when walking outdoors. In addition, GPS necessarily includes error, and accurate estimation of step length cannot be expected unless the user is in an ideal environment having small multipath effects. There is also a technique using ultrasound waves for calculating step length, but incorporating an ultrasound device in smartphones is not practical.

3 PDR ALGORITHM

The processes used for estimating number of steps, step length, direction of travel, and inter-floor movement are constituent elements of PDR. Accelerometer values are used to estimate number of steps and step length and calculate distance moved, while gyrometer values are used to estimate direction of travel. Combining these two types of values enables location information to be obtained by estimating how far and in what direction the target person has moved on a plane. Furthermore, since estimation of inter-floor movement is necessary to perform 3D PDR, the values of an air pressure sensor are used to estimate height in this study.

3.1 Estimating Number of Steps

Accelerometer values are used to estimate number of steps. An accelerometer obtains values along a total of three axes: two axes in the horizontal direction and one axis in the vertical direction with respect to the ground plane. In this study, instead of being affixed to the user's body, the accelerometer is carried around, which means that it is always in a tilted state. Accordingly, to enable acceleration values to be used without having to take sensor tilt into account and to simplify processing, we denote the three-axes composite value of acceleration as A and the values of acceleration along the X, Y, and Z axes as a_x, a_y, and a_z, respectively, and combine those values using their norm as shown in Eq. 1.

$$A = \sqrt{a_x * a_y * a_z} \tag{1}$$

Accelerometer values, however, include noise owing to factors such as the sensor's accuracy and sensitivity, an unsteady hand, etc. Consequently, if accelerometer values were to be used as such, values displaced by walking could not be distinguished from such noise and estimating number of steps would be difficult. For this reason, we perform a smoothing process by using a low-pass filter to remove high-frequency components and extract only low-frequency components. Denoting the three-axes composite value after correction as $lowA$, the three-axes composite value before correction as A, and elapsed time as t, we express this filtering process by Eq. 2.

$$lowA_t = A * 0.1 + lowA_{t-1} * 0.9 \tag{2}$$

Since composite acceleration is affected only by gravity when the accelerometer is at rest, it will stabilize near 0.9 [g], the approximate value of gravitational acceleration. Acceleration when walking, however, will drop below the value of gravitational acceleration after the body moves upward when taking a step. Then, when stepping on the floor, it will rise above the value of gravitational acceleration due to the accompanying shock and the increase in speed. Thus, as shown in Figure 1, we set thresholds at 1.4 [g] and 0.4 [g] to enclose the value of gravitational acceleration and increment the step count by one whenever the value of acceleration reaches the upper threshold, drops to the lower threshold, and again reaches the upper threshold.

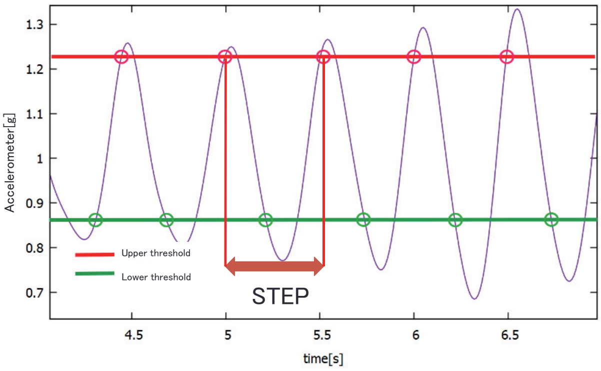


Figure 1: Step estimation using accelerometer.

3.2 Estimating Step Length

Step length constitutes an important piece of data for calculating walking distance, but it differs between pedestrians. One factor that can be offered to explain this is body height since leg length differs according to height. For this reason, body height is often used to estimate step length. Denoting step length as l [m] and body height as h [m], step length can be calculated by Eq. 3.

$$l \approx h - 1 \tag{3}$$

In the event that such height information is not available, calibration data can be obtained beforehand using a PDR-based application. Specifically, if the subject can be asked to hold the smartphone while walking a fixed distance L, that distance can be divided by the estimated number of steps s to give the length of a single step (Eq. 4).

$$l = 14.4/s \tag{4}$$

Step length can therefore be determined as described above by dividing calibration distance by number of steps. However, if error happens to be included in the value estimated for number of steps, error will likewise occur in step length. If location should then be estimated using step length, that error will always be retained when calculating the distance of movement thereby increasing the amount of error. With this in mind, we tried calculating step length using acceleration. Denoting step length as l, maximum acceleration and minimum acceleration of the accelerometer's vertical component per unit time as $max(a_v)$ and $min(_v)$, respectively, and a constant as r_i, we express step length by Eq. 5.

$$l \approx \sqrt{max(a_v) - min(_v)} * r_i \tag{5}$$

However, the results of a preliminary experiment showed that accuracy could not be improved beyond that of Eq. 4. We consider the reason for this to be as follows: since the subject hand-holds the accelerometer, a variety of factors other than walking can be reflected as accelerometer values thereby making step length fluctuate in value. In light of the above, we decided to use Eq. 4 for estimating step length.

On the other hand, a disadvantage in estimating step length by Eq. 4 is that dealing with actual fluctuation in step length is difficult since the trajectory of the walking user is calculated with step length as a fixed value from beginning to end. One location where step length can vary greatly from the norm is a staircase. Human step length generally lies in the range of 60 - 80 [cm], but on a staircase, it takes on a fixed value if walking one step at a time regardless of body height or gait. Since the depth of a single step on a staircase is approximately 30 [cm], step length can be approximated as such. Whether the subject is moving between floors can be estimated from the slope of the regression line plotted from the values of an air pressure sensor as described later. Thus, when estimating that the user is walking on stairs, we change step length from the value previously calculated from Eq. 4 to a fixed value and calculate walking distance accordingly. Then, when estimating that the user has stopped walking on stairs and returned to walking on level ground, we return step length to its original value.

3.3 Estimating Direction

We estimate direction of travel using a gyrometer built into a smartphone. Change in the direction of travel can be computed by integrating the value of angular speed. In this study, we target ordinary buildings and assume a building structure composed only of linear corridors and orthogonal corners. Consequently, to detect a change in the direction of travel, we consider that estimating only whether the user is walking straight ahead or turning at a right angle is sufficient and narrow down user behavior into the four patterns of moving forward, turning right, turning left, and turning around. In judging the direction of rotation, the difference between the value of angular speed measured in real time and that measured 0.5 s earlier corresponds to a right rotation if negative and a left rotation if positive. We use the thresholds shown in Figure 2 to judge the occurrence of right and left turns and a turn-around. For example, the detection of a rotation with a value between 70 ° and 160 ° is judged to be a left turn of 90 °, and a detection of a rotation with a value between -70 ° and -160 ° is judged to be a right turn of 90 °. The reason for using thresholds of 70 ° and -70 °, that is, for setting thresholds ± 20 ° different from a right angle, is to absorb a change in angular speed of about 20 ° to the left or right that may be applied owing to rotation of the torso even when walking straight ahead at the time of measurement.

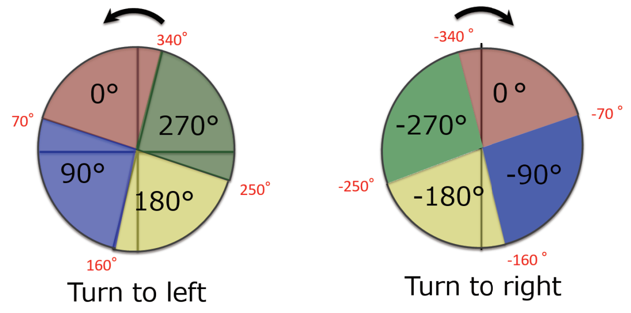


Figure 2: Turn estimation.

3.4 Estimating Height

We use the values obtained from an air pressure sensor to estimate height. First, however, we apply a low-pass filter to remove noise as a form of preprocessing. We also introduce the knowledge that the height of any particular floor is the same throughout that floor [4]. Furthermore, we assume that movement between floors is limited to stairs-elevators and escalators are outside the scope of this study. Finally, we estimate height by combining two processes: judging that the subject is ascending or descending stairs between floors and judging the current floor.

In judging the ascending/descending of stairs, we calculate the slope of the regression line plotted from the time-series values of an air pressure sensor and conclude that the subject is ascending/descending stairs if the value of that slope has risen above or dropped below a certain threshold value.

Here, we use 100 air pressure values at one-second intervals to determine the slope of the regression line. Specifically, we judge that the user is ascending stairs when the slope value is below -0.0877 [hPa/s], descending stairs when the slope value is above 0.0444 [hPa/s], and walking on level ground otherwise. When walking on stairs, height changes incrementally. We use Eq. 6 below to compute height h [m] from air pressure p [hPa].

$$h = 153.8 \times (t0 + 273.2) \times \left(1 - \left(\frac{p}{p0}\right)^{0.1902}\right) \qquad (6)$$

Here, p0 [hPa] and t0 [°C] are atmospheric pressure and ambient temperature, respectively, at h = 0 [m]. We set p0 = 1013.25 [hPa] and t0 = 15 [°C] in this formula to calculate height. We determine the distance ascended or descended by subtracting the value of height calculated in real time from the value of height when beginning to ascend or descend the stairs.

Height is set to a fixed value once it has been determined that the user is walking on the same floor. In short, if height is set to a value of 賊 0 [m] at start time, it can then be set to a value one floor's worth greater after ascending stairs to the next floor. Likewise, if it is estimated that the user has returned to the start-time floor, height will be reset to a value of 0 [m]. In judging the current floor, it is determined that the user has moved between floors if the difference in air-pressure values before and after judging stair ascending/descending is above a certain threshold.

The height of one floor is set to constant H. Here, a low-pass filter is applied to the air-pressure values obtained from the user's smartphone to determine an average value. The value of air pressure when starting the walk is treated as a reference value and the difference between air-pressure values calculated in real time and that reference value is determined. However, as the value of air pressure when starting the walk is essentially unstable, the reference value is updated to the average value of air pressure obtained on the tenth time the process for determining the average value is performed after starting to walk. The difference between two air-pressure values is an absolute value. In this study, we set the threshold for judging floor movement by air-pressure difference to 0.51 [hPa]. If difference in air pressure exceeds this threshold, the reference value is updated to the current average value of air pressure.

When judging that floor movement has taken place, the current height is determined and the height value is fixed to that height. If the reference value is higher than the current air-pressure value, the floor number is considered to have increased. Furthermore, if the current height should satisfy $H - 0.9 < h < H + 0.9$, height is updated to H.

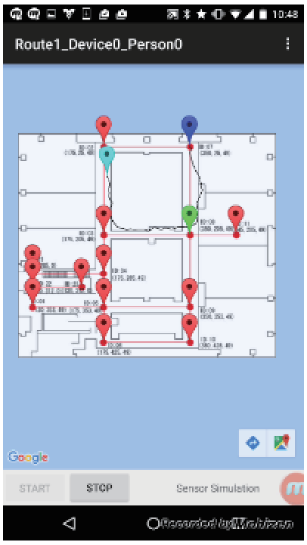


Figure 3: Android application for PDR Challenge.

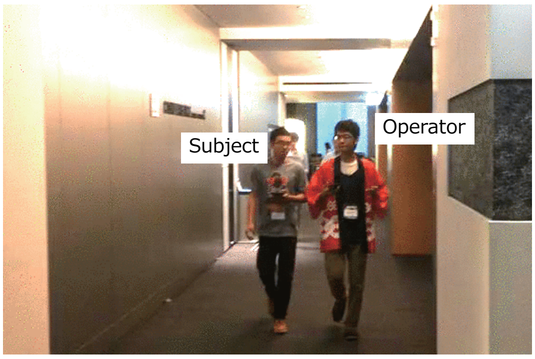


Figure 4: A scene of data collection.

4 UBICOMP/ISWC 2015 PDR CHALLENGE

PDR Challenge was a contest in PDR accuracy that took place in 2015 as part of UbiComp/ISWC held at Grand Front Osaka in Japan [5]. In this contest, subjects walked on five prearranged routes and collected data by carrying a smartphone equipped with several types of sensors. Nearly 100 subjects came together and about 300 sets of walking data were collected.

An Android app designed for this PDR Challenge was used to collect data. This app displays a floor map on the smartphone's screen and plots the subject's trajectory in real time as shown in Figure 3. Purple balloon is start/goal point. Initial direction is purple balloon to green balloon. The other nodes are represented as red balloon. Aqua balloon is current position. The subject proceeds to walk while observing that trajectory. Here, the management side of the PDR Challenge distributes an "app skeleton" in a state that enables a certain amount of PDR in this app to be performed. The participant side adds a program on top of this skeleton and/or makes revisions to incorporate the participant's own algorithm for collecting walking data.

Table 1: Result of PDR Algorithm Category.

	Error Avg. [m]	Error SD.
Team Freshers	12.96	9.21
TUT USL	13.06	7.78
No PDR, No Future.	3.49	1.69
Kohei Kanagu	10.67	6.53
Team UCLAB (unofficial)	46.93	9.37

An operator walks behind the subject at the time of data collection. The subject is briefed on the route to be taken once before the walk beings, but the subject is also given instructions from the operator while walking to move straight ahead, turn right or left, ascend/descend stairs, etc. (Figure 4). The smartphone held by the subject is connected to a button switch by a cord. The operator walking behind the subject holds this button switch and pushes it when giving the subject certain route-related instructions, such as walk, turn right or left, or stop. Pushing the button in this way adds relevant information to sensor time data such as when walking began, where a right or left turn was performed, or when walking ended. This information is used in calculating accuracy.

Final results in the PDR Algorithm Category are shown in Figure 1. Our team, called ”Team Freshers,” is listed at the top of the table shown. With our algorithm, we achieved an average error of 12.96 [m] and standard deviation of 9.21 [m]. These results came in third out of the five teams that competed in the PDR Challenge. Looking only at the results for standard deviation, our value was large coming in fourth among the participating teams. This result reflected a mix of walking data exhibiting relatively high accuracy and walking data exhibiting low accuracy. We consider two factors at play here: the subject’s trajectory was estimated in a linear manner and errors occurred in estimating height.

Among the above factors, the one that helped improve accuracy was the former one, that is, an algorithm that assumes linear walking and orthogonal corners. Examples of location estimation accuracy on a plane are shown in Figure 5. The left and right sides of the figure shows results for traditional PDR and proposed PDR, respectively; the blue plots are the correct trajectories and the red plots are the estimated trajectories. Trajectories obtained by traditional PDR turn out to be curved trajectories owing to the effects of drift in the gyrometer. In contrast, the proposed PDR method generates trajectories close to the correct trajectories as a result of applying the knowledge that corridors are linear and corners are orthogonal [4].

At the same time, there were several failures in estimating right and left turns, which contributed to a drop in accuracy (Figure 6). An error in estimating a right or left turn even at only one location can significantly displace the rest of the estimated trajectory from the correct trajectory and degrade accuracy greatly. There is therefore a need for some means of preventing such mishaps in estimating right and left turns.

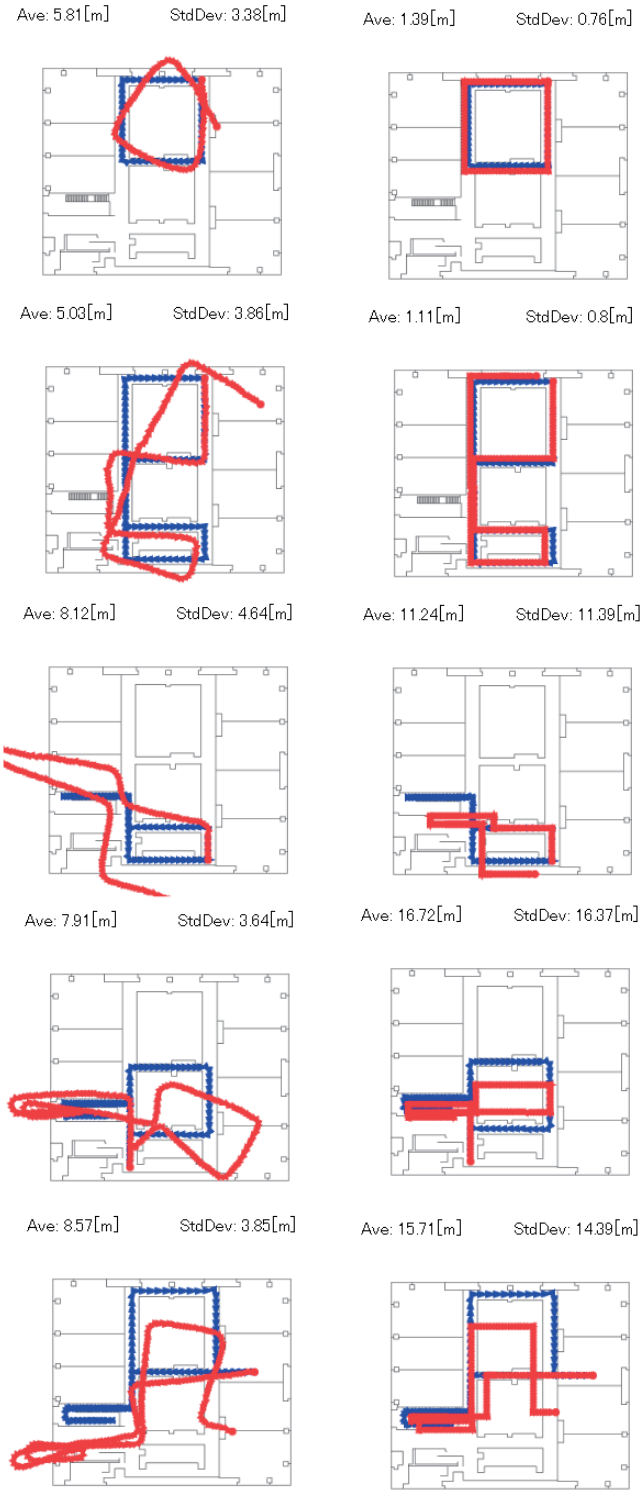


Figure 5: Examples of horizontal positioning (left: traditional PDR, right: proposed PDR)

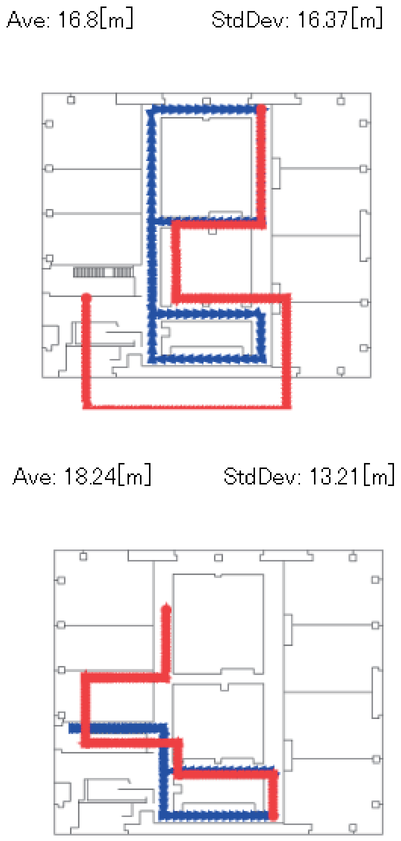


Figure 6: Error of direction estimation

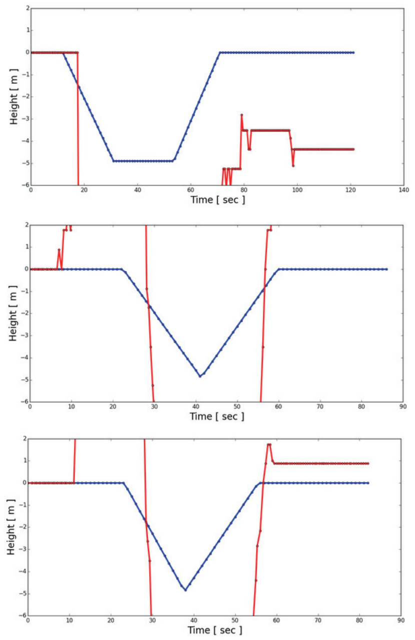


Figure 7: Examples of height estimation.

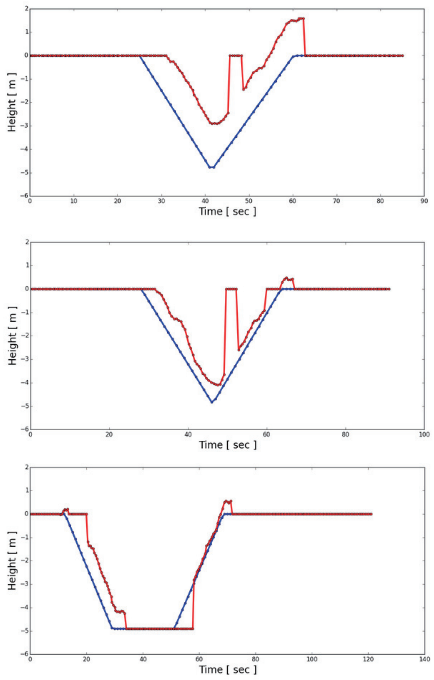


Figure 8: Improvement of height estimation.

Examples of making estimations in the height direction are shown in Figure 7. Unfortunately, the implemented algorithm contained errors preventing correct estimations from being made in the height direction. On the whole, errors in the height direction affect a drop in accuracy the most, so a correct implementation of the proposed algorithm became an issue to resolve.

However, we were given an opportunity to have our algorithm evaluated one more time after the PDR Challenge, so we decided to revise and reevaluate it. Original version contains several bugs especially in height estimation. Therefore, we focused mainly on removing program errors, resolving problems in the passing of function arguments, and improving the height-estimation algorithm. At first, correct fixed floor height is introduced. Second, we improve estimation algorithm of stair descending/ascending stairs because the original version cannot estimate the activity stably. As a result, the revised algorithm achieved an average error of 7.93 [m] and standard deviation of 5.72 [m] thereby improving our standing from third to second place. In particular, improvement could be seen in height estimation. As shown in Figure 8, ascending and descending by stairs could be tracked and change in height estimated correctly to some extent. However, these results still included failures in floor estimation since the current floor was sometimes incorrectly estimated to be the first floor. Further improvement in this algorithm is therefore needed.

5 CONCLUSION

In this paper, we proposed a method for estimating the location of a pedestrian in an indoor 3D space using data on acceleration, angular speed, and air pressure obtained from sensors built into a smartphone. A key feature of this method is the introduction of environmental knowledge that can be applied to ordinary indoor walking such as linear walking and turns of 90 °. Using this method, we competed in the PDR Challenge of UbiComp/ISWC 2015 (a leading international forum for ubiquitous computing) and came in third out of five teams with an average error of 12.96 [m] and standard deviation of 9.21 [m]. Subsequent improvements in the PDR algorithm resulted in an average error of 7.93 [m] and standard deviation of 5.72 [m] thereby improving our standing from third to second place. Since linear corridors can be seen in many buildings, we consider that an algorithm that makes the walking trajectory linear as proposed here should be applicable to many buildings.

In future research, we plan to deal with the occurrence of large error associated with the estimation of travel direction and failure in estimating movement between floors by air pressure with the aim of creating an even more accurate PDR algorithm.

REFERENCES

[1] Kamisaka, D., Muramatsu, S., Iwamoto, T., Yokoyama, H. Design and Implementation of Pedestrian Dead Reckoning System on a Mobile Phone. *IEICE transactions on information and systems*, 94(6):1137–1146, 2011.

[2] Kourogi, M., Sakata, N., Okuma, T., Kurata, T. Indoor/Outdoor Pedestrian Navigation with an Embedded GPS/RFID/Self-Contained Sensor System. In *The 16th International Conference on Advances in Artificial Reality and Tele-Existence*, pages 1310–1321, 2006.

[3] Ban, R., Kaji, K., Hiroi, K., Kawaguchi, K. Indoor Positioning Method Integrating Pedestrian Dead Reckoning with Magnetic Field and WiFi Fingerprints. In *In Proceedings of The Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU2015)*, pages 169–174, 2015.

[4] Kaji, K., Kawaguchi, N. Estimating 3D Pedestrian Trajectories using Stability of Sensing Signal. In *The 7th International Conference on Indoor Positioning and Indoor Navigation (IPIN2016)*, 2016 (to appear).

[5] Kaji, K., Kanagu, K., Murao, K., Nishio, N., Urano, K., Iida, H., Kawaguchi, N. Multi-Algorithm On-Site Evaluation System for PDR Challenge. In *in Proceedings of the Ninth International Conference on Mobile Computing and Ubiquitous Networking (ICMU2016)*, 2016 (to appear).

[6] Seidel, S., and Papport T. 914Mhz Path Loss Prediction Model for Indoor Wireless Communications in Multifloored Buildings. In *Proceedings of IEEE Transactions on Antennas and Propagation*, pages 207–217, 1992.

[7] Manandhar, D., Kawaguchi, S., Uchida, M., et al. IMES for Mobile Users Social Implementation and Experiments based on Existing Cellular Phones for Seamless Positioning. In *International Symposium on GPS/GNSS*, 2008.

[8] Kaji, K., Kawaguchi, N. Design and Implementation of WiFi Indoor Localization based on Gaussian Mixture Model and Particle Filter. In *The 3rd International Conference on Indoor Positioning and Indoor Navigation (IPIN2012)*, pages 1–9, 2011.

[9] Bahl, P., and Padmanabhan, V. N. RADAR: An In-Building RF-based User Location and Tracking System. In *Proceedings of IEEE Infocom 2000*, pages 775–784, 2000.

A Proposal of Life Improvement Method Using SNS Data and Life Log Data

Shota Kuwano, Jun Sawamoto
Software and Information Science Department
Iwate Prefectural University, IPU
Takizawa, Iwate, 020-0193 Japan
sawamoto@iwate-pu.ac.jp

Kota Watanabe, Hiroshi Yajima
Department of Information Systems and Multimedia Design
Tokyo Denki University, TDU
5 Senju-asahicho，Adachi-ku，Tokyo 120-8551 Japan

Abstract- Recently, a systems to use life log data for the improvement of daily living life is paid much attention. Considering that the quality of one’s daily living relates to his/her personal feelings very closely, more effective improvement of living life can be expected by taking account of personal feelings in addition to the life log data. The purpose of this research is to combine user's feelings with life log data and provide more effective improvement method of the quality of daily living. The user lives using the system while paying good attention to provided advices and factors which influences user's feelings, and tries to reproduce a daily living model which generates good feelings. The experiment showed that days which produced better feelings increased with advices provided by our system.

Keywords: Life improvement method, Personal feelings, Life log, SNS, Health care.

1 INTRODUCTION

Recently, a system to use life log data for the improvement of daily living life is paid much attention [1] [2]. Considering that the quality of one’s daily living relates to his/her personal feelings very closely, more effective improvement of living life can be expected by taking account of personal feelings in addition to the life log data. Wearable devices are widely used rapidly by technological progress in recent years. Life log data became easy to collect by the wearable devices such as Apple Watch, SONY SmartBand, etc. Acquired life log data is used for the improvement of living life.

Negative feeling of such as spiritlessness is mentioned as the one which damages healthy consciousness [3]. It is pointed out that it is difficult to motivate people in correction of living habit, but a supporting method which relates living habit to human feelings such as spiritlessness is the effective clue to bring up the self-management power toward the healthy behavior.

The human feelings are related to the living life very closely. For example, a student with appropriate eating habit scored low negative feelings such as depression, anxiety and spiritlessness, and the oppression tendency of the feeling which can be often seen under stress conditions was little [3].

The purpose of this research is to combine user's feelings with life log data and provide more effective improvement method of the quality of daily living.

Takeuchi et al. [4] use life log data for health care by mining the data in their system. Their system asks users to input their personal stress level in five stages manually by themselves. But the problem is that the manual input takes up users’ time and the evaluation of stress becomes subjective above all. In our research, we consider to reduce the input time and make the stress evaluation objective by extracting personal feelings out from users’ daily SNS information.

By using two kinds of data, i.e., user's feelings and life log data, daily life model is constructed. Life log data on the day when good feelings are generated is accumulated as a good model, on the contrary, when bad feelings are generated as a bad model.

Advices to improve daily living habit are given based on the models accumulated. When daily life log data is approaching to a bad model, a warning notice is issued and an advice to bring it back to a better model is provided. The system looks for the tendency of the model and tries to find factors which influence the user's feelings. The user lives using the system while paying good attention to provided advices and factors which influence user's feelings, and tries to reproduce a model which generates good feelings.

By the experiment, life log data and data of personal feelings extracted from SNS have been collected for 2 months for one subject who is a University student. Then, data was collected for one more month while providing life improvement suggestions to the subject. We use standard score (deviation) as the reference value. On a day beyond score value 55 in feelings value, the model is categorized as a good model. When the score value is lower than 45, the model is categorized as a bad model.

This experiment showed a feature that the feelings on the day when the subject went to the university is tend to be a good model. Also considering the fact that a short stay time at home or a short work time at the part time job induced better feelings. It is highly possible that the chance to come in contact with other students is the factor to cause good feelings. After that the subject tried to go to a university as much as possible intentionally and the standard score improved. As an evidence, 5 cases out of 16 good models concentrated in the last 2 weeks of the experiment. The experiment was a short one, but we can confirm that days which produced better feelings increased with advices by our system.

But time of the good feelings may not always be good for health. So good feelings are generated by the proposed method, but there is a possibility of unhealthy living conduct. It is left for the user's choice to compromise good feelings and good healthy living. That is one of the problems of the current system.

The rest of this paper is organized as follows: In section 2, we describe about improvement of living with consideration of feelings. In section 3, we present how to acquire personal feelings in detail. In section 4, details of the acquisition of life log data is described. In section 5, we describe model and model construction. In section 6, we describe how to generate life improvement advices using model data base. In section 7, conducted experiment and its result are described. Finally, the paper is concluded in section 8.

2 IMPROVEMENT OF LIVING HABIT WITH CONSIDERATION OF FEELINGS

A user's daily life log data and feelings are acquired. Life log data on the day when good feelings were produced is saved as a good model and life log data on the day when bad feelings were produced is preserved as a bad model as shown in Figure 1.

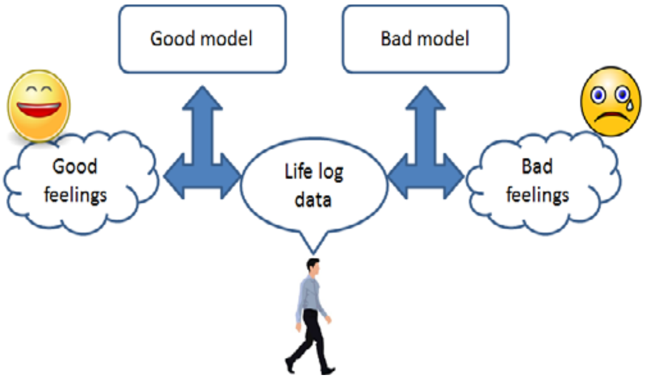


Figure 1: Mechanism of a model construction

Advices to improve user's living habit are provided based on the model accumulated as shown in Figure 2. When user is approaching to "bad model", daily life log data is considered cautionary and the system sends advices to bring it back close to a better model. And it's the purpose of this research, to change the daily lifestyle to a better one.

3 AQUISITION OF PERSONAL FEELINGS

To deal with the user's feelings as data by this research, it's necessary to acquire feelings. After explaining the system configuration for the acquisition of personal feelings, the procedure to extract personal feeling from SNS is explained in detail. A design of the data base with which we deal the feelings acquired in the system is described.

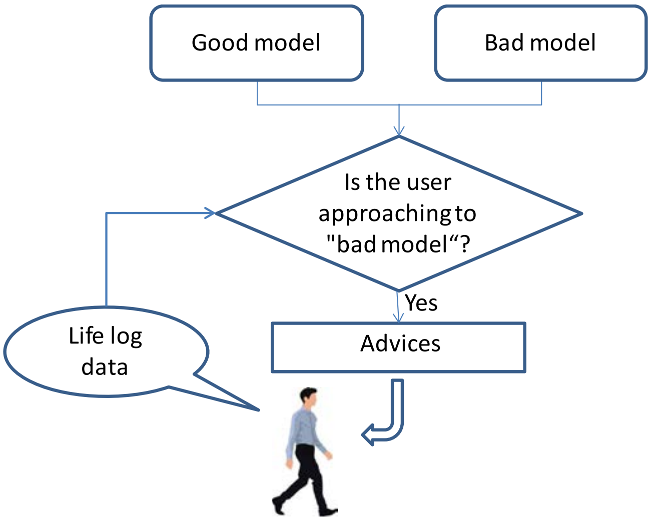


Figure 2: Approach to change the daily life to a better life

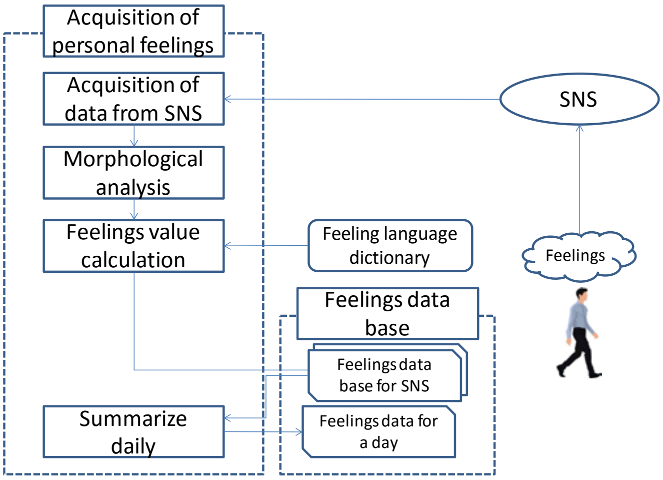

Acquisition of personal feelings
Acquisition of data from SNS
SNS
Morphological analysis
Feelings value calculation
Feeling language dictionary
Feelings
Feelings data base
Feelings data base for SNS
Summarize daily
Feelings data for a day

Figure 3: System configuration of acquisition of personal feelings

As shown in Figure 3, the user contributes the contents about the feelings which has formed from the daily life and events to SNS. Contributed information is acquired by the contribution information acquisition method in the feelings acquisition part and a feelings value of the contribution information is calculated with the procedure described in 3.1. After calculating the feelings value of the contributed information, it is preserved and accumulated. To have to calculate the feelings value of the day at the end of the day, the system refers to the feelings values of the contributed information which are accumulated up to now and calculate the mean value every day. This is preserved as the mean feelings value of the day and accumulated in the feelings data base again.

3.1. Acquisition procedure of feelings

An elementary feeling acquisition and analysis method from Twitter is used. We acquire data of the contribution and its date and time from Twitter.

After morphological analysis of the body of acquired contribution messages, we find the feeling value for every morpheme. It is checked with the feeling language dictionary [5][6] for every morpheme. There are various views about the kind of feelings. In this research, two kinds feelings of positive and negative is used and a feeling language dictionary which consists of only two kinds is used and positive is set to 1 and negative -1.

With the feelings value for a morpheme, the feeling value of the contribution information is calculated as a mean value with the following equation. Here x_i is the feelings value for each morpheme 1 or -1 and n the number of morphemes of the contribution information which were registered by a feeling language dictionary.

$$\text{Feeling value of the contribution information} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The system is able to calculate the feelings value of contribution information as the value between 1 to -1. We preserve the feeling value and contribution message as the feelings value of the contribution information together. The feelings value calculated to each case of contribution information, a mean is calculated once a day. It is processed for the previous day's contribution information when the date changes.

3.2. Feelings data base

The feelings data base holds the feelings value of the contribution information of Twitter and the feelings value for every day. The data base records are as follows;

- The feelings value of the contribution information
  Contribution date and time
  Feelings value of the contribution information
- The feelings value of a day
  Date
  Feelings value of a day

4 AQUISITION OF LIFE LOG DATA

After showing the system configuration in the life log data acquisition part, The contents of acquired data and the data base we deal with are described.

System configuration in the life log data acquisition part is shown in Figure 4.

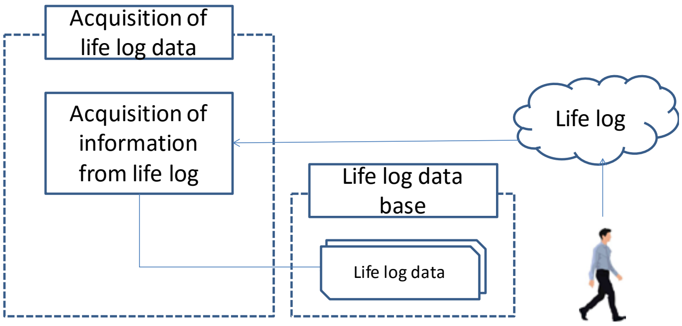

Acquisition of life log data
Acquisition of information from life log
Life log data base
Life log data
Life log

Figure 4: System configuration of life log data acquisition part

4.1. Acquisition methods of life log data

Life log data acquisition part employs SmartBand and Lifelog API which Sony Mobile Communications offers [7]. A smart phone with lifelog application is to be used to log the trends of usage of the smart phone and the position information by the GPS. By the SmartBand, sleeping hours and the depth of the sleep are measured, and the number of steps and consumption calorie can be recorded in detail by the acceleration sensor.

Smart phone is connected with SmartBand by Bluetooth. The information acquired from SmartBand is preserved by a Lifelog server through smart phone. Preserved data can be accessed by Lifelog API on the smart phone.

4.2. Life log data base

Following items are recorded in the life log data base.

- Smart phone usage time
  The general usage hours of the smart phone in a day is measured.
- The genre-sorted smart phone application usage time
  Time is measured about the genre-sorted smart phone applications with the smart phone. Roughly sorted by SNS, a game application and a browser application, etc. and measured application usage time according to the genre.
- Sleep (sleep time and the depth of sleep)
  A record about sleep is acquired. Acquired items are as follows.
  Total sleeping hours
  Time of shallow sleep
  Time of deep sleep
- Position information
  In this research, we measured the sojourn time at the pre-designated points rather than using latitude and longitude position information. Pre-designated points are as follows because we conducted experiment for a subject who is a University student. We estimated the stay at these points by using acquired latitude and longitude value.
  Home stay time
  University sojourn time
  Part-time work sojourn time

-Activity data
  It is acquired as the most basic data of life log. The acquired items are as follows.
  Number of walk steps
  Consumption calorie (calorie consumed by exercise)

5 MODEL CONSTRUCTION

Model construction part is explained here. After explaining about the model, the system configuration in the model construction part is shown.

5.1. Model and model construction

A model is life log data on the day when good feelings or bad feelings are produced. It's preserved with the feelings value of the day.

The system configuration of the model construction part is shown in Figure 5. A model is constructed using and processing data acquired by the respective acquisition part in the model construction part. If feelings value of a day exceeds the threshold then recorded as a good model or is lower than the threshold then recorded as a bad model.

The deviation value is used for the model construction standard. We refer to the feelings value of all days from the feelings data base and calculate the deviation value of the feelings value of the latest day in the whole data base. It's confirmed by the model examination method whether the calculated deviation value satisfies the model construction standard beforehand. When satisfied, we refer to life log data on the day from the life log data base and add it to the model data base.

The model construction part is processed every day. At the end of the day, when feelings acquisition finishes its daily calculation, calculation of the deviation value and model examination are performed.

5.2. Model construction criteria

We first prepare two criteria by which a good model and a bad model are constructed. The deviation value is used for the respective criterion this time.

The reason why we have decided to use the deviation value is because we want to make the model construction criteria somewhat dynamic. When the degree of life is improved by the system, the mean of the feelings value rises. And the chances of producing good models would increase. But at certain degree the living improvement may stop with fixed criteria. In this research we always aim at improving to a better life by use of the system. Therefore, we thought it is always possible to put up a higher target by using the deviation value. When the living degree has gone down, it is the same. In that case, by setting a low target which fits the user as a waypoint, then the good effect can be expected.

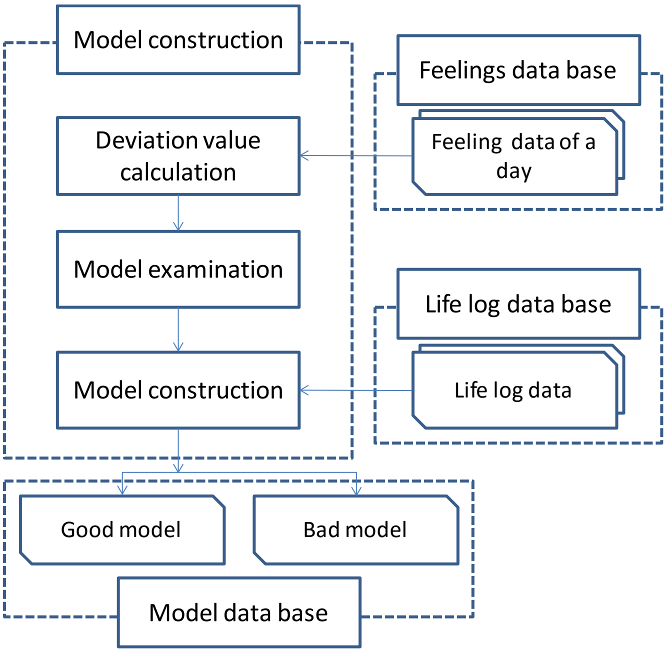


Figure 5: System configuration of model construction part

6 ADVICE GENERATION FOR LIFE IMPROVEMENT

It's explained about the improvement suggestion (advice) part. The system configuration of the improvement suggestion part is shown in Figure 6.

Current life log data is compared with the model in the improvement suggestion part. Models are referred from the model data base and a model chart and an improvement advices are generated. Improvement suggestions (advices) with a model chart are shown to the user.

6.1. Creation of improvement proposals

The following improvement suggestion is made for every item of the living data at the improvement suggestion part.

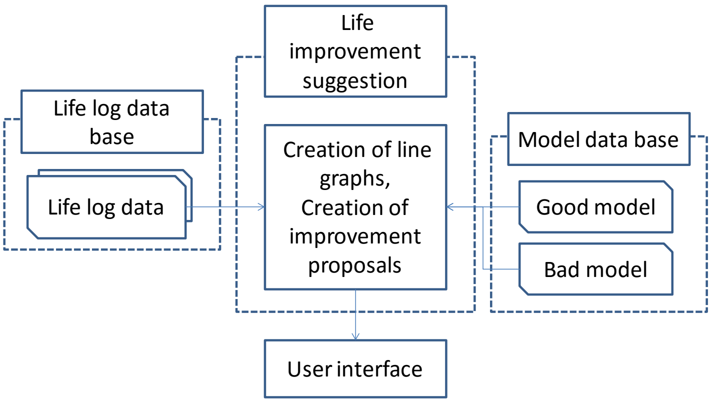

Life improvement suggestion
Life log data base
Life log data
Creation of line graphs, Creation of improvement proposals
Model data base
Good model
Bad model
User interface

Figure 6: System configuration of improvement suggestion part

- Ascertain the reliability of the model
  All the acquired life log data may not be the factor which largely influences the user's feelings. So factors in the good model and bad model are compared and find the degree of difference between them. When the difference is distinctive, there is a possibility that the factor is the item which tends to have an influence on personal feelings, so the advice is mainly generated so that an effort may be made to bring the factor close to that of a good model. On the other hand, it can be considered that the credibility is low as the item which influences feelings when the difference is small.

- Confirm whether it's the raised item or lowered item
  Acquired life log data, more than one, it's different depending on the users whether a higher value tend to induce a good model or a lower value tend to induce a good model. So we refer to the existing models and examine which model has higher value, a good model or a bad model. When the one of the values of the good model is high, it'll be important to raise this value in daily life. When the value of the good model is low, it's necessary to be aware that this value should be lowered.

- See the recent living tendency
  Recent life log data is compared with the model to see whether the recent life is approaching to a good model or a bad model. When approaching a bad model, caution is needed.
- Confirm whether there is a day when you deviated from a good model in recent life
  The user is trying to live to achieve a good model but it's abnormal to deviate from the value of the good model. So it should be confirmed whether there is a day when the value deviates from the model. When noticed that, there is an interface which can read life log data in the nearest seven days, so we recommend the user to look at the life again and look for the cause for the large deviation.

6.2. Design of user interface

The main function of the user interface is to show life trend in terms of life log data in the nearest seven days, to confirm generated models and to receive life improvement advices. The system refers data for the nearest seven days from the life log data base and outputs them. The improvement advices created at the improvement suggestion part are output. Data of the model data base created at the model construction part is referred and outputs by a line graph for the model confirmation. Output data are more detailed below.

- Life trend in the nearest seven days
  Life log data in the nearest seven days are output. It is to see what kind of life the user is living through these data once again, and have the user look back his life trend.

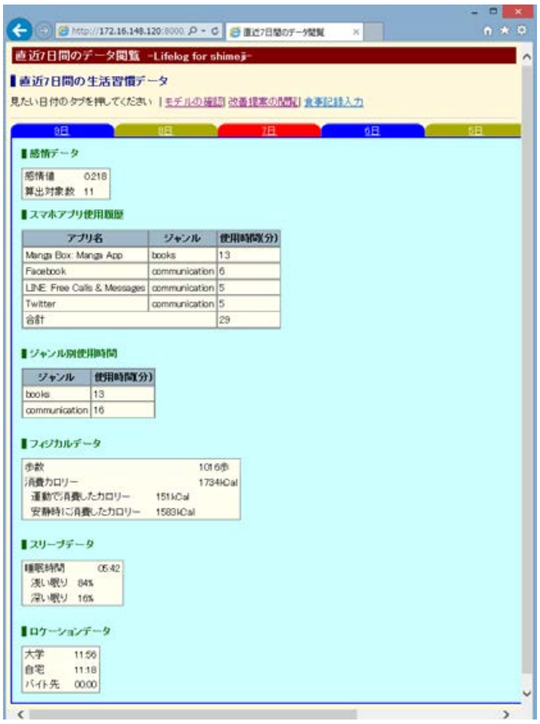
- Confirmation of generated models
  Respective averages of the present good model and bad model and the value of one week life log data of the nearest seven days are indicated by line graphs. In days which invents good feelings or bad feelings the user can confirm which data item is in what degree visually from these graphs. Life log data in the nearest seven days are shown together and comparing the present life with models, it is possible to recognize current status of life.

- Life improvement advices
  The improvement advices created at the improvement suggestion part is presented verbally.

Interface looks as follows.

- Life log data in the nearest seven days

As shown in Figure 7, life log data in the nearest seven days can be accessed here. For the upper part, tabs of a week are shown, and when a tab is clicked, life log data on that day is listed. It is also possible to check sleep information and position information by scrolling down to the bottom.



直近7日間のデータ閲覧 -Lifelog for shimeji-

直近7日間の生活習慣データ

見たい日付のタブを押してください | モデルの確認 | 改善提案の閲覧 | 食事記録入力

感情データ

感情値 0218
算出対象数 11

スマホアプリ使用履歴

アプリ名	ジャンル	使用時間(分)
Manga Box: Manga App	books	13
Facebook	communication	6
LINE: Free Calls & Messages	communication	5
Twitter	communication	5
合計		29

ジャンル別使用時間

ジャンル	使用時間(分)
books	13
communication	16

フィジカルデータ

歩数		1016歩
消費カロリー		1734kCal
運動で消費したカロリー	151kCal	
安静時に消費したカロリー	1583kCal	

スリープデータ

睡眠時間		05:42
浅い眠り	84%	
深い眠り	16%	

ロケーションデータ

大学	11:56
自宅	11:18
バイト先	00:00

Figure 7: Showing Life log data in the latest seven days

- Confirmation of generated models and life improvement advices

As shown in Figure 8, model confirmation and life improvement advices are accessed here. Tabs are used to select data items and to output them. The system aims to have the user check the models and current state of life visually through charts, and has this as the trigger to look back his life again.

The improvement messages created at the evaluation part is shown on the right space of the charts. And English translations are shown on the right hand side boxes in the figure. In this implementation, some of the simple message templates are prepared and output messages are constructed by combining variable parts with the prepared templates. The user is asked to make an effort for improvement of his living habit checking the evaluation and data of the charts.

7 EXPERIMENTS

In this chapter, the experiment to evaluate the usefulness of the system is described. This experiment was conducted for one subject who is a University student.

7.1. Experiment method

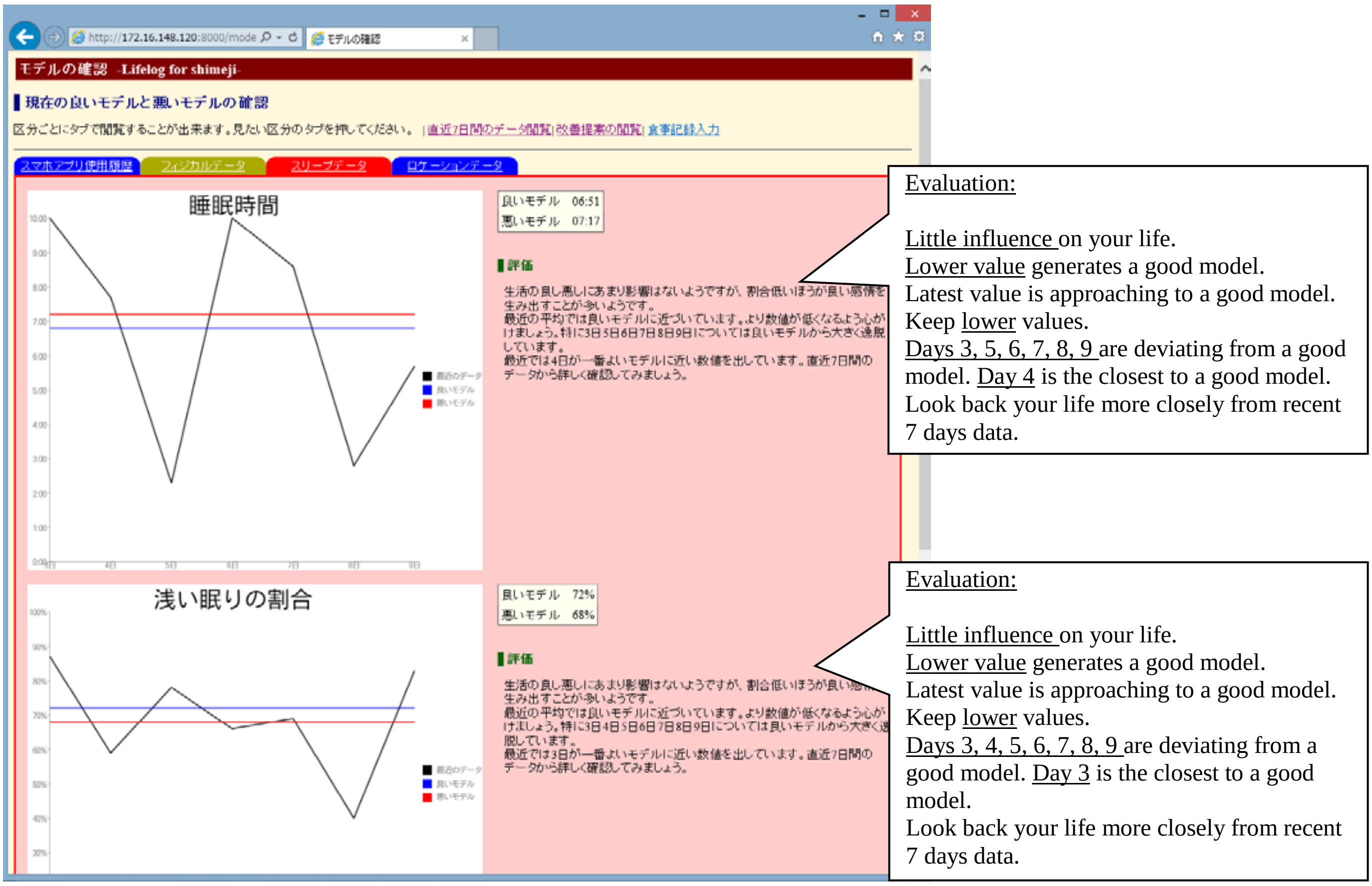
Feelings extraction is performed from Twitter in this experiment, so the subject has to have Twitter account and has certain familiarity to the usage of Twitter. For the experiment, life log data and contribution information to SNS are acquired, so we get permission for the usage of private data in terms of the privacy protection. Lifelog is installed in smart phone of the subject, and the subject puts on Smartband.

The subject is only permitted to remove Smartband during bathing, and he is supposed to wear it even during sleeping. Charge for Smartband is only allowed during bathing when it is taken off and charge it other time is basically not allowed because it isn't possible to acquire life log data during charging it.

As usual, the subject uses Twitter. We did not set any restriction on the lowest or maximum number of tweets because there was a possibility to obstruct usual usage.

The experiment is made for two months first and collection of life log data and construction of models are executed to certain degree. For the model construction criteria, a good model corresponds to deviation value more than 55 and a bad model to deviation value less than 45. From the third month, using the interface of the system, the user receives life improvement advices and looks back his life. The experiment is finished when three months have passed.

In this experiment, the system is evaluated by the frequency of the created models. If the number of good models created during the last one month in the three months of the whole experiment duration exceeds 33% of all good models which are created throughout the experiment, we judge that the system is useful to change user's life to create better feelings in his daily life. On the other hand, if the number of bad models created during the last one month falls below 33% of whole bad models, we judge that the system tends to prevent to create bad feelings in his daily life. The system is evaluated from two points of view.



モデルの確認 -Lifelog for shimeji-

現在の良いモデルと悪いモデルの確認

区分ごとにタブで閲覧することが出来ます。見たい区分のタブを押してください。

良いモデル	06:51
悪いモデル	07:17

評価

生活の良し悪しにあまり影響はないようですが、割合低いほうが良い感情を生み出すことが多いようです。
最近の平均では良いモデルに近づいています。より数値が低くなるよう心がけましょう。特に3日5日6日7日8日9日については良いモデルから大きく逸脱しています。
最近では4日が一番よいモデルに近い数値を出しています。直近7日間のデータから詳しく確認してみましょう。

良いモデル	72%
悪いモデル	68%

評価

生活の良し悪しにあまり影響はないようですが、割合低いほうが良い感
生み出すことが多いようです。
最近の平均では良いモデルに近づいています。より数値が低くなるよう心がけましょう。特に3日4日5日6日7日8日9日については良いモデルから大きく逸脱しています。
最近では3日が一番よいモデルに近い数値を出しています。直近7日間のデータから詳しく確認してみましょう。

Evaluation:

Little influence on your life.
Lower value generates a good model.
Latest value is approaching to a good model.
Keep lower values.
Days 3, 5, 6, 7, 8, 9 are deviating from a good model. Day 4 is the closest to a good model.
Look back your life more closely from recent 7 days data.

Evaluation:

Little influence on your life.
Lower value generates a good model.
Latest value is approaching to a good model.
Keep lower values.
Days 3, 4, 5, 6, 7, 8, 9 are deviating from a good model. Day 3 is the closest to a good model.
Look back your life more closely from recent 7 days data.

Figure 8: Confirmation of generated models and life improvement advices

7.2. Experimental results

The experimental result is shown in Table 1.

Table 1: Experimental result

Month	No. of good models	No. of bad models
First, second months	9	6
Third month	7	3
Total	16	9
% of third month	43.8％	33.3％

For the evaluation of good model, result value 43.8% exceeds 33% and bad model evaluation value 33.3% is over 33%. The system is effective for the improvement of creation of good feelings, but the effect which evades holding bad feelings can't be expected.

We came up with a hypothesis that the good feelings could be easily created when the subject goes to the University. An actual record is as shown in Table 2.

Table 2: Created model related to sojourn time at University

Created models	Average sojourn time at University
Good models	4 hours 21 minutes
Bad models	53 minutes

When the position information is checked, it shows the same trend as indicated in Table 3.

Table 3: Created model related to sojourn time at Home and Part-time work

Created models	Hours at home	Hours at Part-time work
Good models	09:20	01:42
Bad models	11:40	05:02

It is estimated from this result that the chance to interact with others has a high possibility to create good feelings in case of this subject. Therefore, after two months the system advised the subject to go to school as much as possible and that caused this experimental result.

8 CONCLUSION

In this experiment, the number of subject was only one, and the session length was three months and it is a very short period. Sample data is also small, and the reliability of the evaluation is still low. Therefore it's necessary to assume various patterns and conduct an additional long term experiment.

In the limited experiment, it is seen that the proposed system is working for the improvement of the creation of good feelings, but the effect which evades holding of bad feelings can't be expected. In this research, feelings was combined with life log data, and a system was designed, implemented and evaluated its effectiveness for the improvement of daily life to some extent. It was a short period of an experiment, but it is confirmed that days when better feelings are produced increased.

Days of good feelings may not always be good for health. So even good feelings are produced by the proposed method, there is a possibility which induces unhealthy life. It is left for the user's choice to compromise good feelings and good healthy living. That is one of the problems of the current system. It is necessary to consider whether factors which produce good feelings would not influence unhealthily to the user and the proposal should be made.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 16K00429.

REFERENCES

[1] Liao, J., Wang, Z., Wan, L., Cao, Q. C., & Qi, H. (2015). Smart Diary: A Smartphone-Based Framework for Sensing, Inferring, and Logging Users' Daily Life. Sensors Journal, IEEE, 15(5), 2761-2773.

[2] Kasuya, S., Zhou, X., Nishimura, S., & Jin, Q. (2016). A framework of personal data analytics for well-being oriented life support. In Advanced Multimedia and Ubiquitous Engineering (pp. 443-449). Springer Berlin Heidelberg.

[3] K. Takahashi (2009), A Study of Psychological Factors Related to Health Attitudes and Lifestyles of Japanese Undergraduate Students from View Points of Emotional Responses, Coping Behaviors and Health Locus of Control, " Hirosaki University Health Management Summary. 30, 2009, p.14-21.

[4] H. Takeuchi, N. Kodama, T. Hashiguchi, and N Mitsui, (2005) " Healthcare data mining based on a personal dynamic healthcare system," Proc. 2nd Int. Conf. on Computational Intelligence in Medicine and Healthcare, pp.37-43.

[5] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi. Collecting Evaluative Expressions for Opinion Extraction, Journal of Natural Language Processing 12(3), 203-222, 2005.

[6] Masahiko Higashiyama, Kentaro Inui, Yuji Matsumoto. Learning Sentiment of Nouns from Selectional Preferences of Verbs and Adjectives, Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, pp.584-587, 2008.

[7] Lifelog API | Sony Develofper World, https://developer.sony.com/develop/services/lifelog-api/

Session 7:
Data Models

(Chair: Tomoki Yoshihisa)

Proving Anonymity for Timed Systems

Yoshinobu Kawabe[†], and Nobuhiro Ito[†]

[†]Department of Information Science, Aichi Institute ofTechnology, Japan
{ kawabe, n-ito }@kwb.aitech.ac.jp

Abstract - In this study we discuss a proof technique for a timed version of trace anonymity. Even though communication patterns are indistinguishable, the sender's identity might be disclosed by detecting the timing of message emission. To deal with such timing features, this study employs timer variables which range over non-negative real numbers. When verifying the timed anonymity of a security protocol, in this study we employ a theorem-proving tool. In this paper, we also have a discussion on an adversary model to handle stronger adversaries. This paper does not introduce formal definitions, but we believe that the formalization is possible in analogy with the adversary model.

Keywords: Timed systems, Anonymity, Verification, Formal Method, I/O-automaton

1 INTRODUCTION

We say a security protocol is anonymous if an adversary who can observe all the occurrences of events from the protocol cannot determine who is the "actor" of the events. There are many studies to describe and verify the anonymity of security protocols formally; for example, in [3] a proof technique that incorporates theorem-proving is introduced.

In this paper we discuss anonymity of timed systems. There have been many studies based on formal methods that modeled and verified the correctness of timed systems [1][2]. To establish anonymity, we should deal with patterns of communication. However, even though all the communication patterns are indistinguishable, detecting a timing of message emission might break anonymity. That is, the detection of timing information leads to the disclosure of who is an actor.

This study models a timed system with a formal specification language [4]. The language is based on I/O-automaton theory [5][6]. We must handle infinite systems by introducing timer-related variables; however, I/O-automaton theory provides a proof technique called a simulation-based method that can handle infinite-state systems directly. We discuss how the simulation-based method can be applied for proving the anonymity of timed systems.

2 PRELIMINARIES

We explain how to formalize (untimed) anonymity. We assume that readers are familiar with the basic notions and notations for I/O-automaton theory and an I/O-automaton-based formal specification language.

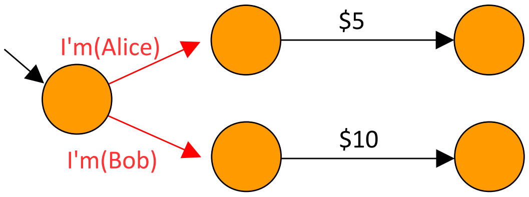


Figure 1: `D1`

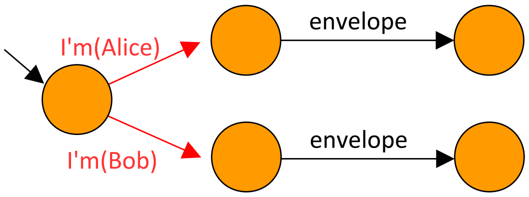


Figure 2: `D2`

2.1 Basic Notion of (Untimed) Anonymity

We explain the basic notion of anonymity with the following example.

Example 1 (Donating anonymously) *There are two people, Alice and Bob, and we assume that only one of them has made an anonymous donation. Alice was going to contribute $5, while Bob was going to contribute $10.*

I/O-automaton `D1` in Fig. 1 describes the above situation. Actions `$5` and `$10` of `D1` are external actions to represent a donation. After the occurrence of `I'm(Alice)` or `I'm(Bob)`, either `$5` or `$10` occurs. Here, `I'm(Alice)` and `I'm(Bob)` are actions that specify the donor. For convenience, we call `I'm(Alice)` and `I'm(Bob)` *actor actions*. We can see that `D1` is anonymous if an adversary who observed all the occurrences of the non-actor actions cannot determine which actor action of `D1` will occur.

If an adversary observed that $5 was posted, then the adversary can deduce that Alice made a donation, since action `I'm(Alice)` can occur in `D1` only when action `$5` occurs. That is, `D1` is not anonymous. One reason for `D1` not being anonymous is that an adversary can know how much money was posted. So, we assume that a donation was posted in an envelope. Suppose f is an operation to replace external actions `$5` and `$10` of `D1` with a fresh external action `envelope`, and we define `D2` as f(`D1`) (see Fig. 2). This operation hides information on how much money was posted. With `D2`, an adversary who can detect the occurrence of event `envelope` cannot deduce which actor action is actually possible. Hence, `D2` is anonymous.

Below, we formally discuss the correctness of communication patterns with regard to anonymity. Let X be an I/O-automaton and A be a family with the following conditions: (i) $\bigcup_{A' \in A} A' \subset ext(X)$; (ii) A' and A'' are disjoint for any distinct $A', A'' \in A$. We call A a *family of X's actor actions*, and an element of $\bigcup_{A' \in A} A'$ is called an *actor action* (on A). The occurrences of different actor actions should be indistinguishable to an adversary. That is, if an eavesdropper cannot distinguish the trace set of system X and that of X's "anonymized" version, then we can see that X is anonymous. This is formalized as follows.

Definition 1 *Let X be an I/O-automaton and A be a family of X's actor actions. We define I/O-automaton $anonym_A(X)$ as follows:*

$$\begin{aligned}
&states(anonym_A(X)) = states(X),\\
&start(anonym_A(X)) = start(X),\\
&ext(anonym_A(X)) = ext(X),\\
&int(anonym_A(X)) = int(X) \text{ and}\\
&trans(anonym_A(X))\\
&\quad = \{(s_1, a, s_2) \mid (s_1, a, s_2) \in trans(X) \wedge a \notin \textstyle\bigcup_{A' \in A} A'\}\\
&\qquad \cup \{(s_1, a, s_2) \mid (s_1, a', s_2) \in trans(X)\\
&\qquad\qquad \wedge A' \in A \wedge a' \in A' \wedge a \in A'\}.
\end{aligned}$$

If $traces(anonym_A(X)) = traces(X)$ holds, we say X is trace anonymous on A.

2.2 How to Prove Anonymity

We describe a proof method for trace anonymity [3].

Definition 2 *Assume X is an I/O-automaton and A is a family of X's actor actions. An anonymous simulation as_A of X on A is a binary relation on $states(X)$ that satisfies the following conditions:*

1. *$as_A(s, s)$ holds for any initial state $s \in start(X)$;*
2. *For any states $s_1, s_2, s'_1 \in states(X)$ and action $a \in sig(X)$, $as_A(s_1, s'_1)$ and $s_1 \xrightarrow{a}_X s_2$ implies the following:*

   If $a \in A'$ holds for some $A' \in A$, for all $a' \in A'$ there is a state s'_2 such that $as_A(s_2, s'_2)$ and $s'_1 \overset{a'}{\Longrightarrow}_X s'_2$; otherwise, there is a state s'_2 such that $as_A(s_2, s'_2)$ and $s'_1 \overset{a}{\Longrightarrow}_X s'_2$.

I/O-automaton X is trace anonymous on A if X has some anonymous simulation as_A. It is easy to see

$$traces(anonym_A(X)) \supseteq traces(X)$$

since $anonym_A(X)$ has all the transitions of X. The other inclusion $traces(anonym_A(X)) \subseteq traces(X)$ of traces can be shown since as_A is a forward simulation from automaton $anonym_A(X)$ to automaton X. A forward simulation from I/O-automaton P to I/O-automaton Q is a binary relation $r \subset states(P) \times states(Q)$ with the conditions shown in [5]; if there is a forward simulation, then we have $traces(P) \subseteq traces(Q)$ ([5], Theorem 3.10).

Proposition 1 *Let X be an I/O-automaton. If there is an anonymous simulation as_A of X on A, then*

$$traces(anonym_A(X)) = traces(X)$$

holds. □

3 ANONYMITY FOR TIMED SYSTEMS

This section discusses a formalization of timed anonymity.

3.1 Example

We use the following example.

Example 2 *There are two people, Alice and Bob. Alice has \$50, while Bob has \$10,000. Charlie has requested only one of them to give him \$10. We do not know which person makes a payment, but one of them actually sends \$10.*

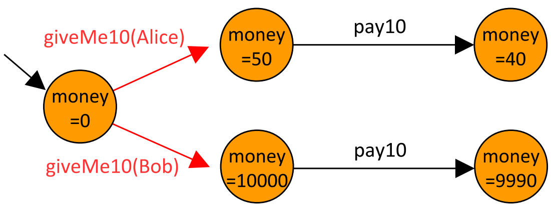


Figure 3: Automaton GMT

Fig. 3's I/O-automaton `GMT` describes this situation. Special actions `giveMe10(Alice)` and `giveMe10(Bob)` represent the actors, and `pay10` is for a payment. The trace set

$$traces(\texttt{GMT}) = \left\{ \begin{array}{l} \texttt{giveMe10(Alice).pay10,} \\ \texttt{giveMe10(Bob).pay10} \end{array} \right\}.$$

is given by automaton `GMT`. We can see that an adversary who observed the occurrence of action `pay10` cannot determine the preceding action. That is, both of `giveMe10(Alice)` and `giveMe10(Bob)` are possible, so the adversary never knows who made a payment. This is the same thing of the second setting in Example 1, so `GMT` is trace anonymous.

In Example 2, Alice possibly pays \$10 even though she has only \$50. In the following, we would like to consider a modified example.

Example 3 *Bob has much money (\$10,000), so he can send \$10 immediately. But Alice has only \$50. When asked by Charlie, she thinks for a moment before sending \$10.*

Fig. 4 shows this situation. Bob can make a decision immediately, but Alice might take some time up to 100 seconds before sending \$10. From this observation, if the payment of \$10 occurs after one second, then the payer is Alice. This means that even though communication patterns are indistinguishable, the sender can be identified by detecting the timing of message emission.

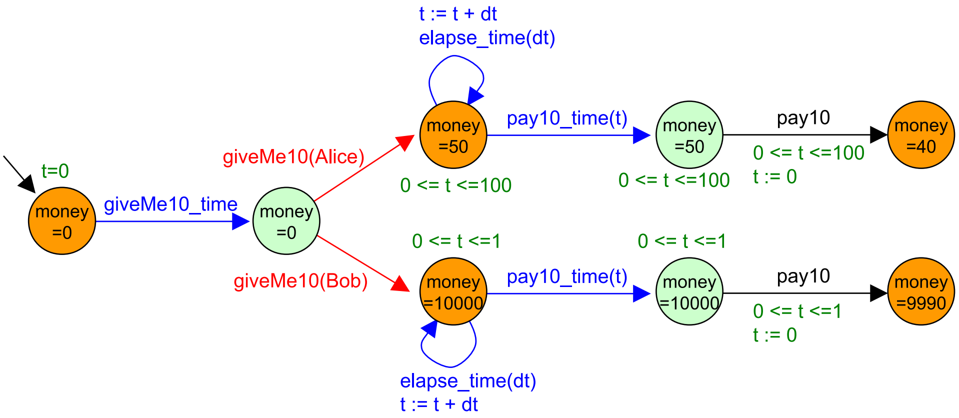


Figure 5: GMTt's Transitions (Expanded in Conventional I/O-Automaton)

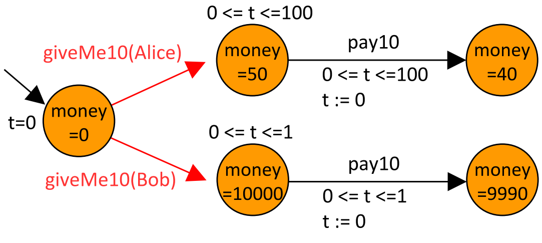


Figure 4: Timed Automaton GMTt

3.2 Modeling Timed System

We model a timed system with IOA [4], which is a formal specification language based on I/O-automaton theory. Automaton GMT in Fig. 3 is written as follows.

```
automaton GMT
  signaure
    output giveMe10(Alice)
    output giveMe10(Bob)
    output pay10

  states
     money: Nat  := 0

  transitions
    output giveMe10(Alice)
      pre money = 0
      eff money := 50;

    output giveMe10(Bob)
      pre money = 0
      eff money := 10000;

    output pay10
      pre (money = 50 \/ money = 10000)
      eff money := money - 10
```

To model a timed system in IOA, this paper uses the following special variables:

- `timer` : a timer variable for elapsing time, and
- `timerFlg` : a flag variable for activating/deactivating the timer.

Also, we use time actions (`giveMe10Time`, `pay10Time(t)` and `elapse(delta)`) for expressing timing constraints and for elapsing time; in the rest of this paper, other actions except actor actions (i.e., `giveMe10(Alice)`, `giveMe10(Bob)` and `pay10`) are called normal actions. We obtain Fig. 4's automaton `GMTt`. We can see that (with Fig. 5) a one-step transition by action `pay10` in Fig. 4 is formalized with a two-step transition sequence with `pay10Time(t)` and `pay10` in IOA language.

```
automaton GMTt
  signaure
    output giveMe10(Alice)
    output giveMe10(Bob)
    output pay10
    output giveMe10Time
    output pay10Time(t)
    output elapse(delta)

   states
     money: Nat   := 0,
     timer: Real   := 0.0,
     timerFlg: Bool   := true

   transitions
     output giveMe10(Alice)
       pre     ~timerFlg
           /\ money = 0
       eff money := 50;
           timerFlg := true

     output giveMe10(Bob)
       pre     ~timerFlg
           /\ money = 0
       eff money := 10000;
           timerFlg := true

     output pay10
       pre     ~timerFlg
           /\ (money = 50 \/ money = 10000)
       eff money := money - 10;
           timerFlg := true

     output giveMe10Time
       pre     timerFlg
           /\ money = 0
       eff timerFlg := false

     output pay10Time(t)
```

```
      pre    timerFlg /\ t = timer
          /\ (money = 50 \/ money = 10000)
          /\ ((money = 50)
              => (0.0 <= timer /\ timer <= 100.0))
          /\ ((money = 10000)
              => (0.0 <= timer /\ timer <= 1.0))
      eff timer := 0.0;
          timerFlg := false

    output elapse(delta)
      pre    timerFlg /\ delta > 0.0
          /\ (money = 50 \/ money = 10000)
          /\ ((money = 50)
               => (   (   0.0 <= timer
                       /\ timer <= 100.0)
                   /\ (   0.0 <= timer + delta
                       /\ timer + delta <= 100.0)))
          /\ ((money = 10000)
               => (   (   0.0 <= timer
                       /\ timer <= 1.0)
                   /\ (   0.0 <= timer + delta
                       /\ timer + delta <= 1.0)))
      eff timer := timer + delta
```

4 ANALYZING ANONYMITY FOR TIMED SYSTEMS

Automaton GMTt is not anonymous because we do not have

$$\begin{aligned}&\texttt{giveMe10Time.giveMe10(Bob).}\\&\quad\texttt{elapse(30).pay10Time(30).pay10}\\&\qquad\qquad\qquad\qquad \in \mathit{traces}(\texttt{GMTt})\end{aligned}$$

but we have

$$\begin{aligned}&\texttt{giveMe10Time.giveMe10(Bob).}\\&\quad\texttt{elapse(30).pay10Time(30).pay10}\\&\quad \in \mathit{traces}(\mathit{anonym}_{\{\{\texttt{Alice,Bob}\}\}}(\texttt{GMTt})).\end{aligned}$$

This means that $\mathit{anonym}_{\{\{\texttt{Alice,Bob}\}\}}(\texttt{GMTt})$'s anonymity does not lead to GMTt's anonymity. If we use:

```
  output pay10Time(t)
    pre    timerFlg /\ t = timer
        /\ (money = 50 \/ money = 10000)
        /\ ((money = 50)
            => (0.0 <= timer /\ timer <= 1.0))
        /\ ((money = 10000)
            => (0.0 <= timer /\ timer <= 1.0))
    eff timer := 0.0;
        timerFlg := false

  output elapse(delta)
    pre    timerFlg /\ delta > 0.0
        /\ (money = 50 \/ money = 10000)
        /\ ((money = 50)
             => (   (   0.0 <= timer
                     /\ timer <= 1.0)
                 /\ (   0.0 <= timer + delta
                     /\ timer + delta <= 1.0)))
        /\ ((money = 10000)
             => (   (   0.0 <= timer
                     /\ timer <= 1.0)
                 /\ (   0.0 <= timer + delta
                     /\ timer + delta <= 1.0)))
    eff timer := timer + delta
```

for defining pay10Time(t) and elapse(delta), we can establish the anonymity. We call the resulting automaton GMTt2.

For GMT, we can find an anonymous simulation easily. The binary relation is:

$$\begin{aligned}\mathit{as}_{\texttt{GMT}}(s,s') \quad\Longleftrightarrow\quad & s.\texttt{money} = s'.\texttt{money}\\ & \vee |s.\texttt{money} - s'.\texttt{money}| = 9950.\end{aligned}$$

For the modified timed automaton GMTt, we can define a binary relation over states:

$$\begin{aligned}\mathit{as}_{\texttt{GMTt2}}(s,s') \quad\Longleftrightarrow\quad & \mathit{as}_{\texttt{GMT}}(s,s')\\ & \wedge s.\texttt{timer} = s'.\texttt{timer}\\ & \wedge (s.\texttt{timerFlg} \Longleftrightarrow s'.\texttt{timerFlg})\end{aligned}$$

which is a candidate binary relation of an anonymous simulation. This formula contains $\mathit{as}_{\texttt{GMT}}$, and we can prove that $\mathit{as}_{\texttt{GMTt2}}$ satisfies some of the conditions to be an anonymous simulation of GMTt2. Below, let $(s,t,p) \in \mathit{start}(\texttt{GMTt2})$ be an initial state of GMTt2, where s is a tuple that represents a state of automaton GMT, t is a value of variable timer, and p is a value of variable timerFlg.

For the initial state condition of Definition 2, it is easy to show $\mathit{as}_{\texttt{GMT}}(s,s)$ implies

$$\mathit{as}_{\texttt{GMTt2}}((s, 0.0, \texttt{true}), (s, 0.0, \texttt{true}))$$

since we have $t = 0.0$ and $u = \texttt{true}$ from GMTt2's definition.

For the step's correspondence condition of Definition 2, we have two more cases. Suppose that a normal action a is enabled in GMTt2. If we have (i) $(s_1,t,p) \xrightarrow{a}_{\texttt{GMTt2}} (s_1',t,p')$ and (ii) $\mathit{as}_{\texttt{GMTt2}}((s_1,t,p),(s_2,u,q))$ then we have (iii) $t = u$, (iv) $p = q = \texttt{false}$ and $p' = \texttt{true}$, and (v) $\mathit{as}_{\texttt{GMT}}(s_1,s_2)$ and $s_1 \xrightarrow{a}_{\texttt{GMT}} s_1'$. Thus, there exists a state s_2' of GMT such that:

- We have $s_2 \overset{a'}{\Longrightarrow}_{\texttt{GMT}} s_2'$ and $\mathit{as}_{\texttt{GMT}}(s_1',s_2')$;
- $a \in \{\texttt{giveMe10(Alice)}, \texttt{giveMe10(Bob)}\}$ implies $a' \in \{\texttt{giveMe10(Alice)}, \texttt{giveMe10(Bob)}\}$; and
- $a = \texttt{pay10}$ implies $a' = a = \texttt{pay10}$.

Therefore, for the state $(s_2', t, \texttt{true})$, we have

$$(s_2,u,q) \equiv (s_2,t,\texttt{false}) \overset{a'}{\Longrightarrow}_{\texttt{GMTt2}} (s_2',t,\texttt{true})$$

and

$$\mathit{as}_{\texttt{GMTt2}}((s_1',t,p'),(s_2',t,\texttt{true})).$$

Consequently, if binary relation $\mathit{as}_{\texttt{GMT}}$ is an anonymous simulation, then binary relation $\mathit{as}_{\texttt{GMTt2}}$ satisfies a step correspondence condition for any normal action.

On the other hand, we suppose that a time action b is enabled at a state. If we have: (i) $(s_1,t,p) \xrightarrow{b}_{\texttt{GMTt2}} (s_1',t',p')$, and (ii) $\mathit{as}_{\texttt{GMTt2}}((s_1,t,p),(s_2,u,q))$ then we have: (iii) $s_1' = s_1$, $t = u$, and $p = q = \texttt{true}$; (iv) If b is elapse(delta) then $p' = \texttt{true}$; otherwise, $p' = \texttt{false}$; and (v) $\mathit{as}_{\texttt{GMT}}(s_1,s_2)$. If we can prove

$$(s_2,u,q) \equiv (s_2,t,\texttt{true}) \overset{b}{\Longrightarrow}_{\texttt{GMTt2}} (s_2,t',p')$$

for state (s_2,t',p'), then $\mathit{as}_{\texttt{GMTt2}}((s_1,t',p'),(s_2,t',p'))$ holds. Hence, $\mathit{as}_{\texttt{GMTt2}}$ satisfies the conditions to be an anonymous simulation of GMTt2 for action b.

5 DISCUSSION: ANONYMITY PROOF WITH STRONGER ADVERSARIES

We have introduced an automaton that has variables `timer` and `timerFlg` in this study. A similar approach is employed in [7] to deal with stronger adversaries.

The technique in [3] can only deal with eavesdroppers, and in [7] an adversary model has been introduced to handle stronger adversaries, which may change the protocol's state in various ways, e.g. by sending dummy messages and by rewriting disk image of a PC. This is formalized as follows.

Definition 3 (Attacker part) *Attacker Atk of system X is quadruplet $(states(X), S_{Atk}, A_{Atk}, T_{Atk})$, where a set of attacker's states S_{Atk}, a set of attacker's actions A_{Atk} and a set of attacker's transitions T_{Atk} should satisfy:*

$$\begin{cases} sig(X) \cap A_{Atk} = \emptyset \text{ and} \\ T_{Atk} \subseteq \{((s_1, v_1), a, (s_2, v_2)) \mid s_1, s_2 \in states(X), \\ \qquad\qquad\qquad\qquad\qquad v_1, v_2 \in S_{Atk}, \\ \qquad\qquad\qquad\qquad\qquad a \in A_{Atk}\}. \end{cases}$$

With the attacker part Atk, automaton (X, Atk) is defined with:

$$\begin{aligned} & states((X, Atk)) = \{(s, v) \mid s \in states(X), v \in S_{Atk}\}, \\ & start((X, Atk)) = \{(s, v) \mid s \in start(X), v \in S_{Atk}\}, \\ & ext((X, Atk)) = ext(X) \cup A_{Atk}, \\ & int((X, Atk)) = int(X), \\ & act((X, Atk)) = act(X), \text{ and} \\ & trans((X, Atk)) \\ & \quad = \{((s_1, v), a, (s_2, v)) \mid (s_1, a, s_2) \in trans(X), v \in S_{Atk}\} \\ & \qquad \cup T_{Atk}. \end{aligned}$$

For automaton (X, Atk), a state $(s, v) \in states((X, Atk))$ has two parts. The second part v is a state of the attacker, and the protocol part X does not change the second part. We can see that, in analogy with the above adversary model, timer-related variables correspond to the attacker's state v, and time actions correspond to attacker's actions T_{Atk}.

This paper has shown a basic idea to prove the anonymity of timed systems, but we have not introduced a formal definition for timed anonymity. We believe that it is possible to introduce such a formal definition as in a similar way of [7].

6 CONCLUSION

This paper discussed a method to verify the anonymity of timed systems. By describing a timed system with an I/O-automaton-based formal specification language, a proof technique for anonymity of untimed systems can be applied to a timed system.

This paper has shown a basic idea to prove the anonymity of timed systems with a small example. As described in Section 5 , the formalization of timed anonymity is not complete, and it is an important future work. Also, it is another interesting future work to deal with a larger example such as Mixnet [8].

Acknowledgment

This study is supported by the Grant-in-Aid for Scientific Research (C), No.26330166, of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

[1] K. van Hee and N. Sidorova, "The Right Timing: Reflections on the Modeling and Analysis of Time", *PETRI NETS 2013*, LNCS 7927, pp.1-20, Springer, 2013.

[2] M. Wehrle and S. Kupferschmid, "Mcta: Heuristics and Search for Timed Systems", *FORMAT 2012*, LNCS 7595, pp.252-266, Springer, 2012.

[3] Y. Kawabe, K. Mano, H. Sakurada and Y. Tsukada, "Theorem-proving anonymity of infinite-state systems". *Inf. Proc. Lett.*, vol. 101, no. 1, pp. 46–51, 2007.

[4] A. Bogdanov, "Formal verification of simulations between I/O-automata", Master's thesis, MIT, 2000.

[5] N. A. Lynch and F. Vaandrager, "Forward and backward simulations — part I: Untimed systems". *Inform. and Comput.*, Vol. 121, No. 2, pp. 214-233, 1995.

[6] N. A. Lynch, *Distributed algorithms*, Morgan Kaufmann Publishers, 1996.

[7] Y. Kawabe and H. Sakurada, "An adversary model for simulation-based anonymity proof". *IEICE Trans.*, Vol. E91-A, No. 4, pages 1112-1120, 2008.

[8] D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms". *CACM*, Vol. 24, No. 2, pp. 84-90, 1981.

Delta ISMS Model to Enhance Company-wide Information Security Management Using Incident Database: The Concept

Hiroshi Horikawa†1 Hisamichi Ohtani†2 Yuji Takahashi†3
Takehisa Kato†4 Fumihiko Magata†5 Yoshimi Teshigawara†6
Ryoichi Sasaki † 6 Masakatsu Nishigaki † 1

†1 Graduate School of Informatics, Shizuoka University, Japan
†2 Security Engineering Department, NTT DATA Corporation, Japan
†3 The Research Institute of Science and Technology, Tokyo Denki University, Japan
†4 Toshiba Corporation, Japan
†5 NTT Secure Platform Laboratories, NIPPON TELEGRAPH AND TELEPHONE CORPORATION, Japan
†6 School of Science and Technology for Future Life, Tokyo Denki University, Japan

hholy0403@gmail.com

Abstract - In this paper, we propose the Delta ISMS model, which strengthens company-wide information security management using an incident database. International standards of information security management systems (ISMS) have been established to provide useful guidelines for information security risk management to organizations so they can suitably respond to information security incidents. ISMS requires feedback and learning from incidents. However, even in ISMS certified organizations, information security incidents do not always diminish. This can indicate that these organizations do not effectively carry out the PDCA cycle of ISMS. We recognize that insufficient detailing of learning procedures hinders organizations from appropriately improving ISMS. Therefore, this paper aims to develop detailed procedures for learning from incidents to run the PDCA cycle. Our purpose is to give organizations a set of procedures to make a positive difference (delta) in their information security management, so we call our method "Delta ISMS". The procedures consist of operation of a company-wide incident database, routinized recalculation of the annual loss expectancy, countermeasure selection using an incident-countermeasure matrix (Delta ISMS table), and decision-making support for top management in the countermeasure selection process. The procedures are routinely applied by the headquarters under the direction of the Chief Information Security Officer (CISO). By following the procedures, the CISO headquarters monitors all kinds of data regarding incidents occurring in the organization and evaluate cost-effectiveness for additional countermeasures against the incidents. The evaluation results are summarized in a Delta ISMS table and used by top management in their decision-making on whether or not to adopt the additional countermeasures. Thus, company-wide information security governance can be achieved.

Keywords: information security management system (ISMS), incident database, information security incident, risk assessment, information security governance.

1. INTRODUCTION

International standards of information security management system (ISMS) have been established to provide useful guidelines for information security risk management to organization so they can suitably respond to information security incidents. ISMS requires feedback and learning from incidents. However, even in ISMS certified organizations, information security incidents do not always diminish [1][2][3]. This can indicate that these organizations do not effectively carry out the PDCA cycle of ISMS. We recognize that insufficient detailing of learning procedures hinders organizations from appropriately improving ISMS.

When an incident has occurred at one section, it is approached by the "ground zero" section (in some case, with the Computer Security Incident Response Team) until the first-order actions (i.e. the negative situation is found and deal) and second-order actions (i.e. the response to remove the cause of the nonconformity is taken). Annex of ISO/IEC27001 requires notes and reports of information security incidents [4] and the standard requires learning the result of the response as the third-order actions, but the procedures have not been given in detail. Even ISMS certified organizations do not maintain procedures to utilize data of incidents that have occurred at each section to improve security countermeasures of the whole organization.

Therefore, this paper aims to develop a detailed method and its procedures for learning from incidents to run the PDCA cycle. Our purpose is to give organizations a set of procedures to make a positive difference (delta) in their information security management, and thus we call our method "Delta ISMS". The procedures consist of operation of company-wide incident database, routinized recalculation of the annual loss expectation, and countermeasure selection using an incident-countermeasure matrix, and decision-making support for top management in the countermeasure selection process.

Information security management should be reinforced in the company-wide structure beyond divisions, shops, factories, and offices. While these respective parts of an organization can certify information security in ISMS, this paper focuses on a company-wide security management beyond a certified scope. This means that the proposed procedures should be routinely applied by a division across the whole company under direction of the Chief Information Security Officer (CISO). This division is called the "CISO headquarters" in this paper.

The CISO headquarters saves data of the incidents that have occurred in the organization into an incident database. By following the procedures, the CISO headquarters monitors all kinds of data regarding incidents occurring in the organization and evaluate cost-effectiveness for additional countermeasures against the incidents. When the first deal and secondary response by the "ground zero" section have ended, the CISO headquarters calculates the effect for the whole organization of adopting this secondary response adopted at the section concerned. Specifically, SLE (Single Loss Expectancy) and ARO (Annual Rate of Occurrence) are estimated for the cause of this incident, and a candidate countermeasure to reduce the risk of reoccurrence is enumerated. In addition, the introduction costs of the candidate countermeasure and residual risk are calculated, and results are stored in an incident database. The CISO headquarters probes the incident database periodically and chooses new countermeasures that should be adopted in the whole organization. The Delta ISMS table, a matrix of causes of an incident and their countermeasures, is used to choose high return-on-investment countermeasures.

As described above, the micro purposes of this paper are to embody control of Annex A.16 Information security incident management in ISO/IEC 27001:2013 (ISMS requirements) and to complement ISO/IEC 27001:2013.

On the other hand, there is less the mechanism to pick out the information from incidents in all sections that contributes to countermeasure improvement for the whole organization, and this is directly connected with stultification of incident reports in the management review. When incident data is reported to top management for a management review, only the information of whether or not the response is complete is shown. Nobody offers top management the judgment information (whether or not a countermeasure should be added) that is needed to improve security countermeasures for the whole organization. As a result, recognition of top management to information security risk management does not improve, and information security management of an organization becomes estranged from information security governance.

To resolve this problem, "decision-making support for top management in the countermeasure selection process" is introduced into the above Delta ISMS procedures. The evaluation results are summarized in the Delta ISMS table and used by top management in their decision-making on whether or not to adopt the additional countermeasures. The Delta ISMS table, where a cause of the incident and a countermeasure are in a matrix, is used to choose a candidate countermeasure that has high cost-effectiveness. Several patterns of candidate countermeasures (high cost and high effect, middle cost and middle effect, and low cost and low effect) are derived, and the top management can choose the most suitable countermeasure in accordance with the management strategy. A candidate countermeasure choice task can be formulated as a discrete optimization problem for a recommendation plan.

The CISO headquarters shows all the candidate countermeasure plans as well as the Delta ISMS table in the event of a management review. Top management uses these as judgement information and decides which countermeasure to adopt to improve security countermeasures of the whole organization. CISO will explain the information to executive officers on the board of directors, which improves the executive officers' recognition about information security risk management. Thus, ISMS of the organization is combined with information security governance, and company-wide information security governance can be achieved.

A described above, the macro purposes of this paper are to show consistent effective procedures to improve information security governance at the whole company level and to complement ISO/IEC 27014:2013 [5][6].

These series of procedures in the method are Delta ISMS. Chapter 2 reports related studies of risk analysis, ISMS, and information security governance. Chapter 3 describes the background of Delta ISMS and its approach. Chapter 4 explains the procedures of Delta ISMS in detail. Chapter 5 concludes the paper.

2. RELATED RESEARCH

Information security incident is an event that has a significant probability of compromising business operations and threatening information security [7]. Roberto stated that excellent leaders do not consider problems to be threatening and think all problems are a chance to improve and learn. He also stated that organizations should develop an incident report system [8]. An incident has a cause and produces an amount of damage. Hoo shows 29 kinds of common computer security incident [9].

By applying a risk management process, ISMS gives confidence to interested parties that risks are adequately managed [4]. Risk is the effect of uncertainly on objectives [7]. In the modeling of an information asset, the method to formulate the ALE (Annual Loss Expectancy) is adopted in the field of risk analyses [10]. ALE is formulated as follows;

$$ALE = SLE\times ARO$$

$$SLE = AV \times EF$$

Here, SLE is Single Loss Expectancy, ARO is Annual Rate of Occurrence, AV is Asset Value, and EF is Exposure Factor. We based our method on that of Nakamura et al. [11].

Since Japan tops the chart in the number of the ISMS certificates, this paper investigates Japanese literatures mainly. The current problems of ISMS can see from the results of questionnaires given to ISMS certified organizations by Nakao and Uchida [1][2] and a comparative study of Eguchi and Yamada [3]. The Ministry of Economy, Trade and Industry (METI) of Japan also issued an introduction to information security governance as guidelines that executive officers should use to make ISMS more effective[5]. These guidelines were updated as ISO/IEC[6].

3. BACKGROUND AND APPROACH OF DELTA ISMS

We consider the reason that information security incidents do not always diminish in ISMS certified organizations as well as the approach to resolve the problem.

3.1 The reason information security incidents do not decrease

Even in ISMS certified organizations, information security incidents do not always diminish. Eguchi and Yamada showed that enterprises that had acquired the international standard certification for ISMS could not always reduce incidents more than enterprises that had not [3]. Nakao and Uchida carried out questionnaires and obtained free answers of ISMS certified organizations (103 free answers in 2013 and 130 in 2010) [1][2]. We searched for "accident" and "incident" from 233 free answers and obtained 19 comments. We read the 19 comments and found that incidents had occurred at 11 organizations. Annex of ISO/IEC27001 requires that knowledge gained from information security incidents should be used to reduce the likelihood or impact of future incidents (A.1.6.1.6). When an organization responds as requested, incidents can be reduced, but there are some organizations in which incidents do not diminish. That suggests these organizations cannot use the knowledge gained from an incident for improvement. We consider one way to reduce information security incidents is to indicate detailed procedures to utilize incident data to the organizations that cannot reduce incidents.

3.2 Utilization of an incident database in the organization

To resolve the problem described in Sec. 3.1, the CISO headquarters uses an incident database, searches for the risks that lurk in the organization from incident data, and derives improved ideas for security countermeasures for the whole organization. In general, potential risks are difficult to search for, but a potential risk source from an incident (obvious risk) can be approached. By using the data of incidents that previously occurred at an organization to improve its security countermeasures, each organization can expect to find suitable countermeasures.

When the first deal and secondary response by an incident-investigation section have ended, the CISO headquarters calculates the effect for the whole organization of adopting the secondary response adopted at a section concerned. Specifically, SLE (Single Loss Expectancy) and ARO (Annual Rate of Occurrence) are estimated for the cause of this incident, and a candidate countermeasure to reduce the risk is enumerated. Additionally, the introduction costs of the candidate countermeasure and residual risk are calculated, and results are stored in an incident database.

The CISO headquarters chooses new countermeasures that should be adopted in the whole organization by probing an incident database at the Plan phase in the (n+1)-th rounds of the PDCA cycle. The Delta ISMS table is used to choose a high return–on-investment countermeasure. It is possible to formulate a countermeasure candidate choice task as a discrete optimization problem. Delta ISMS table is explained in detail in the next chapter.

3.3 The need for information security governance

According to METI's guidelines for information security governance [5], executives, in order to manage the risk associated with information assets, should establish a system to ensure the business activities in consideration of the information security of the whole organization. Figure 1 shows the framework of information security governance to be established in an organization.

Hitherto, the recognition for information security of executive officers and managers/employees is different, and building and operation of an information security risk management system are not totally optimized for the whole organization.

When a section of an organization is certified in ISMS, the PDCA cycle tends to be limited in the certified scope that is the manager and employee layer in the section. Therefore, it is important that the governing body drives the PDCA cycle of ISMS by monitoring, evaluating, and directing the PDCA cycle of ISMS of the section.

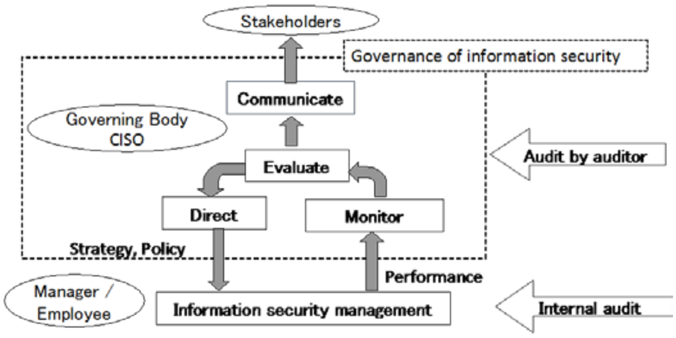


Figure 1 Framework of information security governance [5]

3.4 Dissociation of ISMS and information security governance

To lead information security governance to success, executive officers must understand and participate in the PDCA cycle of ISMS. According to results of a questionnaire given to ISMS certified organizations [1], 78.9% of ISMS operations managers are executive officers, 82.4% of executive officers participate in the PDCA cycle other than a management review. Both percentages are high, and participation of executive officers leads to an effective PDCA cycle of ISMS. However, on the other hand, to a question about "what I am mainly doing to improve the effect of ISMS," "Improving recognition and understanding of executive officers" and "developing a way to explain the cost-effectiveness" were respectively the 10th and 11th most common answers out of 11 items for three straight years. In other words, information for executive officers cannot be disseminated sufficiently from the manager/employee layer.

As the manager/employee layer does not have the method to offer the judgement information necessary to executive officers to improve a security countermeasures for the whole organization, managers report only the state of the incidents ("response is complete” or “response is continuing") for a management review. Therefore, executive officers cannot determine whether to add countermeasures or adopt amendments. There is no way the recognition of the executive officer for the information security risk management improves with this. As a result, the organization information security management is considered to have deviated from the information security governance.

3.5 Binding of ISMS and information security governance

As described in the next chapter, the Delta ISMS model provides the CISO headquarters a method and its procedures to derive multiple candidates for the countermeasures to the whole organization (for example, the three patterns) at the Plan phase in the (n+1)-th round of the PDCA cycle. For this reason, the CISO headquarters can present the multiple candidate countermeasures proposed along with the Delta ISMS table to the top management in the management review. Top management uses this information to determine the countermeasures of security. The CISO explains the information to the board of directors to improve their recognition of information security risk management. As a result, the monitoring cycle for the ISMS of the organization functions substantially, and ISMS and information security governance are joined together. Thus, company-wide information security governance can be achieved.

4 DELTA ISMS MODEL AND PDCA CYCLE

In this chapter, we describe the proposed Delta ISMS model in detail. The Delta ISMS model, using the actual incident data stored within the organization, embodies the method and its procedures for improving the information security risk management of the organization after the second round of the PDCA cycle of ISMS (Figure 2).

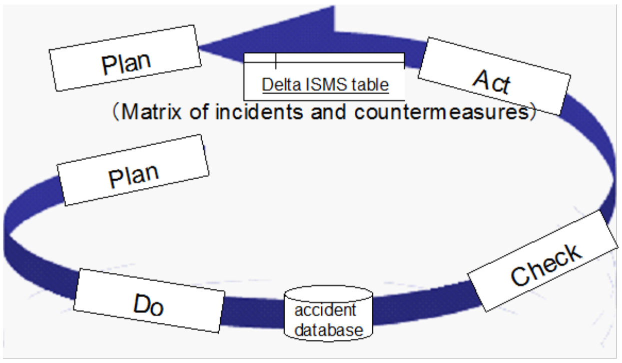


Figure 2 Delta ISMS model

4.1 Response to incident

When an incident occurs in the organization, the Delta ISMS performs not only conventional first and second actions specified in "10.1 nonconformity and corrective action" of ISMS [4] but also recommends performing a third-order action. Note that while the first- and second-order actions are carried out at the "ground zero" section, the third-order actions are carried out at the CISO headquarters periodically or when a serious incident occurs.

- First-order actions (to deal with the discovered nonconformity)
  - Take action to control the nonconformity (note, report, and evaluate).
  - Correct it.
  - Deal with the consequences.
- Second-order actions (to eliminate the cause of the nonconformity)
  - Review the nonconformity.
  - Determine its causes.
  - Determine if similar nonconformities exist or could potentially occur.
  - Implement corrective action.
  - Review the effectiveness of any corrective action taken.
- Third-order actions (the response of the organization as a whole)
  - From the nonconformities that have occurred in one section, assume the potential risk to the entire organization of those nonconformities and calculate the SLE (single loss expectancy) and ARO (annual rate of occurrence).
  - Decide the treatment of risk to the entire organization from SLE, ARO, and ALE (annual loss expectancy).

The SLE and ARO are difficult to accurately assess. Therefore, in general, they are often evaluated in the approximate order. For example, SLE is classified as "$ 100 - $ 1,000, $ 1,001 - $ 100,000, $ 100,001 - $ 1,000,000, $ 1,000,001 - $ 10,000,000, $ 10,000,001 or more" and, the ARO is classified to "once in several decades, once in a few years, once per year, several per year, dozens per year, or more." It is also possible to adopt this method in the third-order actions. ALE is SLE × ARO, when the calculation uses the intermediate value of each range.

4.2 Operation of the incident database

The CISO headquarters stores the incident responses of the results described in Section 4.1 in the incident database. Registration is carried out immediately after the incident response. The incident database contains date and time of incidents, the cause of the incident, the incident route, scope of influence, the content of first-order actions, the amount of damage, the countermeasures of second-order actions, the cost of second-order actions, and the content of third-order actions. Table 1 shows the specifications of the incident database.

Table 1 Incident database

Item	Meaning
Date and time	Occurrence date and time of the incident.
Cause of the incident	The contents of the incident (free format).
Type of the cause	The cause is selected from the following 13 types of classification. erroneous / lost or misplaced / unauthorized access / incorrect information takeout / mismanagement / bug, security hole / theft / internal fraud / misconfiguration / purpose outside use / worm, virus / unknown / other
Incident route	The route is chosen from the following seven types. USB, etc. / paper / PC / Internet / mobile phones and smart phone / e-mail / other
Scope of influence	The range of influence is selected, from near-miss to serious incident.
Content of first-order actions	The contents of the first-order actions (free format).
Amount of damage	The total cost until the incident is converged, including in-house artificial costs. In addition, costs of countermeasures to prevent recurrence are not included.
Countermeasures of second-order actions	The contents of the second-order actions (free format).
Cost of second-order actions	Countermeasure cost for preventing recurrence pile up, including in-house labor costs.
Content of third-order actions	Third-order response of content. Record potential risks to be assumed, SLE, ARO, ALE, and the treatment of risk.

An incident is an event where the risk was obvious. The incident database is not merely a list of the "facts of the incident." It will be used for searching out both potential risks and risk actualized from the incident.

All incidents that have happened in the organization, including information about near-misses, are recorded and stored in the incident database. In addition, for handling risk, not only incidents but also corrective items and suggestions for improving identified through external and internal audits can be used as effective information. Corrective items and recommendations for improving of the external and internal audits can be divided into the following three kinds.

- Items in accordance with the documents, such as rules and records of the organization.
- Items in accordance with the way of the ISMS, such as risk assessment and performance evaluation.
- Items in accordance with the nonconformity that was discovered by the field test.

In particular, nonconformity items discovered by the field test inspection can be treated as equivalent to incidents (actualized risk) to be added to the incident database.

4.3 Selection of the security countermeasures improvement plan

At the Plan phase in the (n+1)-th round of PDCA cycle, the CISO headquarters probes the incident database, overviews of all the countermeasures applied to the sections in which incidents occurred during the Do phase in the n-th round of PDCA cycle, and selects a candidate of the new countermeasures to be adopted as a whole organization.

Cause of the incident and countermeasures are in a many-to-many relationship. To select optimal countermeasures, the accumulated costs for countermeasures must be compared with the amount by which financial losses are reduced. To evaluate this, the Delta ISMS Table, a matrix of the cause of the incident and its countermeasures, is created.

Table 2 shows a Delta ISMS table. The abbreviations have the following meanings.

- LP_j: annual expected loss of the cause of the incident.
- R_{ji}: rate of ALE reduction by the countermeasures (0% to 100%).
- S_i: the presence or absence of each countermeasure (0 or 1).
- C_i: the cost of the countermeasures.

Table 2 Matrix of incidents and countermeasures (Delta ISMS table)

Incident	ALE	Costs of countermeasure 1 (S_1C_1)	Costs of countermeasure 2 (S_2C_2)	・・・	Costs of countermeasure i (S_iC_i)
1	LP_1	R_{11}	R_{12}	・・・	R_{1i}
2	LP_2	R_{21}	R_{22}	・・・	R_{2i}
・	・	・	・	・	・
・	・	・	・	・	・
・	・	・	・	・	・
j	LP_j	Rj_1	Rj_2	・・・	R_{ji}

Selection of the most investment countermeasures in the Delta ISMS is represented as a selection of countermeasures so that value E⊿ in Eq. 1 is the largest. Here, formulation of Eq. 1 refers to Nakamura et al.'s method [11].

Equation 1:

$$E_{⊿} = \sum_{j} \left\{ LP_j \left(1 - \prod_{i} \left(1 - R_{ji} S_i \right) \right) \right\} - \sum_{i} C_i S_i$$

Countermeasures can be distinguished into three kinds.

- Countermeasures already applied to the entire organization.
- Countermeasures partially applied to the organization in the second-order actions.
- Countermeasures unapplied to the organization.

The CISO headquarters selects the countermeasures that would be better applied to the entire organization from the countermeasures applied in the secondary treatment to the affected section, using the Delta ISMS table and Eq. 1. In this countermeasure selection process, by referring to the information security countermeasures collection such as ISO/IEC 27002: 2013 [12] and NIST SP800-53 [13], multiple candidate countermeasures are derived in response to setting costs and benefits of countermeasures. For example, strong countermeasures at a high cost (such as introducing a new information security system) are "high-ranked" candidates, weak countermeasures at low cost (such as education) are "low-ranked," and countermeasures in between are "middle-ranked." Here, the countermeasures selection should be considered, including retaining the risk, sharing the risk, and avoiding the risk [14].

Table 3 shows an example of reducing ALE by choosing countermeasures. If you choose Measure 2 (strap introduction: introduction cost 3,000 USD), cutting the risk of mobile phone loss and the USB memory loss by 30% will save 75,000 USD and 7,500 USD. Note that cost and ALE in Table 3 may change due to company size, business category, and prices.

Table 3 Example of reduction in ALE by choosing countermeasure

	measure	measure1	measure2	・・・	measure n
		change setting	strap	・・・	encryption software
cause of incident	cost ALE	30,000USD	3,000USD	・・・	40,000USD
e-mail wrong transmission	250,000USD	30%	0%	・・・	15%
mobile phone loss	250,000USD	0%	30%	・・・	0%
⋮	⋮	⋮	⋮	・・・	⋮
USB memory loss	25,000USD	0%	30%	・・・	40%

4.4 Determination of the improvement countermeasures by top management

The CISO headquarters, at the time of the management review, presents to the top management the improvement plan of security countermeasures for the entire organization described in Section 4.3. Top management utilizes the calculation results of Delta ISMS table and Eq. 1 as information to determine whether to adopt revised or new countermeasures, and thus can select countermeasures from among the candidates proposed. As a result, security countermeasures can be improved across the organization in a way that is consistent with business management strategy. The CISO explains this information to board of directors, so recognition of the governing body continues to improve with respect to the information security risk management. As a result, the monitoring cycle for the ISMS of the organization functions substantially to bind the ISMS and information security governance.

4.5 Improvement by Delta ISMS

Figure 3 shows improvement of information security management achieved by Delta ISMS. By focusing on the comparison (difference) of incident data between each PDCA cycle, the trend of incidents occurred in the whole organization, which is important for the continuous improvement of ISMS, also becomes easy to observe.

5 CONCLUSION

To improve the ISMS effectively and strengthen the information security governance, this paper proposed a method and its procedures for learning from incidents. The procedures consist of operation of a company-wide incident database, routinized recalculation of the annual loss expectancy, countermeasure selection using an incident-countermeasure matrix, and decision-making support for top management in the countermeasure selection process.

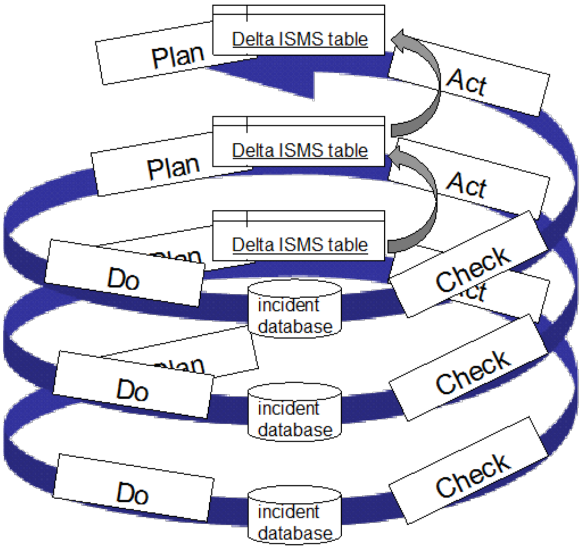


Figure 3 Improvement of information security management achieved by Delta ISMS

By evaluating and applying the method in a real organization in the future, we want to strive to improve the method and its procedures. In addition, we want to consider automating the incident database registration (for example, by fusing it with forensic processes).

REFERENCES

[1] H. Nakao and K. Uchida, Survey of Registered Organizations of Information Security Management System (ISMS), Journal of Tokyo University of Information Sciences, Vol. 17, No. 2, pp. 125–182 (2014) (in Japanese).
[2] New Media Development Association, Survey of Registered Organizations of Information Security Management System (ISMS) (2010) (in Japanese).
[3] A. Eguchi and S. Yamada, Comparative Study of Information Security Incidents Based on ISO 27001 Certification, Japan Society of Security Management, Vol. 27, No. 1, pp. 3–16 (2013) (in Japanese).
[4] ISO/IEC 27001:2013, Information technology - Security techniques - Information security management systems - Requirements, ISO/IEC (2013).
[5] Ministry of Economy, Trade and Industry of Japan, Guidance for introducing information security governance (2009) (in Japanese).
[6] ISO/IEC 27014:2013, Information technology - Security techniques - Governance of information security, ISO/IEC (2013).
[7] ISO/IEC 27000:2014, Information technology - Security techniques - Information security management systems - Overview and vocabulary, ISO/IEC (2014).
[8] M.A. Roberto, Know what you don't know, Wharton School Publishing (2009).
[9] K.J.Soo Hoo, How Much Is Enough? A Risk-Management Approach to Computer Security, Consortium

for Research on Information Security and Policy (CRISP) (2008).

[10] R. Bojanc and B. Jerman-Blazic, An economic modelling approach to information security risk management, International Journal of Information Management, No. 28, pp. 413–422 (2008).

[11] I. Nakamura, T. Hyodo, M. Soga, T. Mizuno, and M. Nishigaki, A Practical Approach for Security Measure Selection Problem and Its Availability, IPSJ (Information Processing Society of Japan) Journal, Vol. 45, No. 8, pp. 2022–2033 (2004) (in Japanese).

[12] ISO/IEC 27002:2013, Information technology - Security techniques - Code of practice for information security controls, ISO/IEC (2013).

[13] R. Ross, et al. (NIST): Security and Privacy Controls for Federal Information Systems and Organizations, NIST SP 800-53 Revision 4 (2013).

[14] ISO Guide 73:2009, Risk management - Vocabulary, ISO (2009).

A Method for Removing Ambiguity in Designing Sequence Diagrams for Developing Communication Programs

Satoshi Harauchi[*], Kozo Okano[**] , and Shinpei Ogata[**]

[*]Advanced Technology R&D Center, Mitsubishi Electric Corporation, Japan
Harauchi.Satoshi@bc.MitsubishiElectric.co.jp

[**]Electrical and Computer Engineering, Shinshu University, Japan
{okano, ogata}＠cs.shinshu-u.ac.jp

Abstract – It is important to eliminate rework of the design for developing software systems. Faults and errors in design should be extracted in order not to leave them to later phases such as implementation or test. When developing communication programs, faults may be into programs against designer's intention. It is difficult to detect them by reviewing them especially in designing complicated and asynchronous communication programs. In this paper, we propose a method to detect faults when designing communication programs. The method focuses on the sequence diagram which represents exchanges of messages between lifelines and aims at removing ambiguity for the order of exchanges. The method consists of following procedures. The method generates model descriptions and test expressions from sequence diagrams, and executes model checking with both of them. Then the method notifies the information in diagrams, at which an error occurs in model checking unless model descriptions satisfy test expressions. The notification enables designers to eliminate inconsistency from diagrams. This paper describes the problem on developing sequence diagrams, our method to solve the problem, implementation with UML 2.0 and evaluation of the method. The result of evaluation shows that the method is effective though the time for generation depends on the complexity of diagrams.

Keywords: Communication programs, Sequence diagrams, Model checking, Promela, Linear Temporal Logic, UML.

1 INTRODUCTION

It is important to remove faults and errors for software systems in order to develop software with high reliability. The faults may be inserted into software not only in implementation but in software design. It is much more important to detect and remove faults in design than in implementation. In general it takes more costs and efforts to rework and modify the design when removing them. Hence, they should be extracted in order not to leave them to later phases such as implementation or test.

It is also important to detect faults for developing communication programs. The bigger communication programs become, the more complicated the design will be. It is difficult to detect them by reviewing the complicated design. Therefore the design is expected to be supported to detect them.

In this paper, we propose a method to detect faults when designing communication programs. The method focuses on the sequence diagrams which represent asynchronous exchanges of messages between lifelines. The diagrams will be complicated when designing complex communications. Complicated exchanges of messages cause faults frequently because ambiguity for the order of exchanges remains in the diagrams. The ambiguity shows that the order to receive messages could not be determined when several messages are asynchronously transmitted to a specific lifeline. The method aims at detecting the ambiguity and helps to remove faults inserted in the diagrams.

The method consists of following procedures. At first, the method generates formal descriptions written in Promela[1] from sequence diagrams. Components such as lifelines and messages described in the diagrams correspond to the elements of Promela. Next, the method generates test expressions with Linear Temporal Logic(LTL). The expressions are obtained from every message for each lifeline. Generated expressions are used for checking the order exhaustively. Then, the method executes model checking with formal descriptions and test expressions. Failing to satisfy the expressions for formal descriptions means the existence of ambiguity related to the order of messages. The method finally indicates the position in diagrams which cause an error in model checking. The indication helps designers to correct diagrams and remove the error. The method is reapplied from the top of the procedure after removal and the designers apply repeatedly until errors do not occur.

We implement the method as a tool with UML 2.0 and evaluate 2 aspects. The first aspect focuses on the number of indications generated by the tool and the time to spend the procedures for various kinds of sequence diagrams. The second goes to the diagrams applied to a product. The result of evaluation shows that the method provides 10 candidates for modifying the diagrams and 4 candidates out of 10 are required to correct it according to interview of the engineer who engaged in the above product.

This paper provides the method in detail. First, we describe the problems on developing sequence diagrams in section 2, and our method to overcome the problems in section 3. We then describe implementation and evaluation of the method in section 4.

2 PROBLEMS ON DEVELOPING SEQUENCE DIAGRAMS

Figure 1 shows an example of sequence diagrams. In the figure, lifeline A, B and C asynchronously communicate each other. The figure is used when designing the specification of communications. The design is considered to be completed after reviewing the diagrams, and then programs are implemented according to the diagrams.

However, a fault may occur in figure 1 after implementation. Figure 2 shows sequence diagrams in case of the fault. According to the Figure 1, “msg6” shall be sent from lifeline C to B in figure 1 after “msg3”. Lifeline B shall receive “msg6” before “msg5”. Nevertheless, the “msg6” may reach lifeline B after “msg5” against the intention. Lifeline B emits an error due to the violation of specifications.

Removing the fault requires time and efforts. The time depends on the causes of the fault. The fault shown in figure 2 cannot be detected in unit test, but can be often detected in integration test. Accordingly, it is necessary to rework of design, implementation and test. The time to repair faults increases in accordance with the number of faults and the complexity of diagrams.

The error shown in figure 1 is caused by ambiguity of sequence diagrams. Figure 1 indicates that the situation which “msg6” reaches lifeline B cannot be determined. This paper describes removing such ambiguity in designing sequence diagrams.

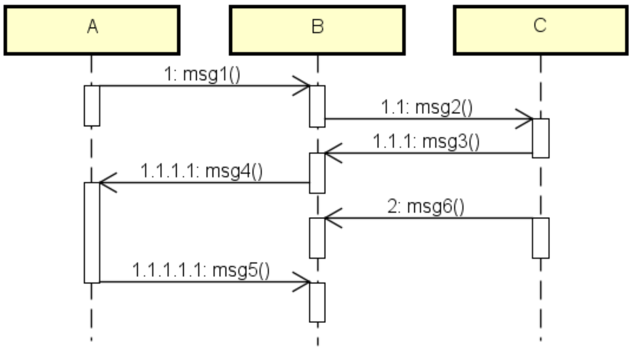


Figure 1: An example of sequence diagrams

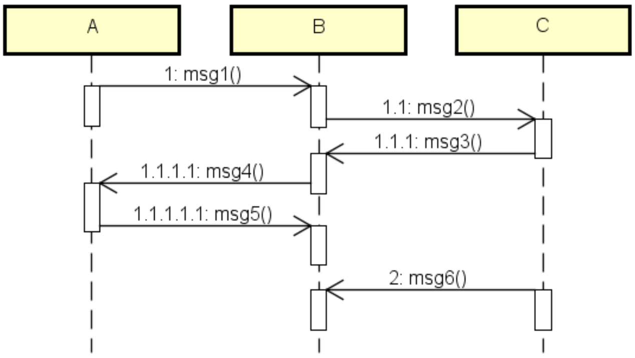


Figure 2: Sequence diagrams in case of a fault

3 PROPOSED METHOD

3.1 Outline

The method eliminates ambiguity for the order of messages with the semi-automatic modification of sequence diagrams. It is possible to correct automatically, but we adopt the procedure which provides the candidates for modifying, enables designers to select an appropriate candidate and corrects the diagrams by the use of the selected candidate.

The input for the method is the diagrams written in XML. The output is the diagrams without the ambiguity. The specification of diagrams uses UML 2.0[1]. The diagrams allow asynchronous representation of messages.

Proposed method consists of four steps shown in figure 3.

STEP 1: Generating formal descriptions

STEP 2: Generating test expressions

STEP 3: Model checking and the generation of candidates for modifying diagrams

STEP 4: Correct diagrams

We describe the details of each step in the following sections.

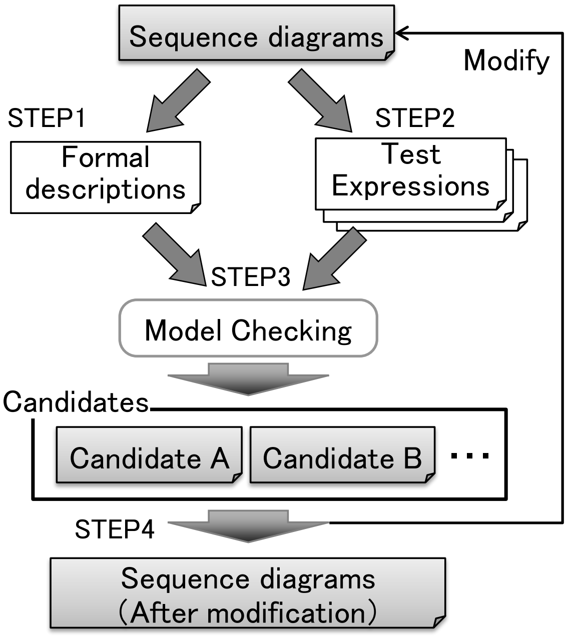


Figure 3: An overview of proposed method

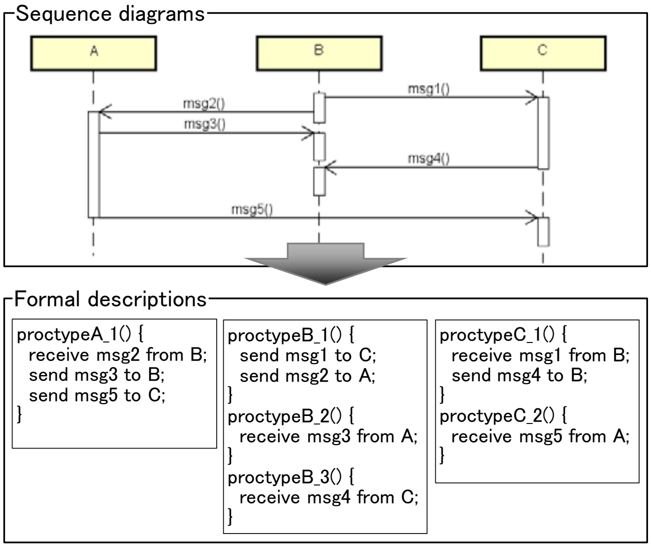
3.2 STEP 1 Generating formal descriptions

This step generates formal descriptions from the input. XML is obtained with astah* professional [2]. The formal descriptions are written in Promela [3] used by SPIN model checker. The method shown by Lima [4] is referred for generating descriptions. A lifeline and a message for each execution specification correspond to a process and a channel with variables defined by Promela, respectively.

Figure 4 shows an example of generation. The upper part of figure shows sequence diagrams and the lower part shows the summary of formal descriptions generated from the diagrams. Process B_1, Process B_2 and Process B_3 are generated since lifeline B has 3 execution specifications. Each message is translated into 2 descriptions. For example, “msg1” generates one description that B sends “msg1” to C and another one that C receives “msg1” from B.

The generation for each execution specification maintains the order of messages within the execution specification. Furthermore, the generated descriptions represent the ambiguity for the order of asynchronous messages. The method by Lima does not generate the description for each execution specification.

The order of messages has an assumption of the communications. Successive messages will be received successively. For example, it is not guaranteed that “msg3” is received before the reception of “msg5”. However, we assume that “msg3” is received before “msg5” because “msg5” is sent successively.



```
proctypeA_1() {
  receive msg2 from B;
  send msg3 to B;
  send msg5 to C;
}
```

```
proctypeB_1() {
  send msg1 to C;
  send msg2 to A;
}
proctypeB_2() {
  receive msg3 from A;
}
proctypeB_3() {
  receive msg4 from C;
}
```

```
proctypeC_1() {
  receive msg1 from B;
  send msg4 to B;
}
proctypeC_2() {
  receive msg5 from A;
}
```

Figure 4: An example of generating formal descriptions

3.3 STEP 2 Generating test expressions

This step generates test expressions from the input. These are written in Linear Temporal Logic(LTL) expressions. The expressions are used to check whether diagrams have the ambiguity for the order of messages or not. Each expression is generated from 2 messages which adjoin each other.

We can show the example using figure 4. Lifeline B in sequence diagrams has 4 exchanges of messages. The item to be checked is extracted from two adjacent messages such as “msg2” and “msg1”. The items in the lifeline are obtained by all of adjacent messages. Therefore the items in relation to lifeline B are described as follows.

(a) Whether “msg2” is sent before “msg1” is sent
(b) Whether “msg3” is received before “msg2” is sent
(c) Whether “msg4” is received before “msg3” is received

The method generates expressions below from (a) to (c).
(a’) (send “msg2”) before (send “msg1”)
(b’) (receive “msg3”) before (send “msg2”)
(c’) (receive “msg4”) before (receive “msg3”)

The method then translates above 3 items into test expressions below.
(a’’) ¬ (send “msg2”) ∪ (send “msg1”)
(b’’) ¬ (receive “msg3”) ∪ (send “msg2”)
(c’’) ¬ (receive “msg4”) ∪ (receive “msg3”)

3.4 STEP 3 Model checking and the generation of candidates for modifying diagrams

This step executes model checking with formal descriptions and test expressions. Then, the method provides candidates which indicate how to modify diagrams. Failing to satisfy the expressions means the existence of ambiguity. The result of failure gives a pair of 2 messages described in test expressions. The candidate shows diagrams with the message inserted between the pair of 2 messages.

Figure 5 shows an example of the candidate for modification. The diagrams shown in figure 5 turn out to have ambiguity for following item.

(c) Whether “msg4” is received before “msg3” is received

Therefore lifeline A is added to transmission of “msg6” after sending “msg3” and lifeline C is added to the reception of “msg6” before sending “msg4”.

Figure 6 shows the procedure of generating candidates for all test expressions. First of all, the method selects a test expression among all expressions. The method then executes model checking with formal descriptions and selected expression by SPIN model checker. The execution moves to following processes depending on to the result of checking.

If no ambiguity exists : The method executes model checking with same formal descriptions and another test expression.

If ambiguity exists : The method generates a candidate for modifying from the test expression. The candidate is showed to a designer and diagrams will be corrected if he decides to apply the candidate. The method then executes model checking with corrected formal descriptions and another test expression.

The procedure is repeatedly applied and terminates if model checking is executed for all test expressions.

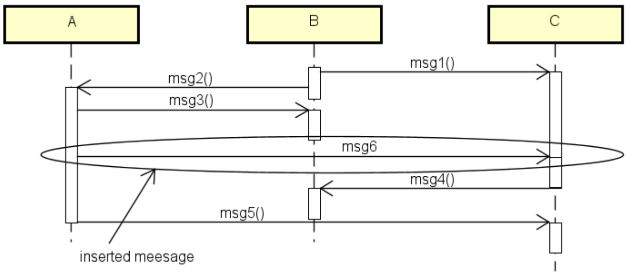


Figure 5: An example of modification

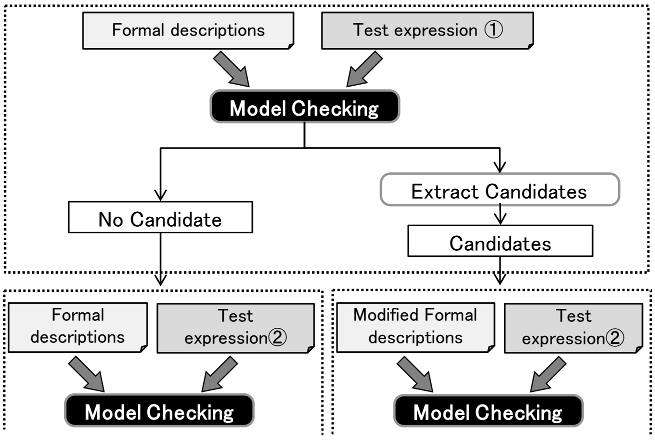


Figure 6: Procedure of generating candidates

3.5 STEP 4 Correcting diagrams

This step corrects XML with selected candidate in previous section. The candidate has the information for additional message such as name of the message and the place to insert. The method corrects the definition of messages and the information related to lifelines. The procedure of this step is over if all candidates indicated by designers are reflected into the diagrams.

If all test expressions pass model checking after the diagrams are corrected, no existence of ambiguity for the order of messages is proved for specified diagrams.

4 EVALUATION

We implement the method as a tool with Java and shell scripts in order to evaluate the performance of proposed method. We evaluate following 2 aspects with the tool.

Aspect 1: The number of candidates for modification generated by the method and the time to spend the execution of the method

Aspect 2: The evaluation of candidates

Aspect 1 focuses on various kinds of sequence diagrams and aspect 2 focuses on the diagrams applied to a product. Computer specification used for evaluation is described as follows.

OS: Windows 7 Professional
CPU: Intel Xeon E5607 2.27GHz×2
Memory: 16 GB
SPIN: Version 6.3.2

The size of state vector for SPIN model checker is defined as 1024 bytes.

4.1 Evaluation method

4.1.1. Aspect 1

We collect the sequence diagrams described as examples in existing researches [5]-[9] and applications [10],[11] related to sequence diagrams. Furthermore, we produce another sequence diagrams with the additional lifelines and messages for specific diagrams of collected examples. We apply the tool to those diagrams. We then enumerate lifelines, messages and candidates for modifying, and measure the time spent for the execution.

The measurement is executed assuming that designers adopt all candidates shown by the tool.

4.1.2. Aspect 2

We apply the tool to the diagrams used for the product. The diagrams is rewritten with astah* professional[2]. This aspect checks the ability to detect faults shown in figure 2. We confirm that the candidates generated by the tool are appropriate to correct them.

4.2 The result of evaluation

4.2.1. Aspect 1

We apply the tool to 11 sequence diagrams. The result of evaluation is shown in table 1. 11 diagrams consist of 7 diagrams collected from references and 4 diagrams where the number of lifelines in reference [9] is edited (described as [9]-1, 2, 3, 4). Columns 1 to 3 in table 1 show the information in the diagrams and columns 4 to 7 show the result applied to the tool. Columns 1, 2 and 3 indicate the source of diagrams, the number of lifelines in diagrams and the number of messages, respectively. A column 4 describes the number of candidates generated by the tool. Columns 5 to 7 show the time for the execution applied to the tool. The time is measured for each step. A column 5 indicates the sum of time spent for STEP 1 and 2 since both of them are executed in parallel with the same input.

No significant differences can be seen in the time spent for STEP 1 and 2. However, the time of [9]-1, 2, 3 and 4 is larger. A large number of lifelines and messages causes the large time. The time spent for STEP 3 becomes large as the number of lifelines or messages in diagrams become large. The larger the number of candidates is, the larger the time spent for STEP 4 will be although no major differences can be observed in the time.

The time spent for STEP 3 in [9]-4 is quite smaller than in [9]-3 though the number of lifelines and messages are very large. The reason is that model checking in that case could not be executed due to insufficient memory. Hence, the number of candidates becomes zero.

4.2.2.Aspect 2

Some diagrams used for the product are supplied for evaluating aspect 2. We select 2 diagrams from supplied diagrams and apply the tool to 2 diagrams where ambiguity for the order of messages may exist. The applied diagrams are shown in figure 7 and 8. The result of application is shown in table 2. Columns in table 2 are same as in table 1.

We describe one candidate of 10 candidates obtained by the tool. The candidate is shown in figure 9. The candidate indicates the modification inserting “msg8” which lifeline B sends and lifeline D receives between “msg7” and “msg6”. The candidate is generated from the ambiguity that “msg7” may reach lifeline D after “msg6”. Therefore, proposed method is able to detect the fault in figure 7.

Table 1: The result of evaluation for aspect 1

	Lifelines	Messages	Candidates for modifying	The time spent for the execution(seconds)		
				STEP 1,2	STEP 3	STEP 4
[5]	7	11	3	0.44	20.21	0.43
[6]	4	12	2	0.42	25.19	0.41
[7]	3	8	0	0.39	15.55	0.41
[8]	3	4	0	0.37	5.96	0.40
[10]	5	6	1	0.42	8.95	0.41
[11]	6	24	0	0.49	56.15	0.42
[9]	7	22	6	0.46	52.08	0.48
[9]-1	12	44	9	0.58	124.77	0.45
[9]-2	22	88	18	0.98	9136.59	0.52
[9]-3	32	132	24	0.97	14054.17	0.56
[9]-4	37	154	0	1.20	744.73	0.56

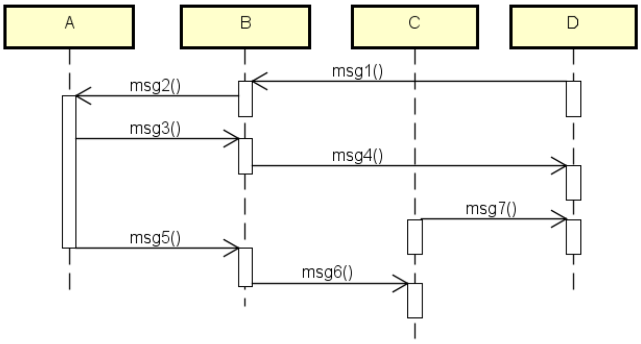


Figure 7: An example of diagrams used for aspect 2

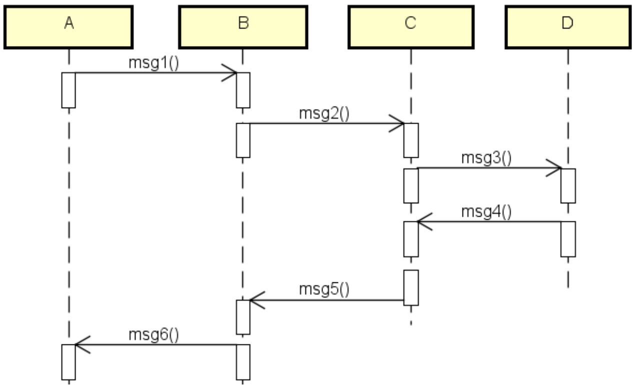


Figure 8: An example of diagrams used for aspect 2

Table 2: The result of evaluation for aspect 2

	Lifelines	Messages	Candidates for modifying	The time spent for the execution(seconds)		
				STEP 1,2	STEP 3	STEP 4
Figure 7	4	7	5	0.40	14.04	0.62
Figure 8	4	6	5	0.41	12.15	0.52

We request the engineer who engaged in the product to check 10 candidates including figure 9. We ask them to confirm the validity of candidates by selecting one appropriate answer out of following 3 answers.

(1) This modification should be done.
(2) This modification need not to be done
(3) This modification should not be done

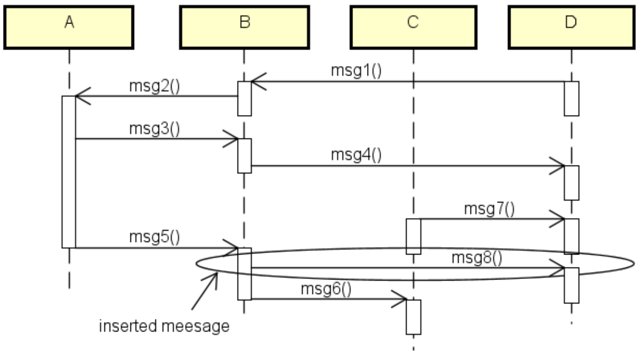


Figure 9: A candidate for modifying

The result shows that (1) is selected for 4 candidates including figure 9. Another result shows that (2) is selected for 5 candidates and (3) for 1 candidate.

4.3 The validity of evaluation

In aspect 1 we collect the sequence diagrams showed in existing researches and tools, and evaluate the number of candidates for modifying and the time spent for the execution by applying the tool to the diagrams. In this paper we apply the tool to only 11 sequence diagrams. The number of lifelines and messages written in the diagrams is limited. Hence, we may obtain another result when applying the tool to the large-scaled diagrams.

In aspect 2 we confirm the possibility to detect faults and generate the candidates by using the diagrams with the ambiguity developed for the product. We use only 2 diagrams for evaluation. It is necessary to obtain a large variety of diagrams used for various products and evaluate them in order to acquire more general results.

5 CONCLUSION

This paper proposed the method to detect faults when designing sequence diagrams which describe asynchronous exchanges of messages. The method transforms formal descriptions written in Promela and test expressions written in LTL from sequence diagrams. The method then executes model checking for all expressions with the descriptions. When an error occurs on the execution, it provides information in diagrams. The information enables designers to remove faults and keep consistency.

We implement and evaluate the method with 2 aspects. In the first aspect, we measure the number of information and the time to spend the execution of the method. In the second one, we applied the method to the diagrams used by a product. The application generates 10 information and evaluates the validity of the information. According to the interview of the engineer, 40% among the information is effective for correcting the diagrams. Applying the method to various development of diagrams and increasing the number and the kinds of candidates are considered as future issues.

REFERENCES

[1] "UML2.0". http://www.omg.org/spec/UML/2.0/. <referred on 13th September 2016>

[2] "astah* professional". http://astah.net/editions/professional <referred on 13th September 2016>

[3] G.J. Holzmann, "The model checker SPIN", IEEE Transactions on software engineering, vol.23, no.5, pp.279-295,1997.

[4] V. Lima, C. Talhi, D. Mouheb, M. Debbabi, L. Wang, and M. Pourzandi, "Formal Verication and Validation of UML 2.0 Sequence Diagrams using Source and Destination of Messages", Electronic Notes in Theoretical Computer Science, vol.254, pp.143-160, 2009.

[5] P. Baker, P. Bristow, C. Jervis, D. King, R. Thomson, B. Mitchell, and S. Burton, "Detecting and Resolving Semantic Pathologies in UML Sequence Diagrams," In the Proceedings of the 10th European software engineering conference held jointly with 13th ACM SIGSOFT international symposium on Foundations of software engineering, pp.50-59,ACM, Sep. 2005.

[6] S. Bernardi, S. Donatelli and J. Merseguer "From UML Sequence Diagrams and Statecharts to Analyzable Petri Net models" In the Proceedings of the 3rd international workshop on Software and performance, pp.35-45, ACM, July 2002.

[7] D. Harel and S. Maoz, "Assert and Negate Revisited: Modal Semantics for UML Sequence Diagrams" Software & Systems Modeling, vol.7, no.2, pp.237-252, 2008.

[8] H. Shen, R. Krishnan, R. Slavin, and J. Niu, "Sequence Diagram Aided Privacy Policy Specication", IEEE Transactions on Dependable and Secure Computing, pp.1-1, 2014.

[9] B. Mitchell, "Characterizing Communication Channel Deadlocks in Sequence Diagrams" IEEE Transactions on Software Engineering, vol.34, no.3, pp.305-320, 2008.

[10] "Lucidchart". https://www.lucidchart.com/.<referred on 13th September 2016>

[11] "tracemodeler". http://www.tracemodeler.com/.<referred on 13th September 2016>

[12] H. Shen, R. Krishnan, R. Slavin, and J. Niu, "Sequence Diagram Aided Privacy Policy Specification" IEEE Transactions on Dependable and Secure Computing, pp.1–1, 2014.

[13] D. Harel and S. Maoz, "Assert and Negate Revisited: Modal Semantics for UML Sequence Diagrams" Software & Systems Modeling, vol.7, no.2, pp.237–252, 2008.

A New Proposal of Generating Counter-example in Model Checking Using Test Automaton

Chikyu Yanagisawa[†], Shinpei Ogata[‡], and Kozo Okano§
[†‡]§Shinshu University, Japan
[†]15tm535d@shinshu-u.ac.jp
[‡]ogata@cs.shinshu-u.ac.jp
§okano@cs.shinshu-u.ac.jp

Abstract - In recent years, "post-verification" of systems attracts attention. The post-verification of systems creates models from a system when a fault has occurred in the system. It then finds out the cause of the fault using formal methods such as model checking. Model checking inspects logically and exhaustively whether a given property is satisfied or not. It creates a model from given source code or systems, then derives a logical expression from the requirements specification and an inspection item, and finally enters them into a model checker. A counter-example is a trace information to help localization of bugs and it is generated by a model checker when the property does not hold. Current model checkers, however, often cannot generate users' expected counter-examples due to mainly their searching algorithms. This paper derives a counter-example which is expected by a user. The derive method first creates test automata which represent roughly behavior of the expected counter-example. Then it performs model checking on a parallel composition of the original automaton with the test automata. We have applied the proposed method to a case study of water tanks of a chemical plant and confirmed its usefulness.

Keywords: counter-example, model checking, automaton

1 INTRODUCTION

Advanced information society needs more reliable software and techniques to develop such software. Software testing has been performed and studied for long year to ensure quality in the process of production of software. The conventional tests are, however, known to take a lot of resources. Therefore, formal methods are attracting attention as a way of improving the quality of software. A formal method is a method which describes the requirements and design of information systems (software and hardware) using a mathematical based language and provides mechanism to infer that the system satisfies users' requirements. This study uses model checking [2], which is one of the formal methods.

Model checking inspects logically and exhaustively whether a given property is satisfied or not. It creates a model from source code or systems, then derives a logical expression from the requirements specification and an inspection item, and finally enters them into a model checker. A "post-verification" of the system is an application of model checking. The post-verification of systems models the system when a fault has occurred in the system [10]. It then finds out the cause of the fault using formal methods, while the conventional approaches carry out cause isolation by log analysis of a particular system of failure.

Modern society needs post-verification because sometimes faults of systems might give serious impacts on the society. As a specific example for a fault due to a system malfunction, a system trouble of Japanese airline that occurred in 2016 can be enumerated.

A counter-example becomes a key when we perform the post-verification using model checking. A counter-example specifies in general that "an example that refutes or disproves a hypothesis, proposition, or theorem." When using counter-example for post-verification, we regard part of "hypothesis, proposition, or theorem" as properties which must be fulfilled. Thus, counter-example becomes an example of not satisfying the properties, thus counter-example can be a diagnosis of how the system fails by tracing it. Research has been conducted for generating a counter-example that easy to understand for humans [8, 9].

A counter-example, however, may not be one which a user expects due to searching algorithms used in the model checker. This trend is especially noticeable in cases where the counter-example including loop structures. This paper, in order to solve the problem, creates test automata to guide counter-examples for the model represented by time automata [4] which is used in UPPAAL [5], an integrated tool for modeling, validation and verification of real-time systems. For that purpose, our proposed method creates a coarse behavior series of counter-examples represented in test automata [16–18]. A parallel composition of the test automata and the original automaton lead counter-examples of the original model. In addition, we have applied our technique to a chemical plant system example, and we confirmed the method is effective.

Section 2 describes the time automata and test automaton. Section 3 describes the proposed method. Section 4 performs a diagnosis of system failure that occurred in the chemical plant systems using our method. Discussion is given in Section 5. Summary and future works are also given in Section 6.

2 PRELIMINARIES

2.1 Model Checking

Model checking [2, 3] of an automaton can be formulated as follow.

Definition 2.1 (Model Checking)
Input1: an automaton A

Input2: a temporal logic expression p
Output: $A \models p$ or $A \not\models p$
Output(optional): If $A \not\models p$, then a counter-example CE

In usual, Computational Tree Logic (CTL) is used as a temporal logic for a timed automaton [5].

Intuitively $A \models p$ means that the behavior (possible runs) of A satisfies the property expressed in p. Automaton A is also called a model. Thus, model checking is checking process whether a logic expression p holds under the model represented in A.

Typical properties are $\mathbb{AG}q$, $\mathbb{EF}q$ and so on. $\mathbb{AG}q$ and $\mathbb{EF}q$ mean that "for any path, always q holds," and "for some path, eventually q holds," respectively. $\mathbb{AG}$ and $\mathbb{EF}$ are called temporal operators.

For a state s, we can consider a property $\neg\mathbb{EF}s$, which means that starting from the initial state, the automaton cannot reach the state s.

Figure 1 represents model checking process.

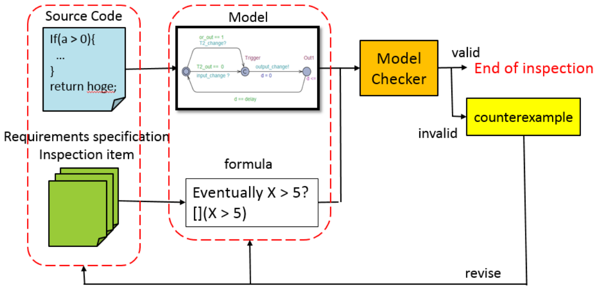


Figure 1: Model Checking Process

2.2 Timed Automaton

A timed automaton uses clocks to refer time. The clocks can be regard as precise analog clocks. Every clock autonomously uniformly and at the same rate increases the value, independently from the behavior of timed automaton. A timed automaton cannot control the clocks except for reset; it can neither put some clocks forward, backward nor stop them. It can only reset some of clocks. The reset clocks make their values 0. they, however, immediately increase their values again.

Definition 2.2 (Clock set C) *By C we denote a finite set of clocks. By x_i $(0 \le i \le |C|-1)$ we denote an element (each clock) in C.*

When there is no confusion we might use literals (without index) x, y, z, and so on to denote clocks.

Since each clock has its time value as a non-negative real, notion of "clock evaluation" is needed.

Definition 2.3 (Clock Evaluation) *Clock evaluation $\nu (\in \mathbb{R}_{\ge 0}^{|C|})$ for clock set C is a $|C|$-dimension vector over $\mathbb{R}_{\ge 0}$.*

An i-th element ν^i of ν corresponds to the time value of clock x_i.

We use the term "evaluation" according to the original paper [15]. Paper [15] defines the evaluation as a mapping from clocks to reals, however, we define ν just as a real vector, in this paper. Since clock evaluation changes according to the elapsed time, and a timed automaton might reset some of clocks to 0 when a transition fires, we introduce two operations on clock evaluation.

Definition 2.4 (Operations on Clock Evaluation) *For a real value d, $\nu + d = (\nu^0 + d, \nu^1 + d, \ldots, \nu^{|C|-1} + d)$.*

For a set of clocks r, $r(\nu) = (r(\nu^0), r(\nu^1), \ldots, r(\nu^{|C|-1}))$, where

$$r(\nu^i) = \begin{cases} 0 : x_i \in r, \\ \nu^i : \text{otherwise}\ . \end{cases} \tag{1}$$

The first operation $+d$ means that every clock increases its value uniformly and at the same rate. The second operation $r(\cdot)$ means that every clock specified in r are reset.

Next we define clock constraints on C, which are used as guards and invariants of a timed automaton.

Definition 2.5 (Differential Inequalities on C) *Syntax of a differential inequality in on a clock set C is given as follows:*

$$in ::= x_i - x_j \sim a$$

$$\mid x_i \sim a,$$

where x_i and $x_j \in C$, a is a literal of an integer constant, and $\sim \in \{\le, \ge, <, >\}$.

Differential inequalities $x_i \sim a$ and $x_i - x_j \sim a$ are true *iff $\nu^i \sim a$ and $\nu^i - \nu^j \sim a$ are* true*, respectively.*

Definition 2.6 (Timed Automaton) *A timed automaton $\mathscr{A}$ is a six-tuple (A, L, l_0, C, I, T), where*
A: a finite set of actions;
L: a finite set of locations;
$l_0 \in L$: an initial location;
C: a clock set;
$I : L \to c(C)$: a mapping from a location to a clock constraint, called a location invariant, or simply an invariant; and
$T \subset L \times A \times c(C) \times 2^C \times L$ is a set of transitions, where $c(C)$ is a set of clock constraints; and 2^C is a super set of sets of clocks.

Elements of the first and last L stand for locations the transition starting from and going to, respectively. An element of A is an action associated with the transition. A clock constraint in $c(C)$ of the transition is called a guard. An element in 2^C is called a set of clocks to be reset.

We denote $(l_1, a, g, r, l_2) \in T$ by $l_1 \overset{a,g,r}{\to} l_2$.

2.3 Test Automaton

In this paper, the test automaton [16–18] is used as guideline for deriving a desired counter-example by tracing the attention to the transition of the original model. A test automaton is in usual used in a parallel composition of the original automaton, in order to check complex property. Model checking uses a logical temporal expression as a property to check.

The test automaton, in general, has more expressive power than a logical temporal expression. Test automaton TA_{id} using this paper is just a timed automaton. We create a desirable test automaton from test automaton components using several operators. We will describe test automaton components and these operators in Section 3.2. Our method creates a single test automaton TA_{id} for the original model M. Then, it generates a desired counter-example by performing model checking on parallel composition of M and TA_{id}.

3 PROPOSED METHOD

3.1 Motivation Example

We use an example of the chemical plant system (described in more detail in Section 4) as our motivation example.

Please note the model has been obtained from description of the chemical plant system, which has some bugs in its design.

We use the following expression as property to check.

$$(In.C2 \wedge T2\_out == 0) - - > T1.Out1) \qquad (2)$$

This expression uses an operation "lead to" $-- >$ and the expression is equivalent to

$$\mathbb{AG}((In.C2 \wedge T2\_out == 0) imply \mathbb{AF} T1.Out1), \qquad (3)$$

which means that "when the proposition $(In.C2 \wedge T2\_out == 0)$ is satisfied, sometimes surely to reach the state that proposition $T1.Out1$ is true."

When this property does not hold, a counter-example with more detailed information as a foothold for subsequent failure diagnosis. In this example, as a counter-example, we would expect the path which cannot reach to the state $T1.Out1$ starting from state $In.C2 \wedge T2\_out == 0$.

The counter-examples, however, which are really generated just specifies an initial state only. Clearly it is insufficient for use in fault diagnosis.

In order to solve the above problem, we propose a method using test automata in the next sub section 3.2.

3.2 Method for Generating Test Automata

Our proposed method obtains a useful counter-example from parallel composition of the original model which represents the behavior of the target system and test automata which also represents the expected behavior of counter-example that a user roughly expects.

In more detail, the method creates a rough sketch of counter-example that users expect as a test automaton. The test automaton is obtained by synthesizing "test automaton components." Then, using parallel composition of the original model and the test automaton synthesized, we obtain desired counter-examples.

Procedures are given in the follows steps.

Let M be an original automaton (or automata) to be verified.

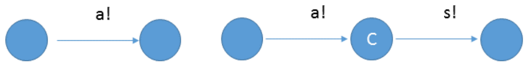


Figure 2: Modification of Event a!

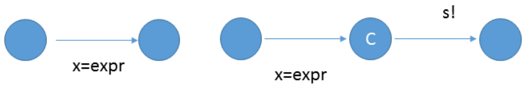


Figure 3: Modification of Update

1. A user considers an outline of the counter-example s/he wants to obtain. S/he predicates the final counter-example from the system documents and faults report documents. S/he also has to know some domain specific knowledge.
2. S/he compose a single test automaton T by synthesizing test automaton components. Each of test automaton components is weaved using rename and fusion operators.
3. S/he performs model checking on a parallel composition of M and T with property P which means "$M||T$ will not reach the final state." If P does not hold, we obtain a counter-example which s/he wants.

We will describe the detail of test automaton components and operators for weaving.

The original model M should be preferably as little as possible changed, however, in order to communicate to the test automaton, the following modification has to be performed. Please note that these modification is can be performed automatically.

If M has some event $a!$, $a?$ or some variable x to check in test automata, then following modification are performed.

1. If M has a transition with event $a!$ $(a?)$, then add a transition which has a synchronization signal $s!$ to a test automaton. Figure 2 shows the modification. Here the location with "C" denotes a committed location. In a committed location, the control does not stay. Thus the event $s!$ is performed immediately after the event $a!$
2. If M has a transition with a variable update $x = exp$ and variable x should refer in a test automaton, then add a transition which has a synchronization signal $s!$ to a test automaton. Figure 3 shows the modification. Also variable x should be declared as a global variable.

3.3 Test Automaton Components and Operators for Weaving

Figure 4 shows general form of a test automaton component

$$TA_{id}(L_{in}, a, guard, update, reset, L_{out}, I_s, I_t)$$

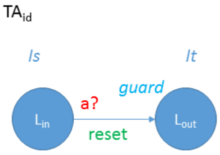


Figure 4: Test Automaton Component

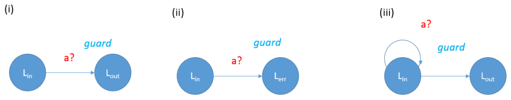


Figure 5: Typical Test Automaton Components

, where L_{in} and L_{out} are an enter location and an exit location, respectively; $a, guard, update, reset, I_s$, and I_t are an event, a guard, an update, a clock set to reset, invariants for an enter and an exit locations, respectively.

Figure 5 shows typical test automaton components.

First we explain the simple component. Figure 5 (i) is a test automaton component to receive signal a.

It can be presented in

$$TA_1(L_{in}, a, true, \emptyset, \emptyset, L_{out}, \emptyset, \emptyset).$$

Using (i) type test automaton components, we can compose a test automaton which can receive consecutive signals.

For example, let us consider test automaton components in Figure 6.

$$TA_1 = (L_{in}, a1?, x > 0, \emptyset, \emptyset, L_{out}, \emptyset, \emptyset)$$

$$TA_2 = (L_{in}, a2?, true, \emptyset, \emptyset, L_{out}, \emptyset, \emptyset)$$

Rename operator ($TA@label\backslash label2$) changes the label of a test automaton component.

For example $TA_1@L_s\backslash L_{out}$ and $TA_2@L_s\backslash L_{in}$ are shown in Figure 7.

Fusion operator $(+)$ is a binary operator which mrges locations in both terms.

For example, $(TA_1@L_s\backslash L_{out}) + (TA_2@L_s\backslash L_{in})$ is shown in Figure 8.

Figure 5 (ii) forces to error state. It used when the conditions required for the counter-example are not met.

Figure 5 (iii) stays in a location. It cannot reach location L_{out} while the event happens with a condition guard.

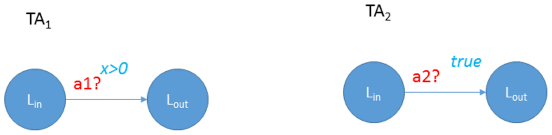


Figure 6: Test Automaton Components 1 and 2

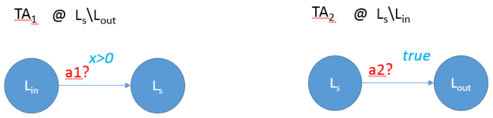


Figure 7: Renamed Test Automaton

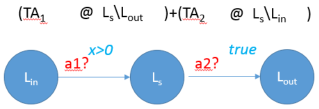


Figure 8: Fusion Test Automaton

4 CASE STUDY

We have applied the method proposed in Section 3 to verification of a chemical plant system of reported in IPA [10] as a case-study for a “post-verification of the system.”

Figure 9 shows a schematic diagram of the system. Functionality of the system is as follows:

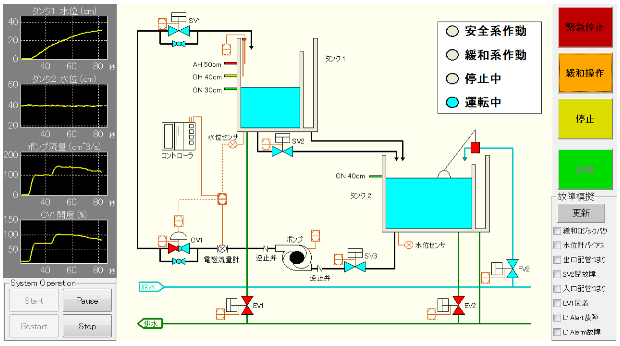


Figure 9: A schematic view of the system from “Fault diagnosis method for the large-scale and complex the embedded system (in Japanese)”[10]

- When the water level is more than an alert level (40cm), the system opens discharge valve for 5 seconds for overflow prevention. Next 15 seconds (5 seconds included) does not accept a new open instruction.

- It also performs the same discharge operation as an instruction of the operator.

- An instruction of the operator always takes over precedence over the other instructions. System accepts it even when the prohibited interval of 15 seconds. It has priority even for the past instructions made of the operator.

This system has a failure that the system cannot drain water and does not accept instruction issued from the operator even when the water level has exceeded the alert level and it has past the prohibited interval. We have performed modeling

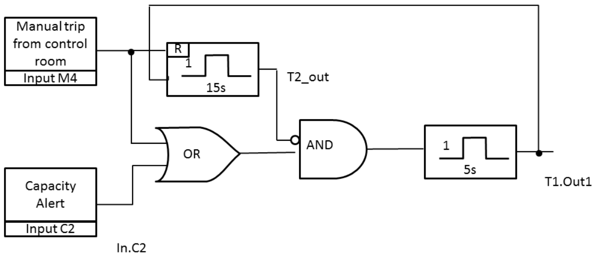


Figure 10: Control Diagram of The Chemical Plant System

using UPPAAL in order to verify this failure. Figure 10 show model diagrams showing the control system.

Signal $M4$ represents a manual input from the operator. Signal $C2$ represents a signal which shows that the water level exceeds the alert level. For both signals, 1 indicates an input of a signal for staring discharge. Also 0 shows no input. Signal $T1$ is representing the state where the system is finally in draining water. Signal $T2$ represents a model of a timer which means when the manual operation is prohibited an overall transmission of sensor input for 15 seconds immediately after the start forced discharge. In is a model indicating that changing the state of the system by an input from the $M4$ and $C2$.

In usual, formal methods decide whether properties which we expect to be hold really hold on the model. In this case, we consider the negation of the failure that has occurred in the system as a property to check using the model checking. In other words, the property to check in this case is "whether the system sometimes becomes discharge mode when draining water is not suppressed and the water level sensor is in a state of warning level."

We check Expression 2 on the model.

A counter-example, however, is a very simple, we cannot understand why the property is not satisfied. Therefore, it cannot be used for fault diagnosis of the system.

Our method creates a test automaton, thereby to generate a human expected counter-example.

Figure 11 describes the test automaton that has produced by the method. The test automaton is derived by coarse behavior series of counter-examples in previous work [11]. Let TA_1 be a test automaton component obtained from type (i) in Figure 5. $TA_1 = TA_{id}(start, input?, C2 == 1\&\&M4 == 0, isT1end = 1, \emptyset, end, \emptyset, \emptyset)$

Similarly, TA_2, TA_3, and TA_4 can be obtained from test automaton components in Figure 5.

Figure 11 summarizes these test automaton components.

A test automaton obtained by the following expression is the final automaton we want.

$$(TA_1@L_s\backslash end) + (TA_2@L_s\backslash start, L_t\backslash end)$$

$$+(TA_3@L_s\backslash start) + (TA_4@L_t\backslash start)$$

Figure 12 shows the obtained test automaton.

By combining a model representing the system and a test automaton, detailed counter-example which can be used in the diagnosis of the failure is obtained. The generated Counter-example has 8 steps and it describes transition of variables.

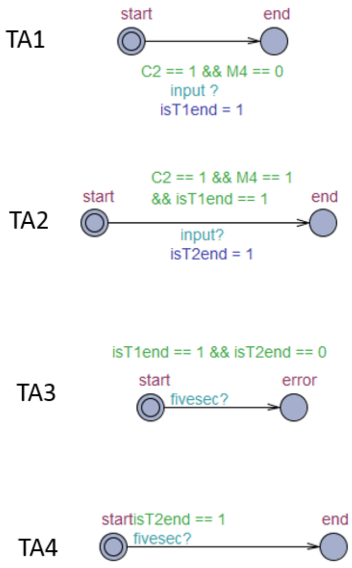


Figure 11: Test automata

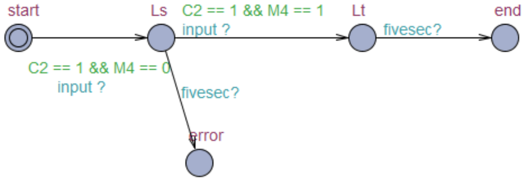


Figure 12: Test automaton Composed

5 DISCUSSION

We discuss here on a coverage and time taken in generating a counter-example.

For coverage, the counter-example the authors are considered can be covered by the proposed test automaton components for the case-study because a counter-example is single path. They would be cover for other cases.

For the time required for counter example generation, the generation was performed in less than 1 second in this example. This method uses a test automaton as a guideline of creating counter-example. Therefore, it is assumed that scalability of the time falls within the appropriate range even counter-example and scale of the time automaton becomes bigger. In addition, the creation of test automaton in this time was by the authors. Therefore, it should be evaluated that whether a novice for fault diagnosis can create a correct automaton and how long it takes to create an automaton. This is a simple method of assembling the test automaton from the parts that are provided. However, it is required knowledge of model checking to create test automaton. Therefore, target of the method is a user with the prior knowledge of model checking. This is referred to as one of the future challenges.

6 CONCLUSION

This paper proposed a method for generating a counter-example of as user expected using test automata. We also applied the technique to a case study of the chemical plant system, in order to verify its validity.

We do not propose a method generating a counter-example automatically. This is an approach that combines model checking and test automata to generate counter-example. Therefore, a method automatically generating counter-examples is needed in future. The follows become problems when we generate counter-example automatically. First, it cannot fully search the model space when it is an infinite state transition system. Then, even in the finite state transition system, it might suffer from sate-explosion problems.

Therefore, model abstraction technique to properly reduce the number of states of the model for each property [13, 14], attracts attentions. Bounded Model Checking (BMC) [12] would one of another promising approaches. Bounded Model Checking is a technique of model checking that prevents state explosion by limiting the search range of the finite state space. When BMC finds violation on a finite state space, counter-examples are generated as a finite length. In general, counter-examples have infinite length. However, users usually want finite counter-examples and therefore, they think a counter-example generated by BMC is enough for their purpose. BMC is considered to be a promising approach for generating a counter-example suitable length within a reasonable time.

ACKNOWLEDGMENTS

This work is partially being conducted as Grant-in-Aid for Scientific Research S (25220003) and C (16K00094). Funding from Mitsubishi Electric Corp. is gratefully acknowledged. Finally the authors wish to acknowledge Dr. Nakajima, Professor of NII for discussion.

REFERENCES

[1] C. Newcombe, T. Rath, F. Zhang, B. Munteanu, M. Brooker, and M. Deardeuff : “How Amazon Web Services Uses Formal Methods,” Communications of the ACM, Vol.58, No.4, pp.66-73, (2015).

[2] E. M. Clarke, O. Grumberg, and D. A. Peled: “Model Checking,” MIT Press, (2000).

[3] K. Okano, T. Nagaoka, T. Tanaka, T. Sekizawa, S. Kusumoto: “Parallel Multiple Counter-Examples Guided Abstraction Loop —Applying to Timed Automaton—,” International Journal of Informatics Society, Vol. 8, No.3, to appear, (2016).

[4] R. Alur: “Timed Automata,” In Proceedings of 11th International Conference of Computer Aided Verification, (CAV ’99), Vol.1633, pp.8-22, (1999).

[5] J. Bengtsson and W. Yi: “Timed automata: Semantics, algorithms and tools,” in Lecture Notes on Concurrency and Petri Nets, Vol.3098, pp.87-124, (2004).

[6] G. Behrmann, A. David, and K.G. Larsen: “A Tutorial on Uppaal,” Formal Methods for the Design of Real-Time Systems, Vol.3185, pp.200-236, (2004).

[7] T. Sekizawa, K. Okano, A. Ogawa, and S. Kusumoto: “Verification of a Control Program for a Line Tracing Robot using UPPAAL Considering General Aspects,” International Journal of Informatics Society, Vol.6, No.2, pp.79-87, (2014).

[8] F. Weitl and S. Nakajima: “Incremental Construction of counterexamples in Model Checking Web Documents,” in Proceedings of the 6th International Workshop on Automated Specification and Verification of Web Systems, EPiC Series, Vol.18, pp.61-75, (2010).

[9] E. Clarke, O. Grumberg, S. Jha, Y. Lu, and H. Veith: “Counterexample-guided abstraction refinement for symbolic model checking,” Journal of the ACM, Vol.50, Issue 5, pp.752-794, (2003).

[10] IPA: “Fault diagnosis method for the large-scale and complex the embedded system (in Japanese),” https://www.ipa.go.jp/files/000045158.pdf, pp.68-78, (2015) ⟨accessed June 05, 2016⟩.

[11] K. Okano and J. Kitamiti: “Fault diagnosis method for the large-scale and complex the embedded system (in Japanese),” Software symposium 2015, http://sea.jp/ss2015/paper/ss2015_C1-2(2).pdf, (2015) ⟨accessed July 28, 2016⟩.

[12] A. Biere, A. Cimatti, E. Clarke, and Y. Zhu: “Symbolic model checking without BDDs,” In Proceedings of 5th International Conference of Tools and Algorithms for Construction and Analysis of Systems (TACAS ’99), pp.193-207, (1999).

[13] E. Clarke, O. Grumberg, S. Jha, Y. Lu, and V. Helmut: “Counterexample-guided Abstraction Refinement,” In Proceedings of 12th International Conference of Computer Aided Verification (CAV ’00), Vol.1855, pp.154-169, (2000).

[14] E. Clarke, A, Gupta, J. Kukula, and O. Strichman: “SAT based Abstraction-Refinement using ILP and Machine Learning Techniques,” In Proceedings of 14th International Conference of Computer Aided Verification (CAV ’00), Vol.2404, pp.695-709, (2002).

[15] R. Alur: “Techniques for automatic verification of real-time systems,” Ph.D. dissertation, Stanford University, (1991).

[16] L. Ageto, P. Bouyer, A. Burgueño, and K. G. Larsen: “The Power of Reachability Testing for Timed Automata,” Journal of the Theoretical Computer Science, Vol.300, Issue 1-3, pp.411-475, (2003).

[17] B. Bordbar, R. Anane, and K. Okano: “An Evaluation Mechanism for QoS Management in Wireless Systems,” International Workshop on Performance Modelling in Wired, Wireless, Mobile Networking and Computing 2005, In Proceedings of International Conference on Parallel and Distributes System (ICPADS 2005), Vol.2, pp.150-154, (2005).

[18] B. Bordbar and K. Okano: “Verification of Timeliness QoS Properties in Multimedia Systems,” In proceedings of 5th International Conference on Formal Engineering Method (ICFEM 2003), LNCS Vol.2885, pp.523-540, (2003).

Session 8:
Multimedia Systems

(Chair: Katsuhiko Kaji)

Electronic Smell Picture Book for Children Using Olfactory Display

Shohei Horiguchi[†], Sayaka Matsumoto[†], Hiroshi Shigeno[‡], and Ken-ichi Okada[‡]

[†]Graduate School of Science and Technology, Keio University, Japan
[‡]Faculty of Science and Technology, Keio University, Japan
{shohei, matsumoto, shigeno, okada}@mos.ics.keio.ac.jp

Abstract - Human get a lot of information through the five senses. Information that can be gained using smell, one of the five senses, has a deep connection with memory and affect; therefore, presentation of scents with visual information has an effect that enhances realistic sensations. In addition, olfaction is also used to detect dangers such as rotten food or gas leaks. For these reasons, sense of smell is a very important in daily life. It is important to use smell many times because it is formed primarily in childhood. However, there is little opportunity to use it compared with eyesight and hearing. In this paper, we developed the electronic picture book with which children can experience scents repeatedly. It is also possible to present stimulations of other senses through wind, vibrations and sounds. In addition, it can read to children automatically with recorded voice and we can change effects presented for children. This study is expected to enrich sensibility and support growth of children.

Keywords: Electronic picture book, Children, Olfactory display, Human computer interaction.

1 INTRODUCTION

Humans have five senses (eyesight, hearing, smell, touch, and taste), all of which are very important and are processed by the sensory system. In five senses, olfactory stimulation has a strong, direct connection to the limbic system, which controls emotion and memory. Therefore, videos accompanied by scents increase viewer concentration and retention of the video [1]. Olfaction is also used to detect dangers such as rotten food or gas leaks. For these reasons, olfaction is an essential sense in daily life. It is formed mostly in childhood owing to feeling scents many times. Typically, there are many educational toys for children concentrating on sight and hearing. On the other hand, there is little opportunity to experience smell. Hence, it is important to use sense of smell many times in childhood. Furthermore, it is also important for growth of children to be read picture books by parents. It results in development of linguistic competence and imagination. In our study, we focused on the experiences of scents and a picture book, and developed the electronic picture book that infants can experience odors repeatedly. This book consists of the olfactory display which can present scent by pulse ejection and a tablet device. It can provide stimulations of other senses through wind, vibrations and sounds. Moreover, children can be read automatically by recorded voice. We conducted two experiments for which were in order to evaluate the usability of our electronic book. In Section 2, we introduce related works about the plays for children and an electronic picture book. In Section 3, we propose our electronic picture book using olfactory display. Section 4 explains the book's implementation, and Section 5 assesses its usability. Finally, in Section 6, we present our conclusions.

2 RELATED WORK

Humans recognize an environment around them owing to using information that can be gained using five senses. These senses are formed primarily in childhood through experiences a lot of new things for them. Experiences using olfaction, touch and taste are important to enhance sensibilities of them [2]. In particular, children get many experiences thorough plays. Typically, there are a lot of plays giving stimulations to children's senses. As examples of such plays, there are building blocks, clay, simple instruments, paintings, picture books and so on. Building blocks and clay enhance a child's touch. The sound made by playing an instrument stimulates children's ears. It is said that music is essential for children's growth [3]. "I'm toy music station" sold as one of the intellectual toys contains nine instruments [4]. When children draw pictures, they use many colors. Humans have their likes and dislikes in colors naturally. Though color sense is an innate instinct, it is affected by the environment in which humans have grown up [5]. It is known that humans are more sensitive to familiar colors than unknown colors. The more we have come in touch with colors in childhood, the more we can recognize a variety of color. Therefore, there are many toys for children contain a lot of colors. Additionally, picture books also contain some colors, and stimulate a child's color sense. It is said that picture books develop a child's imagination and language ability [6]. There are many picture books recommended for a variety age, hence they play an important role for growth of a child [7]. On the other hand, the method of using picture books is not only reading by oneself but storytelling by parents. Storytelling by parents encourages not only the communication between parent and a child, but also the development of a child [8] [9]. Keiko, D et al. reported that children had been given storytelling by parents for a year got better score in the confabulation test than children of control group [10]. In recent years, electronic picture books come into vogue. They are composed of not only pictures but sounds or animations. They also have a function of automatic storytelling. It is reported that children showed more interest in electronic picture book than normal picture book when they use automatic storytelling [11]. There are these toys stimulating eyesight, hearing and touch without olfaction. Additionally, it is known that children have the least number of opportunities to use the sense of smell in five senses [12]. Therefore, children need to feel more odors. There is a smell book as one of the few toys that a child can experiences olfactory stimulations [13]. This book presents scent by breaking capsule including perfume. For this reason, children can never feel the odor after a capsule broken.

3 ELECTRONIC PICTURE BOOK USING SCENTS

It is important that human experiences many things about the five senses in the infancy when they are said to be the most sensitive. Children use sight, hearing and touch in play positively. However, there are fewer toys using sense of smell than toys using other senses. Therefore, it is said that children have the fewest opportunities to use olfaction in the five senses in daily life. On the other hand, it is hoped that the sensitivity of children is raised by smelling a fragrance. In this paper, we focus on a picture book which is familiar with children, and develop the application of electronic smell picture book for children using olfactory display. It is only once that children can experience a fragrance in the existing smell book because an odor is shown by breaking a capsule enclosing it on the illustration. Our electronic smell picture book can present an odor repeatedly by using the olfactory display which has technique of scent presentation in short duration. In addition, we use a tablet device for our book in order to make operation of it easy for children. Olfactory display is worn around the neck. Therefore, our application can smell a scent while operating the tablet. We assume that the target age of our book is 2-6 years old. The 2-3 years old children are often read by parents, contrary to this, the 4-6 years old children often read a book themselves. There are two methods of reading we prepare in order to entertain both younger and older children. One mode is that children use the automatic storytelling with recorded voice of parent to read. It is said that the voice of parents makes a child's mental condition comfortable. Another one is that they read themselves. These modes can decrease a burden to parents. Moreover, our application can provide stimulations of other senses through picture, wind, vibrations and sounds. It can stimulate four senses, olfaction, sight, hearing and touch, in total. In spite of reading, it can edit effects of the book. To evaluate this application, we conduct two experiments about reading part and editing part.

4 APPLICATION

4.1 Olfactory display

The olfactory display we used in our electronic smell picture book is "Fragrance of Jet for Mobile." It is worn around the neck as shown in Figure 1, and only the user can smell the emitted odors. Figure 2 and 3 show the plane and the slide views of it, respectively. This device adopts the thermal method used in ink-jet printers to emit odors. It has an ejection head, storing one large tank and three small tanks. Each tank stores an odorant, thus four kinds of odors can be contained. Odorants are emitted in picoliter (pl) quantities from small holes in the head on the wind by a fan. The average ejection quantity from the large tank's hole and each of the small tank's hole are 7.3 pl and 4.7 pl respectively. There are 255 holes in the head connected to the large tank and 127 minimum holes in the head connected to the small ones. Since this device can emit odorants from multiple holes at one time, the ejection intensity is controlled by the number of holes, of which the range is 0-255 in large tank or 0-127 in small tanks. In addition, the time of ejection can be controlled by 667 microseconds, the unit time. Hence, shown in Figure 4, the total ejection quantity for one pulse is determined by the ejection time and the ejection quantity per unit time, which is determined by the average ejection quantity of using tank (7.3 pl or 4.7 pl) and the number of holes. The device is capable of this pulse ejection, allowing it to emit odors in a manner to avoid sensory adaptation [14]. Since pulse ejection can also prevent odors from lingering in the air, the device can also emit different odors for each page and switch odors quickly. It is therefore suitable for presenting odors with picture book.

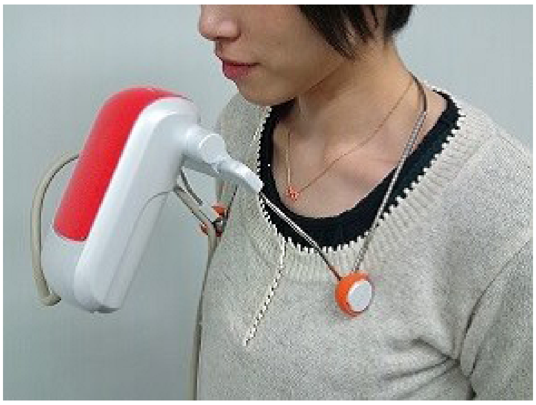


Figure 1: Wearable olfactory display

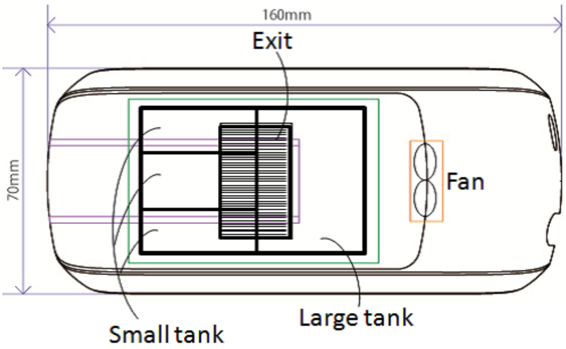


Figure 2: Overhead view of olfactory display

In this paper, we use the wearable olfactory display in order that a child can smell odors while operating a tablet device. Almost children can not stand still. Therefore, if a normal free-standing, rather than wearable, olfactory display is used by a child, position of nose would move to the area where he or she can not smell scents.

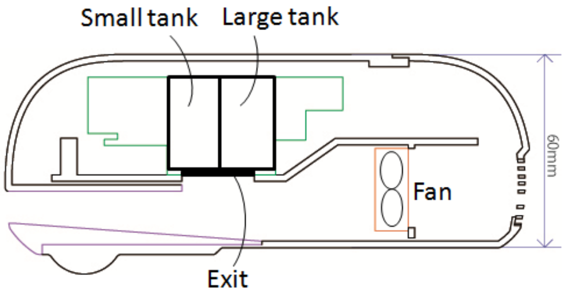


Figure 3: Sectional view of olfactory display

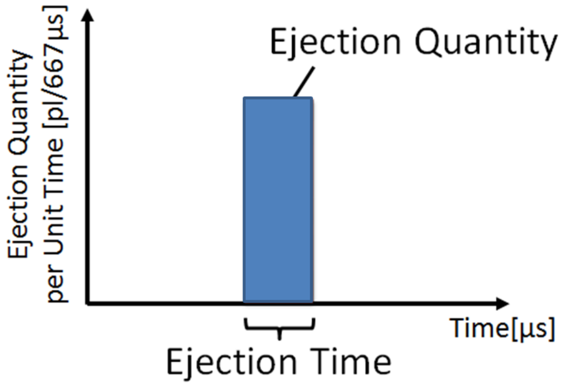


Figure 4: Pulse ejection

4.2 Electronic picture book

We suppose that our application is used by a child who is 2-6 years old. Therefore, we developed the application of an electronic smell picture book which such children can use easily. It is comprised of the olfactory display, a tablet device (Nexus9 [15]), and a computer which controls all devices. To send information of scent between a tablet and a computer, we use the Bluetooth communication system. Figure 5 shows a conceptual scheme of application. The application is implemented by two programs. One program is to send information about scent presentation which contains kinds of odor and ejection time from a tablet to a computer. Another one is that a computer orders the olfactory display to present scent.

4.2.1 Method of presentation

Our book can present four effects: scents, sounds, wind, and vibrations. Scents and wind are presented by using olfactory display. A tablet presents sounds and vibrations. In this paper, we made an original picture book, rather than published one, in order that children can experience four effects. In our picture book, we prepared five kinds of pages presenting stimulations: we named them odor page, sound page, wind page, vibration page, and the mixed stimuli page. There are two presentation methods of odors in the odor page. They are showed in Figure 6. First method is presenting an odor repeatedly by using pulse ejection while odor page is displayed in order that the child can feel it unconsciously and concentrate on reading. Another method is a scent is emphasized momentarily when the child taps display of a tablet. The example of odor page is showed in Figure 7. While the child see this page, scent of banana is presented to him or her. In addition, when the child taps the mark of hand on banana's picture, scent of banana is emphasized instantly. In the wind page, there are similar two presentation methods like in the odor page. They are showed in Figure 6. The wind blows continuously from the olfactory display while the wind page is displayed. Moreover, when the child taps mark of hand in the page, strong wind blows momentarily. In vibration page, there are similar two presentation methods too. They are showed in Figure 6. The tablet vibrates continuously while the vibration page is displayed. When the user taps illustration of hand in the page, other kind of vibration is presented. Contrary to these pages, the sound page has only one method of presentation. When the child taps the mark of hand, a tablet produces the sound.

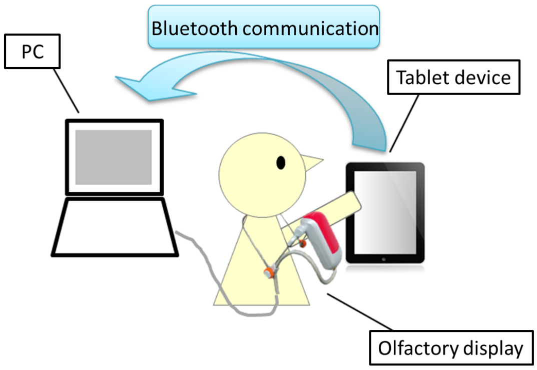


Figure 5: Conceptual scheme of application

4.2.2 Three modes of the application

The operating flow of this application is showed in Figure 8. There are three modes: the mode of reading together, the mode of reading alone, and the edit mode in our electronic book. The mode of reading together is used to read to the child by parent. The mode of reading alone is used when the child reads alone. In the edit mode, parents can record their voice which used by the automatic storytelling, and add effect to book. When the application starts, the user needs to choose among three modes at first.

- The mode of reading together

At first, in the mode of reading together, the page which explains how to read the electronic smell book is showed in Figure 9. In this page, the child can experience four effects: odor, sound, vibration, and wind. For example, if the user touches the mark of hand on the note, he or she could hear the sound effect. After feeling four effects, the user can go next page by sliding the screen with a finger from right to left. Next page is the cover page of picture book. After that the user can read a book by sliding the screen. If the child wants to go back previous page, he or she need to slide the screen from left to right. There are some marks of hand in various place of picture book. The user can get some effects by tapping the

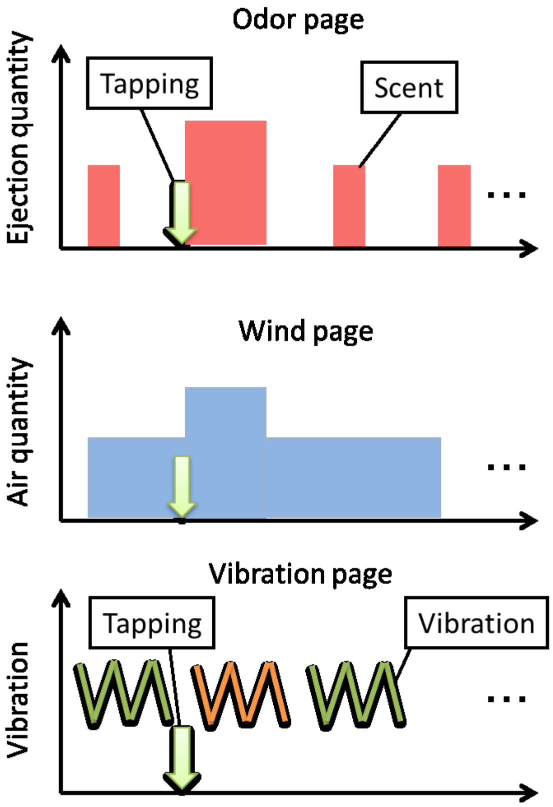


Figure 6: Presentation methods of odor, wind, and vibration

marks. When tapping the mark, a tablet vibrates in order to teach the user the timing of presentation. When "finish" button is pressed in the last page, the screen return to the page of choosing modes.

- The mode of reading alone

In the mode of reading alone, the child can use automatic storytelling with recorded voice in order that he or she can read electronic smell book alone. The user can choose whether using automatic storytelling or not. First screen of this mode have "not use storytelling" button and "use storytelling" button. When "not use storytelling" button is pressed, explanation of how to play is displayed. When "use storytelling" button is pressed, new two buttons: "lady's voice" button and "parents' voice" button are appeared. Lady's voice is prepared at first. Parents' voice means recorded voice by parents. The user can choose which voice using in storytelling. After the user choose the voice, the explanation page of how to play is displayed. The automatic storytelling begins whenever the user slides page of picture book.

- The edit mode

We prepare the edit mode for parents. Parents can record their voice for storytelling, and conform effects to new recorded voice. Screen of the edit mode is shown in Figure 10. Target of edit page is displayed at the lower left of the screen. There is a page number at the upper part of the screen. In both sides of the page number, there are "Previous page" button

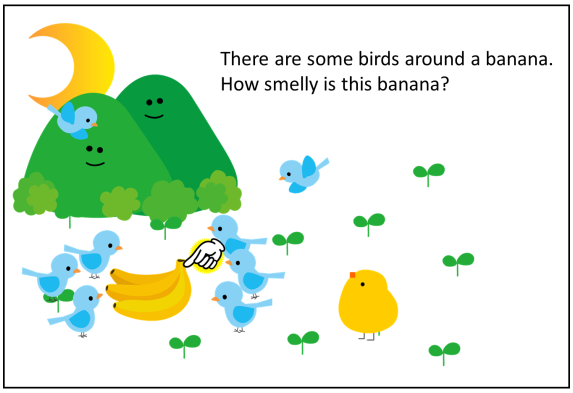

There are some birds around a banana.
How smelly is this banana?

Figure 7: Odor page of picture book

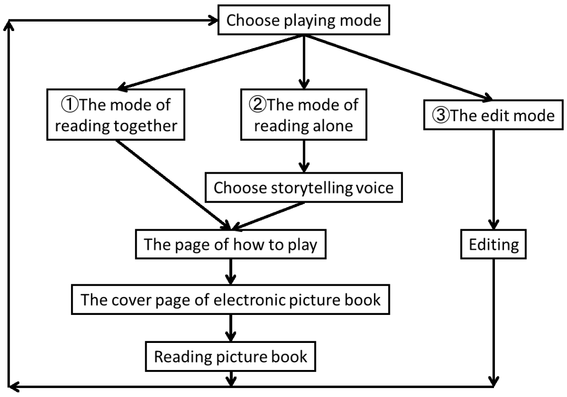
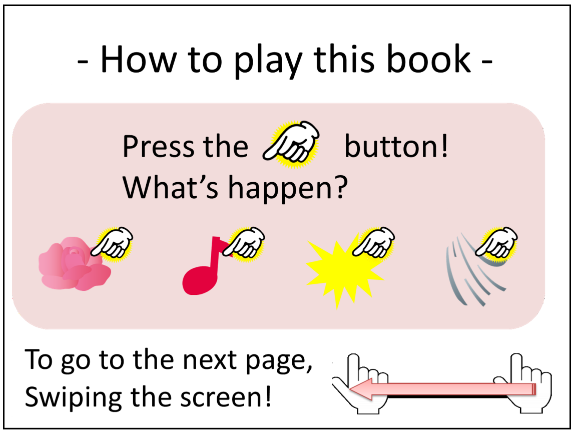

Choose playing mode
①The mode of reading together
②The mode of reading alone
③The edit mode
Choose storytelling voice
The page of how to play
Editing
The cover page of electronic picture book
Reading picture book

Figure 8: Operating flow of application

and "Next page" button to change the editing page. There are "Start recording" button, red "Stop" button, "Play back" button, and green "Stop" button under the page number. When "Start recording" button is pressed, recording is started. To stop the recording, parent needs to press red "Stop" button. The user can listen to their recorded voice by pressing "Play back" button. If parents want to add sound effect to storytelling, they need to push the sound buttons under "Start recording" button during recording the voice. In addition, if they want to add other effect, they need to use the area of "Edit all time effect" and the area of "Edit momentarily effect". The area of "Edit all time effect" is used to add the effect which is presented while reading target page. They choose the kind of effect from pull-down menu. If odor or wind is chosen, it is necessary to select the intensity of effect. In contrast, if vibration is chosen, the user needs to decide the kind of it. Next, when "Fix" button is pressed, effect the user chooses is fixed. The area of "Edit momentarily effect" is used to add the effect which is emphasized temporarily. If parents want to emphasize the effect, they need to press "Add" button while playing back the recoded voice. If the user wants to delete the fixed effect, he or she could delete it by tapping "Delete" button under the "Add" button. This application can use four odors(banana, apple, rose, pineapple), two levels of wind, three vibrations, and four sound as the effect.



- How to play this book -

Press the button!
What's happen?

To go to the next page,
Swiping the screen!

Figure 9: Screen of the page introducing how to play

5 EVALUATION

We conducted two experiments for evaluation of the application. First experiment evaluated whether the mode of reading alone could be used by children. Second experiment assessed whether adults could edit the effect by using the edit mode.

5.1 Experiment for picture book

5.1.1 Experiment outline 1

We conducted the experiment in order to investigate whether children could use our electronic smell book. Subjects were 12 children (5 boys and 7 girls) who went to Hiyoshi Benesse nursery school. Their ages were from 5 to 6 (mean: 5.75, SD: 0.43). In this experiment, we used the original electronic smell picture book which has eight pages: four odor pages (banana, apple, rose, pineapple), a sound page, a wind page, a vibration page, and only picture page. We ordered subjects to read our book by using the automatic storytelling. We measured the number of tapping and the playing time of each page during experiment. After finishing experiment, we asked a question: “ Which page did you enjoy better? ” as a questionnaire. Subjects could choose multiple pages in this question. When subjects used our application, they wore olfactory display, and took a seat. If it was difficult for them to wear olfactory display, we set it on the table. An experimental environment is showed in Figure 11. The experiment for each subject took about ten minutes.

5.1.2 Result of experiment for picture book

First, we considered about the number of tapping. The averages and standard deviations for the number of tapping in each odor page showed in Table 1. We analyzed whether or not the average was different based on the one-way analysis of variance. Then, there was not a significant difference ($p = 0.78 > 0.05$). Therefore, we calculated the tap number's average of odor page from four odor pages. Moreover, we showed the averages and standard deviations for the number of tapping in each effect page in Table 2. We analyzed the result in Table 2 by using the one-way analysis of variance. As a result, there was not a significant difference ($p = 0.41 > 0.05$). Hence, subjects could feel four effects on the same level by using our application.

Secondly, we considered about the playing time. The averages and standard deviations for the playing time in each odor page showed in Table 3. We analyzed whether or not the average of the playing time was different based on the one-way analysis of variance. Then, there was not a significant difference ($p = 0.78 > 0.05$). Therefore, we calculated the playing time's average of odor page from four odor pages. In addition, we showed the averages and standard deviations for the playing time in each effect page in Table 4. We analyzed the result in Table 4 by using the one-way analysis of variance. As a result, there was a significant difference ($p = 0.0007 < 0.05$). Hence, we used Tukey's test as multiple comparison. The result of it indicated that there is a significant difference between odor page and only picture page ($p < 0.05$). Accordingly, we found that children could enjoy odor pages better than only picture page.

Thirdly, we considered about the result of questionnaire which is showed in Figure 12. We found that nine out of twelve subjects chose the odor page as an interesting page. For this reason, children tend to be interested in the odor page. Meanwhile, all subjects could have read our book and felt various effects by tapping screen. From the above results, we confirmed that the electronic odor picture book, we developed, could be used by children.

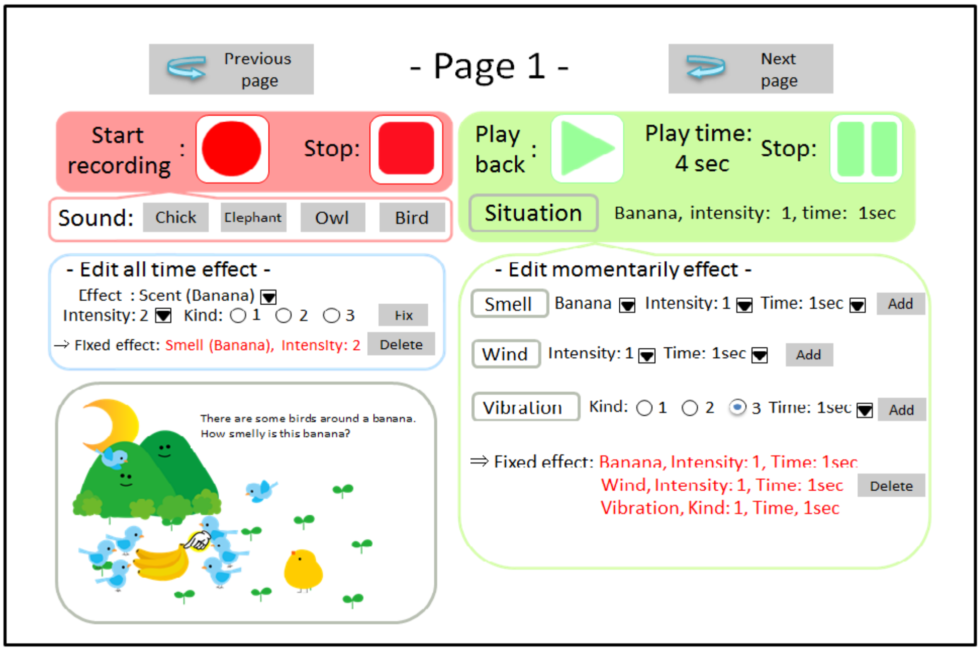
5.2 Experiment for the edit mode

5.2.1 Experimental outline 2

We conducted the experiment in order to investigate whether adults could use the edit mode of our application. 15 subjects (10 men and 5 women) participated in this experiment. The participants were graduate and undergraduate students who were majoring in information engineering. At first, subjects practiced editing the effect of the picture book after lectured how to use the application by us. Next, we gave each subject six tasks: recording voice, adding wind, adding sound, adding vibration, adding scent of banana, and adding scent of rose. However, we also gave the rule in this experience. It is that subjects could try only one time per one task. We evaluated the usability of the edit mode by a percentage of correct edit. After finished experiment, we asked some questions about usability of the application as a questionnaire.

5.2.2 Result of experiment for the edit mode

A percentage of correct edit was 97.8 % over 95 %. There were only two false edit: forgetting to press “fix” button and miss choice of effect. These mistakes could be caused by the rule we gave to subjects. In fact, the user can edit many times, in consequence, we thought that the user could decrease the miss edit. Next, we considered about the result of a questionnaire showed in Table 5. We used the five rated evaluation (1: bad - 5: good). The averages of score are showed in Table 5. The results indicated that our application had a good usability owing to both scores over 4. From the above results, we



Banana, intensity: 1, time: 1sec

Effect : Scent (Banana)
Intensity: 2 Kind: ○ 1 ○ 2 ○ 3
→ Fixed effect: Smell (Banana), Intensity: 2

Smell Banana Intensity: 1 Time: 1sec
Wind Intensity: 1 Time: 1sec
Vibration Kind: ○ 1 ○ 2 ● 3 Time: 1sec
⇒ Fixed effect: Banana, Intensity: 1, Time: 1sec
Wind, Intensity: 1, Time: 1sec
Vibration, Kind: 1, Time, 1sec

There are some birds around a banana.
How smelly is this banana?

Figure 10: Screen of the edit mode

Table 1: Result of tap number in four odor pages

Odor	Banana	Apple	Rose	Pineapple
Tap number	1.83 ± 1.85	1.67 ± 0.99	1.42 ± 0.79	2.00 ± 1.76

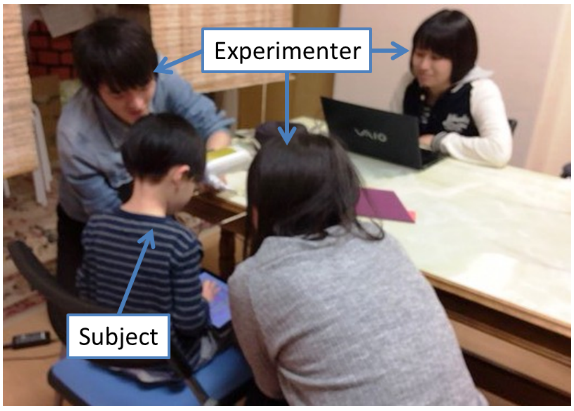


Figure 11: Experimental Environment

confirmed that the edit mode of our electronic book could be used by adults.

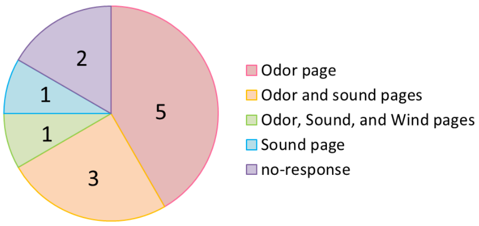


Figure 12: Result of questionnaire for children

6 CONCLUSION

It is important for personal development to get a lot of information thorough the five senses in childhood in order to enhance sensitivity. Though there are many toys stimulating sight, hearing, and touch, there are a few things for olfaction. Children need to use the sense of smell many times. On the other hand, it is said that a storytelling is also important for a child's growth. In our study, we focused on the child's olfaction and storytelling, and developed the application of electronic smell picture book using olfactory display for children. The existing smell book presents a fragrance only one time, contrary to this, our application could eject a scent many times owing to the olfactory display. This electronic smell picture book can also stimulate other senses by presenting the wind, the vibration, and the sound. Moreover, it has a function of automatic storytelling using the recorded parent's voice. Therefore, a child can read this book alone. Parent can also edit the effect of this book. We conducted the two experiments for children and adults in order to evaluate the usability of this application. As a result, subjects of children could operate this application, and feel some stimulations. We found that children tended to be interested in

Table 2: Result of tap number in four effect pages

Effect	Odor	Sound	Wind	Vibration
Tap number	1.73 ± 1.08	1.33 ± 0.65	1.67 ± 0.89	2.00 ± 1.13

Table 3: Result of playing time in four odor pages [sec]

Odor	Banana	Apple	Rose	Pineapple
Playing Time	12.9 ± 9.50	10.3 ± 5.60	10.8 ± 6.75	12.5 ± 6.78

Table 4: Result of playing time in five effect pages [sec]

Effect	Odor	Sound	Wind	Vibration	Only picture
Playing time	11.6 ± 5.81	7.32 ± 4.62	10.4 ± 3.82	9.50 ± 3.21	4.57 ± 1.58

Table 5: Score of questionnaire

	Score
Controllability	4.13 ± 0.52
Comprehensibility	4.60 ± 0.51

the odor page of book. The result of experiment for adults indicated that they could edit the effect of the book correctly. From the above results, we confirmed that our application has a good usability. We hope that many children's sensitivity will be enhanced by using our electronic smell picture book.

7 ACKNOWLEDGMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research (C), No. 26330229 (2014) from the Ministry of Education, Culture, Sport, Science and Technology Japan, Hiyoshi Benesse nursery school, and Takasago International Corporation.

REFERENCES

[1] M. Bowman, S. K. Debray, and L. L. Peterson, Reasoning about naming systems, ACM Trans. Program. Lang. Syst, Vol. 15, No. 5, pp. 795–825 (1993).

[2] Development of smell, touch and taste. (in Japanese)
http://www.jomf.or.jp/html/childcare_pdf/13.pdf

[3] K. Mochida and C. Kaneko, Effect of preschool teachers for creative music expression of children, Technical Report. University of Bunkyogakuin, Vol. 10, No. 1, pp. 37–47 (2008). (in Japanese)

[4] How to choose an intellectual toy. (in Japanese)
http://life.pintoru.com/intellectual-toy/

[5] W. A. Nagel, Observations on the color-sense of a child, Jounal of Comparative Neurology and Psychology, Vol. 16, pp. 217–230 (2004).

[6] Y. Imai, et al., A Comparative Study on Opinions about Methods of Reading Picture Books in Japan and Taiwan, Technical Report. Nara University of Education, Vol. 42, No. 1, pp. 211–223 (1993). (in Japanese)

[7] K. Sato, Child Development and Picture Books, Technical Report. University of Ehime, Vol. 51, No. 1, pp. 29–34 (2004). (in Japanese)

[8] Y. Ishikawa, Consideration of Picture Book: Production of the Zone of Proximal Development and Promotion of Cognitive Development in Children, The journal of Seigakuin University. Seigakuin University, Vol. 22, No. 1, pp. 165–179 (2004). (in Japanese)

[9] Y. Imai and J. Boi, Effect of Reading Aloud Picture Books on Young Children's Comprehension of Emotions, Technical Report. Nara University of Education, Vol. 43, No. 1, pp. 235–245 (1994). (in Japanese)

[10] K. Dono, The Effects of Experiences of Listening to Storytelling of Picture Books ('ehon-no-yomikikase') on the Development of Prosociality in Nursery School Children, Technical Report. Yasuda Woman's University, Vol. 36, pp. 81–91 (2008). (in Japanese)

[11] Report of experiment "storytelling using electronic books". (in Japanese)
http://densholab.jp/page-29/page-1085

[12] S. Yamashita, A Research about Using Olfactory Sense in Life Environment Studies, Departmental Bulletin Paper. Aichi University of education, Vol. 5, pp. 109–116 (2007). (in Japanese)

[13] Y. Kimura and C. Egawa, Otomodachi curry, Sekaibunka Corporation. (in Japanese)
http://www.sekaibunka.com/book/exec/cs/12810.html

[14] A. Kadowaki, J. Sato, Y, Bannai, and K. Okada, Presentation Technique of Scent to Avoid Olfactory Adaptation, Proc.of ICAT 2007, pp. 97–104 (2007).

[15] Nexus9 google.
https://www.google.com/nexus/9/

Examination of the Bass Pitch Detection Algorithm in the Electric Bass

Eiichi Takebuchi[†], Tomoki Kajinami[††], Ichiro Tokuhiro[†††], Haruo Hayami[†††]

[†]Doctoral Course of Informatics, Kanagawa Institute of Technology, Japan
[††]Okayama University of Science, Japan
[†††]Kanagawa Institute of Technology, Japan

Abstract - In this study, we propose a bass pitch detection algorithm for electric guitar. Detection of the bass pitch is one of the most challenging in music information processing, because the bass pitch is the lowest musical tone in musical content. Further, the root note is a musical tone of the criteria in the chord, i.e., the root note and the bass pitch are a different concepts. Electric bass is typically one octave lower than electric guitar. Therefore, we require frequency resolution for electric bass rather than electric guitar. Our research challenge was to detect electric bass from the bass pitch in real time. When the electric bass of the bass pitch can be detected, building systems for practicing electric bass are feasible. Further, our bass pitch detection algorithm is effective for fingering practice systems. In a previous study, we performed a static analysis of the bass pitch of electric bass, thus as mentioned adove, our proposed method is to detect bass pitch for electric bass in real time. In this study, we also examine the effectiveness of our proposed method through experimentation. From results of our evaluation experiments, we detected bass pitch at an average rate of 98.2%.

Keywords: Pitch Detection, Bass Pitch Detection, Electric Bass, Fast Fourier Transform, Overtone, Inharmonicity

1 INTRODUCTION

Detection of the bass pitch is one of most important challenges of music information processing. The bass pitch is the lowest musical tones produced by musical instrument. The root note is a musical tone of the criteria in the chord. Further, the root note and the bass pitch are different concepts.

The bass pitch is primarily used in chord recognition; however, electric bass is used to produce bass tones in band instruments. The electric bass is not typically used to play a chord, but rather a single note.

In this work, we developed and examined a bass pitch detection algorithm for electric bass. As a result, it was detected the bass pitch at an average rate of 98.2%. We applied work from our previous studies, i.e., applying the bass pitch detection algorithm to electric bass.

Our key research challenge was to detect the bass pitch from performances given by electric bass. The electric bass is lower than the electric guitar. Therefore, the electric bass is required frequency resolution than the electric guitar. To increase frequency resolution, we must increase the number of samples for our Fast fourier transform (FFT). In recent years, processing speeds of FFTs have improved via general-purpose graphics processing units (GPGPUs) [1][2][3]. In this current study, we endeavor to detect electric bass from content played in real time.

Therefore, we describe our proposed method and results of evaluation experiments using our method in conjunction with the electric bass.

2 REAL TIME AUDIO RECOGNITION

Composers working with desktop music (DTM) make use of a piano roll or MIDI keyboard. A piano roll is a rotating keyboard-like interface often used for creating digital music. Using a piano roll, composers make and modify music via a simple interface, thus many composers make use of piano rolls. In addition, composers might use a MIDI keyboard for implementation.

Nonetheless, the piano roll is not a suitable interface who player of other musical instrument. The MIDI keyboard was developed for the piano player, but Player of musical instrument are not necessarily good at the piano. Therefore, we introduce to achieve MIDI implemenation for the electric guitar.

Here, a MIDI pickup converts a string vibration to a MIDI signal. Reference [4][5] is mounted with a MIDI pickup tied to the electric guitar. The owner of musical instrument have to use the device during the performance, it is time consuming.

In Reference [6][7], the system has a built-in MIDI controller for electric guitar. In fact, when electric guitar and the MIDI controller are integrated, it is easy to use in most MIDI implementations; however, Reference [6][7] of devices are used as the electric guitar is not possible. For this reason, it is limited only to MIDI implementations. Further, it is different from the electric guitar, which can-not achieve these various expressions.

Reference [8][9], string vibrations are analyzed in two dimensions, realized by crossing the piezo pickups. In [10], it is fitted with a conduction band on the side of the neck of the electric guitar. These devices are difficult to mount on electric guitars, and as such, mounting is often required during manufacting. Further, wiring is complicated and can-not be removed immediately.

Unlike the adove, our proposed method is realized by software, achieving what other studies have attempted via hardware. Our proposed method can be used in environments consisting of a typical electric guitar player.

As noted adove, the electric bass is lower than the electric guitar. In previous research, detection accuracy of the electric guitar has been identified [11]; however, previous research has not reported on the detection accuracy of electric bass.

Therefore, in this study, we report on the detection accuracy of the electric bass.

3 BASS PITCH DETECTION ALGORITHM

Our proposed method detects the bass pitch from a given acoustic signal. In this chapter, we describe the algorithm behind our proposed method.

3.1 Converting the Frequency Spectrum to the MIDI Array

In our proposed method, we convert the given frequency spectrum to the MIDI array. In this way, the search process is in dependent of sampling frequency.

First, the acoustic signal of the musical instrument sound is $x = \{x_0, x_1, ..., x_{w-1}\}$, which represents Fourier transform

$$x_i = \frac{1}{w}\sum_{k=0}^{w-1} X_k \cdot e^{i2\pi\frac{i}{w}}, \qquad (1)$$

where w is the number of analysis frames. Here, (1), represents a set of the power spectrum in $X = \{X_0, X_1, ..., X_{w-1}\}$.

Converting the power spectrum into a MIDI array yields

$$M := \left\{ M_p = X\left(\text{floor}\left(\frac{wf}{s}\right)\right)\right\}, \qquad (2)$$

where p is the MIDI number and s is the sampling frequency. The MIDI number is in the range $0 \leq p \leq 127$.

3.2 Search for the Bass Pitch

Our proposed method searches for a plausible bass pitch from M. More specifically, this bass pitch search finds the maximum value repeatedly from M.

Given general set of real numbers $T = \{t \in \mathbb{R}\}$, we search for t from T, indicating the operation to return the index of t as

$$\text{index}(t) := T_i \Leftrightarrow t \Rightarrow i, \qquad (3)$$

Further, set S to store the maximum value is represented as

$$S := \left\{ \begin{array}{ll} S_0 & = \text{index}(\max_{0\leq j \leq 127} M_j) \\ S_{i+1} & = \text{index}(\max_{0 \leq j < S_i} M_j) \end{array} \right\}. \qquad (4)$$

The index set to be a $(S_i | i \in J)$ defined by J. Function $B(M)$ returns the MIDI number of the bass pitch, with $B(M)$ defined as,

$$B(M) = \left\{ \begin{array}{ll} J_{S_i} & M_{S_{i+1}}\beta + z \leq M_{S_i} - M_{S_{i+1}} \\ & \text{or} S_i - S_{i+1} \leq \alpha \\ -1 & otherwise \end{array} \right. , \qquad (5)$$

where α is the difference in interval $S_i - S_{i+1}$, β is the magnification of the threshold, z is power supply noise and J_{S_i} is the number of the MIDI of the bass pitch.

4 EVALUATION EXPERIMENTS

In this chapter, we discuss the results of applying our proposed method to the electric bass of an acoustic signal.

In our evaluation experiments, we conducted an experiment using a prototype system. we experimented with this prototype system by interacting with it as a player picking a single note of the electric bass. Experimental time consisted 1000 frames in the prototype system, which equated to approximately 16.6 seconds. Here the player repeatedly picked a quarter note at 120 beats per minutes (BPM). In our evaluation experiment, we evaluated the detection rate of the bass pitch of using our proposed method.

4.1 Experimental Environment

Table 1 shows the machine configuration of our evaluation experiments.

Table 1: Machine configuration of our evaluation experiments.

	overview
OS	Windows 7 Enterprise (64 bit)
CPU	Intel Core-i7 2600k
Memory	16 GB
GPU	NVIDIA GTX970
Audio IF	Presonous Audio Box USB

Table 2 shows specific settings of the Audio Box USB.

Table 2: Settings of the Audio Box USB

	overview
Input Volume	10 o'clock
Sampling Frequency	44,100 Hz
Buffer Size	256 samples

Here, the sampling frequency of the Audio Box USB was fixed at 44,100 Hz.

In our proposed method, we used cuFFT [1] as the FFT library with analysis frames of the FFT set at 32,768 points. Further the acoustic signal was multiplied by the Hanning window [12].

We used two types of electric basses in our experiments. i.e., a Fender Custom Shop Jazz Bass and a Music Man's Sting Ray. The Jazz Bass was equipped with two single-coil pickups, whereas the Sting Ray was equipped with a humbucker pickup and a three-band equalizer. Specifications for these electric basses are summarized in Table 3.

Table 3: Specifications of the electric basses used in our experiments.

	Sound Creation	Frets	Pickups
Jazz Bass	Passive	20	2
Sting Ray	Active	21	1

One of the authors served as the performer for the evaluation experiments; this individual had 10 years of experience playing the electric bass. For our experiments, we selected a player who could mute strings and accurately pick. The player used a guitar pick for all experiments.

4.2 Prototype System

Figure 1 shows the graphical user interface (GUI) of our prototype system.

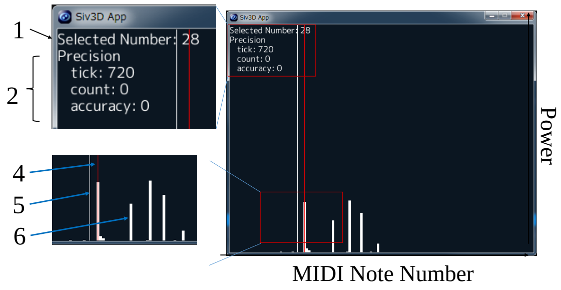

Siv3D App
Selected Number: 28
Precision
tick: 720
count: 0
accuracy: 0

Figure 1: GUI of our prototype system.

In the figure refers to the MIDI number of the detection target, while the lines of 2 show detection results. Here, tick is the number of elapsed frames, count is the number of detected frames, and the accuracy is the precision ratio of counts per tick. The red line shown in 3 is the detected bass pitch via our proposed method. Further, the white line of 4 in the figure shows the detection target on the MIDI number, while the white band of 5 is the power spectrum of MIDI frequencies. Note that the power spectrum is displayed in linear.

The is precision ratio shown in 6 of Figure 1 is defined as

$$P(c) = \frac{c}{t}, \tag{6}$$

where c is the count of 2, t is the tick of 2. $P(c)$ is the precision ratio of 2.

The prototype system updates 60 times per second, with an analysis of the number of frames of the FFT set is 32,768 points. The FFT shifts the analysis frame for each frame; the analysis frame shifts 1024 points at a frame.

We inspected the processing fall before the evaluation experiment. It was recorded 216,000 frames in the investigation. It was not able to confirm the processing fall in the results.

For our prototype, we used the Siv3D [13], TinyASIO [14][15], and cuFFT [2][1] libraries.

More specifically, Siv3D is a C++ library for media art that wrapping DirectX. Siv3D can be implemented using relatively few lines of source code as compared with other existing libraries. We therefore used Siv3D for the GUI of our prototype.

TinyASIO is a wrapper library for Audio Stream Input/Output (ASIO), which is a standard for communicating between an audio interface and software. We used this to receive Audio Box USB signals in the prototype.

Finally, cuFFT is a FFT library that NVIDIA offers; here, cuFFT is calculated by the GPU via parallel processing. Parallel processing is an advantage inherent to GPUs as compared to CPUs [3]. Note that cuFFT uses Stockham [16], which is more efficient than Cooley-Turkey [17]. We used cuFFT to obtain a frequency spectrum in the prototype system.

4.3 Results of our Evaluation Experiments

Table 4 summarizes the results of our evaluation experiments. Note that numbers in parentheses show standard deviation values. The table 4 is organized to show results for each tone control.

Table 4: Precision ratio of the bass pitch via our proposed method (%)

	Open	Close
Jazz Bass	98.5 (2.0)	98.0 (2.7)
Sting Ray	98.8 (3.2)	98.3 (5.8)

The precision ratio of the entire evaluation was 98.5% with a standard deviation of 3.8.

4.4 Examination

In previous our study, three electric guitar by a similar evaluation experiment yielded the precision ratio of 98.9%. Our proposed method converged at approximately 98%.

As noted above, the Sting Ray was equipped with a humbucker pickup and three-band equalizer. Of particular note, the Sting Ray experienced erroneous detection between the four-string zero to four frets. Erroneous detection occurred when the second harmonic overtone was extremely large.

Figure 2 shows the power spectrum of the Jazz Bass and Sting Ray instruments; in the figure, we present the four-string zero-fret for each electric bass.

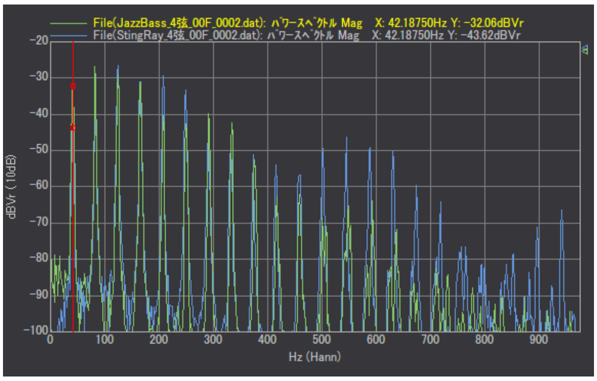


Figure 2: The power spectrum of the Jazz Bass and Sting Ray instruments.

Note that Figure 2 was generated by Ono-Sokki DS-0320 with a sampling frequency of 40,000Hz and FFT resolution of 16,384 points. Further, the green line represents the Jazz Bass, whereas the blue line represents the Sting Ray.

Each electric bass's f_0 frequency was 42.2875Hz. The Jazz Bass's amplitude was 0.278V, while the Sting Ray's amplitude was 0.270V. The Jazz Bass's f_0 power spectrum was -43.06 dBV, whereas the Sting Ray's f_0 power spectrum was -43.62 dBV, however, the Jazz Bass's f_1 power spectrum was -26.60 dBV, Sting Ray's f_1 power spectrum was -29.05 dBV. Here, a difference in the power spectrum of more than 10 dB was considered erroneous detection.

The overtone series disappeared due to inharmonicity of the high position [18][19]. Inharmonicity is a phenomenon that shifts the frequency of the harmonic by the stiffness of the string. The part sound series frequency is defined as

$$f_n = nf_0\sqrt{1+Bn^2} \tag{7}$$

$$B = \frac{\pi^3 Qd^4}{64l^2T}, \tag{8}$$

where B is an inharmonicity value, n is the number of the overtone series, Q is young's modulus, d is diameter, T is tension, and l is the length of the string. The overtone series will be a part sound series with inharmonicity. The overtone of the high position was affected by our proposed method, thus our proposed method has the property that it can remove overtones at or near the high position.

The player picked the same sound repeatedly in our evaluation experiments, however, the player was also able to pick complex melody at an elevated scale when the result had 0.3-second delays display in our prototype. Frequency resolution in the environment of the evaluation experiment was too high, thus the power spectrum remains for a long time. To detect complex melodies, it is better to apply the FFT to more analysis frames. For example, the number of analysis frames was only 32,768 points, 16,384 points, and 8,192 points, it is necessary to consider higher values still.

5 CONCLUSION

In this work, we focused our efforts on bass pitch detection of the electric bass. The typical six-string electric bass has the lowest note, i.e., that is 30.9Hz. It is difficult to detect 30.9Hz in one analysis frame, thus solving this detection problem requires the use of analysis frames or beat tracking [20].

The significance of our study was the discretization of a given performance, which is useful in a variety of applications. Our proposed method first converted a performance into MIDI. The audio signal uses a few kilobytes, whereas, MIDI uses a few bytes, thus MIDI is excellent for efficiency transmitting audio signals. The combination of MIDI and the sound source model may mask poor performance. When a player's performance is good, player motivation goes up. Clearly, players typically want to play g̈oods̈ounds.

For our future work, we aim to focus on transplanting our proposed method to Visual Studio Technology (VST) plugins, thereby making our proposed method available to the digital composers. Our proposed method converts an acoustic signal to MIDI. And composer efficiency of the electric guitar player is considered to be improved.

ACKNOWLEDGEMENTS

The authors would like to thank Rob for the English language review.

REFERENCES

[1] CUDA Nvidia. Cufft library, 2010.

[2] Akira Nukada, Yasuhiko Ogata, Toshio Endo, and Satoshi Matsuoka. Bandwidth intensive 3-d fft kernel for gpus using cuda. In *High Performance Computing, Networking, Storage and Analysis, 2008. International Conference on*, pages 1–11. IEEE, 2008.

[3] Jianbin Fang, Ana Lucia Varbanescu, and Henk Sips. A comprehensive performance comparison of cuda and opencl. In *Parallel Processing, 2011. International Conference on*, pages 216–225. IEEE, 2011.

[4] GraphTech Guitar Labs. Ghost modular pickup system.

[5] Roland. Roland gr-55.

[6] Inspired Instruments. You rock guitar midi guitar controller.

[7] Zivix. Jamstik+.

[8] Ichiro Tokuhiro, Daisuke Otsu, Isoharu Nishiguchi, Kurokawa Masaki, Kiyohiko Yamaya, and Takazawa Yoshimitsu. Measurement of the guitar technique by acoustic guitar mounted the 2 dimensional pickup. *Japan Ergonomics Society*, 45:192–193, 2009 (written in Japanese).

[9] Ichiro Tokuhiro. Necessity of approach with ergonomics to measure playing information of stringed instrument. *Japan Ergonomics Society*, 45sql:107–106, 2009 (written in Japanese).

[10] Hayami Tobise, Yoshinari Takegawa, Tsutomu Terada, and Masahiko Tsukamoto. Construction of a system for recognizing touch of strings for guitar. In *New Interfaces for Musical Expression, 2013. International Conference on*, pages 261–266. NIME, may 2013.

[11] Shoichiro Saito, Hirokazu Kameoka, Keigo Takahashi, Takuya Nishimoto, and Shigeki Sagayama. Specmurt analysis of polyphonic music signals. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(3):639–650, 2008.

[12] Andreas Antoniou. *Digital filters*. McGraw Hill, 1993.

[13] Ryo Suzuki. Play siv3d!

[14] Eiichi Takebuchi and Haruo Hayami. リアルタイム音声処理を 5 行で実現するためのライブラリの開発. *Information Processing Society of Japan. GN, [Groupware and Network services]*, 9(40):1–8, 2015 (written in Japanese).

[15] Eiichi Takebuchi. Tinyasio.

[16] Thomas G Stockham Jr. High-speed convolution and correlation. In *Proceedings of the April 26-28, 1966, Spring joint computer conference*, pages 229–233. ACM, 1966.

[17] James W Cooley and John W Tukey. An algorithm

for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[18] Harvey Fletcher, E Donnell Blackham, and Richard Stratton. Quality of piano tones. *The Journal of the Acoustical Society of America*, 34(6):749–761, 1962.

[19] Hanna Järveläinen, Vesa Välimäki, and Matti Karjalainen. Audibility of the timbral effects of inharmonicity in stringed instrument tones. *Acoustics Research Letters Online*, 2(3):79–84, 2001.

[20] Masataka Goto and Yoichi Muraoka. A real - time beat tracking system for musical acoustic signals. *nformation Processing Society of Japan. MUS, [Special Interest Group on MUSic and computer]*, 94(71):49–56, 1994 (written in Japanese).

A Zero Interruption-Oriented Mobile Video-on-Demand System by Hybrid Broadcasting Environments

Tomoki Yoshihisa*, Tomoya Kawakami**, Yusuke Gotoh***

* Cybermedia Center, Osaka University, Japan
** Graduate School of Information Science, Nara Institute of Science and Technology, Japan
*** Graduate School of Natural Science and Technology, Okayama University, Japan
yoshihisa@cmc.osaka-u.ac.jp

Abstract - Due to the recent proliferation of mobile devices and video-on-demand delivery, mobile video-on-demand delivery, i.e., the users watch videos using mobile devices, gets great attention. In mobile video-on-demand delivery, the users often watch videos while moving outside such as riding cars or trains. Current mobile video-on-demand delivery faces the problem that interruptions of video playback occur in some situations. So, some interruption reduction techniques have been studied. However, these techniques are originally designed for non-mobile devices and it is difficult to solve the problem for mobile devices under various situations. Hence, in this paper, we propose some techniques aiming to zero interruption. To reduce the video interruptions effectively, our proposed techniques use hybrid broadcasting environments. We develop a mobile video-on-demand system with our proposed techniques and report our experiments of mobile video-on-demand delivery using the developed system[1].

Keywords: Internet Broadcasting, Continuous Media, Streaming Delivery, Mobile Devices

1 INTRODUCTION

Due to the recent development of Information and Communication Technologies (ICT), mobile devices such as smart phones or compact PCs become popular. Mobile devices are small lightweight and can connect to Internet at most places. So, the users can get information from Internet using mobile devices at most outside places even though they do not seat in front of non-mobile devices such as desktop PCs. Meanwhile, video-on-demand delivery such as Internet broadcasting by YouTube or TV companies becomes popular due to the communication speed up of Internet. In video-on-demand delivery, the users select their preferable videos from homepages and request the video deliveries to the delivery server. The delivery server sends the video data to the requested clients according to their requests. The users can watch the video from the playing position of the received data without waiting for the reception of all data by using streaming technique, in that the clients play the video data while receiving them. The proliferation of mobile devices and video-on-demand delivery leads mobile video-on-demand delivery, i.e., the users watch videos using mobile devices. In mobile video-on-demand delivery, the users often watch videos while moving outside such as riding cars or trains since they can select and watch videos at most places.

Current mobile video-on-demand delivery faces the following problems.

Problem 1: Interruptions occur when there are a large number of clients.

Interruptions of video playback occur when there are a large number of clients since the bandwidth between the server and each client decreases. For example, suppose the case when a user requests playing a popular video delivered in YouTube when he/she is stopping at a red traffic signal. In this case, the video playback sometimes does not start or interrupts even if it starts. This problem also occurs for non-mobile devices, but frequently occurs for mobile devices since there are a large number of mobile devices.,

Problem 2: Interruptions occur when the condition of electric wave gets worse.

Mobile video-on-demand delivery is often used while moving and the condition of the electric wave for Internet connection changes. A worse electric wave condition causes a lower bandwidth. So, the video playback interrupts when the bandwidth becomes lower than the bit rate. For example, a user requests playing a video while riding on a train at a station and starts watching it. When the train moves to a far distance from the station, the electric wave does not reach to his/her mobile device and the video playback interrupts.

Problem 3: Video playback stops when the remaining battery is low.

Mobile devices are often used outside and it is difficult to charge their batteries outside. So, the users may reduce the battery consumptions caused by video playbacks to lengthen the running time of the devices. The battery consumptions for playing high bit rate videos are high since the processing data size per time increases. Hence, in mobile video-on-demand delivery, some users prefer playing low bit rate videos to high bit rate videos so as to lengthen the time to play videos.

[1] This research was supported in part by Grant-in-Aids for Scientific Research (B) numbered 15H02702, and Grant-in-Aids for Challenging Exploratory Research numbered 26540045.

Some interruption reduction techniques have been studied. But, these techniques are originally designed for non-mobile devices and it is difficult to solve these problems. A simple approach to solve the problem 1 is exploiting streaming delivery techniques such as CDN (Contents Delivery Network) or P2P streaming. However, it is difficult to solve the problem by only these techniques when there are too many mobile devices, e.g., many mobile devices in Japan play video recording disaster situations. Regarding the problem 2, mobile devices can keep playing videos even when they lack Internet connections by sufficiently buffering the video data. However, mobile devices cannot sometimes get enough time to buffer the data, e.g., the train moves fast or the power of the electric wave is weak. Regarding the problem 3, battery consumptions can be reduced by stopping other applications running on mobile devices. However, in this simple approach, the convenience of mobile devices degrades.

Hence, in this paper, we propose three techniques aiming to zero interruption. To reduce the interruptions of playing videos effectively, our proposed techniques use hybrid broadcasting environments. In hybrid broadcasting environments, the clients can receive data both from the broadcasting system and the communication system. The broadcasting system can deliver data to all clients concurrently. So, by broadcasting the data that are requested by many clients, the interruptions of playing videos are reduced effectively. Also, we develop a mobile video-on-demand system using our proposed techniques and report the experiments of mobile video-on-demand delivery using our developed system.

2 RELATED WORK

2.1 Video-on-Demand Services

Some video-on-demand systems have been developed for services.

One of the famous system for delivering the videos that are submitted by the users is YouTube [1]. YouTube service started on 2004 at United States and widely used for a user submitted video delivery on demand service. USTREAM also uses a video-on-demand system for delivering the user submitted videos [2]. USTREAM is often used for live broadcasting on Internet. Niconico-douga is a Japanese video-on-demand service for user submitted videos [3].

Netflix is a service delivering the videos that are provided in the medias of Blu-ray or DVD by contents companies [4]. Netflix originally started a DVD rental service on 1997. Using the stored DVDs, Netflix started video-on-demand service. Hulu also uses a video-on-demand system for delivering the provided video contents [5]. Hulu provides the service with charge. There are many other video-on-demand services all over the world [6].

Techniques used in the above methods and services are originally designed for non-mobile devices and it is difficult to solve the problem for mobile devices under various situations.

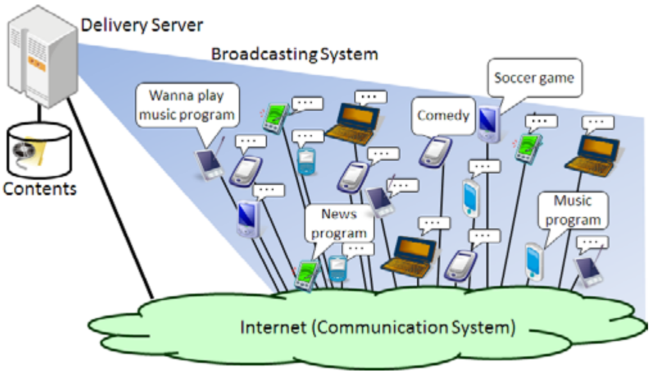


Figure 1: A hybrid broadcasting environment

2.2 Types of Video-on-Demand Delivery

Video delivery combining the video-on-demand and the near video-on-demand techniques is called unified video-on-demand delivery. Some methods for unified video-on-demand delivery have been proposed [7, 8]. In these methods, the delivery serer fixes the broadcast schedule and send the data of that time to broadcast is long after via Internet. A method that generates the broadcast schedule dynamically is proposed in [9].

3 PROPOSED TECHNIQUES

In this section, we explain our proposed techniques. First, we explain our assumed hybrid broadcasting environments. After that, we explain our proposed techniques to solve the problems described in Section 1.

3.1 Hybrid Broadcasting Environments

Figure 1 shows our assumed hybrid broadcasting environment. The clients in the broadcasting area can receive data from the broadcasting system. Also, they can request their preferable data to the server and can receive them from the communication system. The broadcast station delivers data via some broadcast channels and is managed by the server. The server has streaming data and can broadcast the data to the clients using the broadcast station. Also, it can send the data to the clients using the communication system by unicasting.

3.2 Stream Merge

To solve the first problem, we propose a mobile video-on-demand system using stream merge technique. When there are many mobile devices that receive video data, the video-on-demand system can avoid decreasing the bandwidth between the server and the clients by sending the data to some clients concurrently and reducing the communication traffic. For example, suppose the case when Client 1 starts playing a news program in a train on 8:00 p.m. as shown in the left side of Figure 2. The duration of the video is 5 minutes. One minutes after this, Client 2 starts playing the same video. In this case, the server delivers 2 video streams (a set of video data from the begging to the end) for 4 minutes from 8:01 to 8:05 in the simple conventional

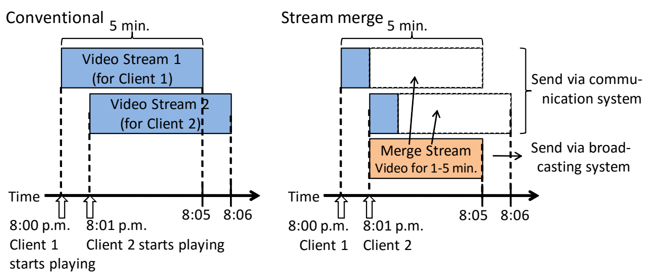

Conventional
5 min.
Video Stream 1 (for Client 1)
Video Stream 2 (for Client 2)
Time
8:00 p.m. Client 1 starts playing
8:01 p.m. Client 2 starts playing
8:05
8:06
Stream merge
5 min.
Send via commu-nication system
Merge Stream Video for 1-5 min.
Send via broad-casting system
Time
8:00 p.m. Client 1
8:01 p.m. Client 2
8:05
8:06

Figure 2: An example of stream merge

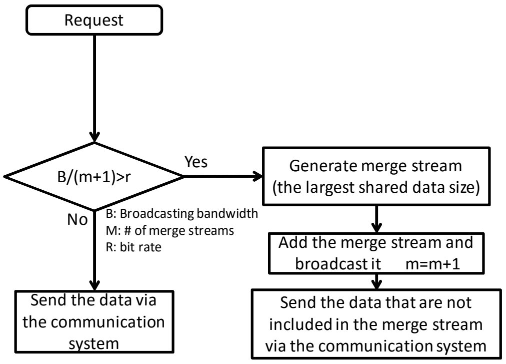

Request
B/(m+1)>r
Yes
No
B: Broadcasting bandwidth
M: # of merge streams
R: bit rate
Generate merge stream (the largest shared data size)
Add the merge stream and broadcast it m=m+1
Send the data that are not included in the merge stream via the communication system
Send the data via the communication system

Figure 3: Flow chart for merging streams

method. However, in our proposed technique, the server merges 2 streams from one minutes after the beginning to the end and delivers the merged stream to all clients as shown in the right side of Figure 2. Client 2 can receive all data since it receives the data for one minutes from the beginning from the server directly. In this case, the server only delivers 2 streams for 1 minutes and can reduce the communication traffic. This is a simple example for the case of 2 clients. Actually, the server merges some streams for multiple clients. Broadcasting systems are suitable for delivering data to all clients. So, in our proposed technique, the server delivers the merged stream via the broadcasting system and other streams via the communication system. For this, we use hybrid broadcasting environments.

Figure 3 shows the flow chart for merging streams in our proposed technique. *B* denotes the broadcasting bandwidth and *m* denotes the number of the currently merged streams. The broadcasting bandwidth for each merge stream in case of adding another merge stream is $B/(m+1)$. When a client requests playing the data, the server compares this value and the bit rate r. If this is larger than r, the server generates a merge stream and sends it via the broadcasting system since interruptions do not occur by receiving the merge stream. Otherwise, the server sends the requested data to the client via the communication system since interruptions can occur by adding a merge stream.

3.3 Spare Data Delivery

To solve the second problem, we propose the spare data delivery technique. In the spare data delivery, for the case when the condition of the electric wave gets worse, the system delivers spare data beforehand. Different from the traditional technique that the clients buffer the requested video data, spare data are the data to keep the motivation for the users to watch the video. We use two types of spare data.

One is a data not related to the requested video such as commercial or weather forecasting. For example, the mobile device plays a commercial movie when the user riding train enters a tunnel and cannot catch the electric wave. The clients can download such spare data before requesting playing videos since the contents of the spare data is not related to video requests. The clients play spare data if the video playback interrupts.

The other is data related to the requested video but the data size is small compared with the video played when the clients have enough bandwidth, e.g., text or static image data. For example, the mobile device shows news by text when the user watches a news program and the video playback interrupts. The clients can download such spare data within a short time after requesting playing videos since the data size is very small.

When the users want to watch the whole requested video, the former type is suitable since the clients play the requested video although other videos can be played midstream. When the users want to grasp the content of the requested video, the latter type is suitable. In case of delivering the spare data not related to requests, broadcasting systems is suitable to deliver the data such as commercial or weather forecasting since these data are used as spare data for all clients. So, in our proposed technique, the server delivers such spare data via the broadcasting system on hybrid broadcasting environments. The detail algorithm such as how to get spare data and determine them depends on the implementation. We will explain the implementation for our developed system in the next section.

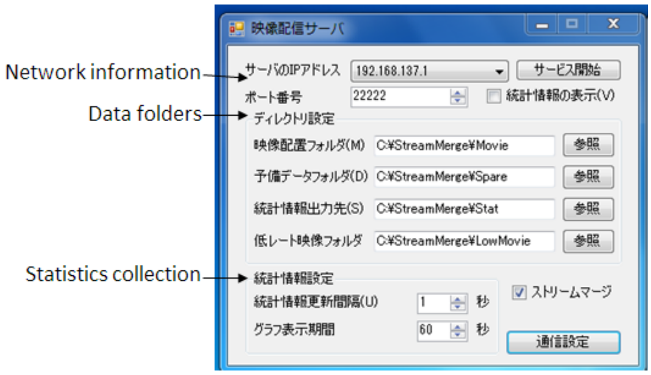
3.4 Remaining Battery based Bit Rate

To solve the last problem, we propose the remaining battery based bit rate. In the remaining battery based bit rate, the bit rate of the video is controlled on the client side. The clients play the video with normal bit rate when they have sufficient remaining battery to play the video. Otherwise, the clients reduce the battery consumption by decreasing the bit rate of the video. To change the bit rate on the client side, our proposed technique uses multi-bit rate encoding. In multi-bit rate encoding, the clients can play the video with some bit rates determined beforehand. The resolution or the size decreases when playing the video at a lower bit rate. For example, by using multi-bit rate encoding, the clients can play the video at 128Kbps even when the server delivers the video at 1Mbps bit rate. In our proposed technique, the users set their preferable remaining battery and the corresponding bit rate. The server delivers the video at the highest bit rate for all the bit rate used by the clients so as to decrease the traffic.

4 DEVELOPED SYSTEM

We develop a mobile video-on-demand system using our proposed techniques. We call our developed system Metreamer, the abbreviation of mobile ever streamer.

4.1 Overview



サーバのIPアドレス 192.168.137.1 [サービス開始]
ポート番号 22222 □ 統計情報の表示(V)
ディレクトリ設定
映像配置フォルダ(M) C:¥StreamMerge¥Movie [参照]
予備データフォルダ(D) C:¥StreamMerge¥Spare [参照]
統計情報出力先(S) C:¥StreamMerge¥Stat [参照]
低レート映像フォルダ C:¥StreamMerge¥LowMovie [参照]
統計情報設定
統計情報更新間隔(U) 1 秒
グラフ表示期間 60 秒
☑ ストリームマージ
[通信設定]

Figure 4: A screenshot of Metreamer server software

Metreamer uses the broadcasting address of UDP protocol, that sends the same data to all the clients connected to the same network, as the broadcasting system and the unicasting of TCP protocol as the communication system. The delivery server has the video data and the spare data. If the server sends the spare data at the same time with the video data when the clients request playing the video, interruptions easily occur. So, in Metreamer, the server sends the spare data when there is a remaining bandwidth capacity such as after stopping the service or finishing all video data deliveries. Metreamer can use video, image, and text data as the spare data. The video encoding type is widely used MPEG2. Metreamer can measure some statistic information such as the number of clients, and so on.

4.2 Software for Servers

Figure 4 shows a screenshot of the software for the servers of Metreamer. The software runs on Windows 7 and uppers. The server sometimes has some IP addresses and so the users can change the IP address on the software. The users can also change the directories for the video, spare, and statistic data. To show the statistic information, the users click the show button and set the interval to get the information. The users can set many other parameters for the video delivery as shown in the figure. By clicking the service start button, the software starts the delivery service and waits for the requests from the clients.

4.3 Software for Clients

Figure 5 shows a mobile device playing a video using Metreamer. The mobile device is a special device for our p developed system. But the software for clients can run on other Android (4.1 or upper) mobile devices. The software for the clients of Metreamer first shows the login screen as shown in the left side of Figure 6. By logging in, the client software can retrieve their settings. To enabling logging in, the users first register themselves to the server and after that they get IDs and passwords. The screen for setting some parameters is shown in the right side of Figure 6. On this screen, the users can set the IP address of the server. In our developed system, the users directly input the server's IP address, but it can be set automatically by getting it from the server list wrote in some homepages in Internet. On the setting screen, the users can set the data amount for cashing, the spare data types for the spare data delivery technique, the bit rates and the remaining battery to change the bit rate for remaining battery based bit rate.

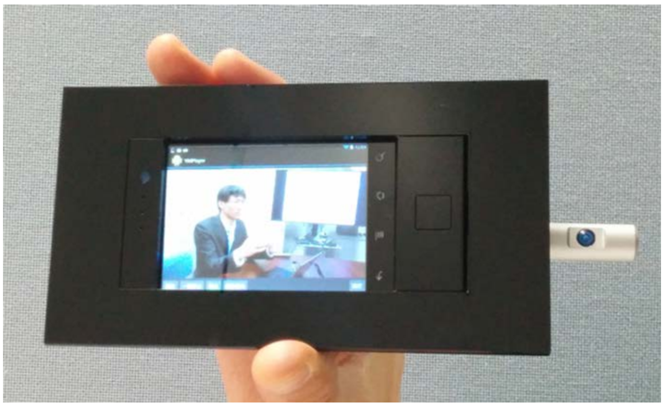
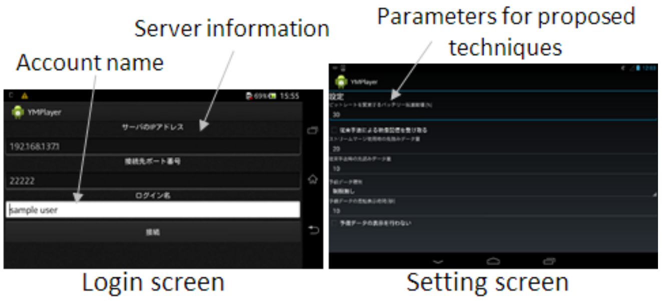


Figure 5: A mobile device running Metreamer



サーバのIPアドレス
192.168.137.1
接続先ポート番号
22222
ログイン名
sample user
接続

設定

Figure 6: A screenshot of Metreamer client software

4.4 Implementation of Proposed Techniques

4.4.1 Stream Merge Implementation

The server gets the broadcasting bandwidth B from its specification beforehand. If the actual broadcasting bandwidth largely differs from the specification, the server modifies the value. The server knows the number of merge stream N since it sends the merge streams. By checking the duration and the data size of each video, the server can get the bit rate r of the video data. Using these values, the server calculates $B/(m+1)$ when it receives the requests to play video from the clients. If the value is smaller than r, the server generates the video stream for the client and send it to the client. If the value is larger than r, the server generates a merge stream so that the duration of the merge stream becomes the longest. The merge streams are not merged to other merge stream again to make the merging algorithm simple in our developed system. Then, the server sends the merge stream to all clients via the broadcasting system and stops sending the streams that are merged to the merge stream. At the same time, the server sends the remaining data for the client, i.e., the data that the requested client cannot receive from the merge stream, to the client.

4.4.2 Spare Data Delivery Implementation

When a user runs the client software first, the client does not have spare data. So, the client requests the spare data to

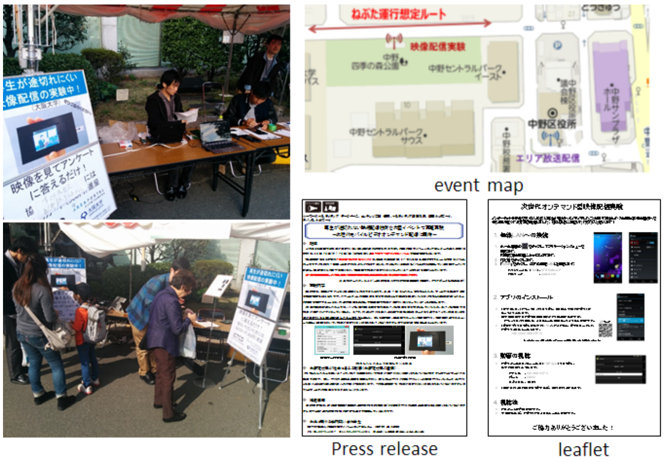


Figure 7: Our experiment on Nakano central park

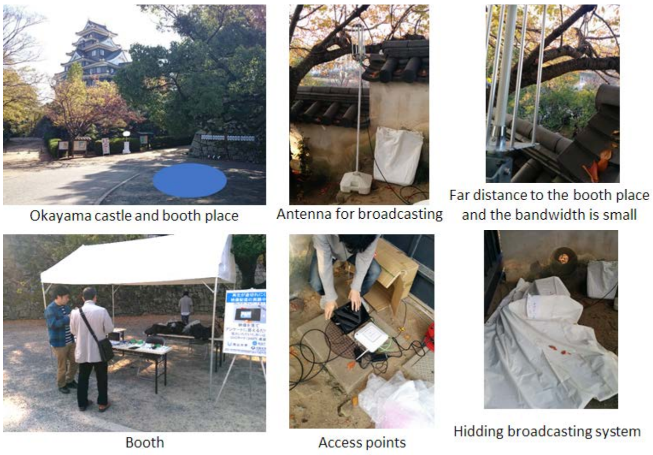

Okayama castle and booth place
Antenna for broadcasting
Far distance to the booth place
and the bandwidth is small
Booth
Access points
Hidding broadcasting system

Figure 8: Our experiment on Okayama castle

the server if it has no spare data. If the user changes the spare data type by setting screen, the client requests the spare data. The spare data are stored in the server and the clients receive some spare data from the server. Thus, the clients have some spare data before starting playing the video. If a client does not have the subsequent video and an interruption will occur, it selects a spare data randomly and shows it. If the number of the spare data is largely less than the number of the interruptions, the clients show the same spare data sometimes.

4.4.3 Remaining Battery based Bit Rate Implementation

The client software can get the remaining battery by asking it to the OS. When the remaining battery changes and the bit rate of the video changes according to the user's setting, the client changes the bit rate. If the client is playing a video, immediately changes the bit rate. Otherwise, the client uses the changed bit rate from the next video playback.

5 EXPERIMENTS

To investigate the effectiveness of our proposed techniques, we used our developed system in two practical situations since computer simulations do not reflect actual situations completely.

5.1 Experiment at Nakano Central Park

With the cooperation of an industrial promotion organization in Nakano, we got a chance to provide a mobile video-on-demand service on practical field.

Before the full experiment, we did a preliminary experiment on July 2nd, 2014 at Nakano central park. In the preliminary experiment, we provided a mobile video-on-demand service using our developed Metreamer for 2 clients. We used the broadcasting equipment installed in the park. The broadcasting equipment uses wireless LAN. In the environment for the preliminary experiment, the client could receive the data both from the broadcasting system and the communication system when they were in the broadcasting area. In case where the clients were out of the broadcasting area, the clients could receive the data only from the communication system. We checked the number of interruptions using the measuring function of Metreamer and confirmed that Metreamer realized zero interruption in the case where the clients were in the broadcasting area. However, otherwise, interruptions frequently occurred. This was because many electric waves were emitted around the park and the communication bandwidth decreased down to the bit rate frequently. So, we decreased the bit rate of the video and did an experiment again on July 17th. In the experiment, Metreamer realized zero interruption even where the clients were out of the broadcasting area. To further investigate the performances of Metreamer, we did a full experiment on a large event.

Considering time and scale, the full experiment was done on Tohoku-Fukkou-Daisaiten-Nakano held on Oct. 25th and 26th, 2014 at the same place. The situations are shown in Figure 7. For the experiment, we made the press release shown in the figure. To show the press release as it is, this is Japanese. The event is the largest one in Nakano area and the attendees are 170,000. Nebuta (large paper made statures) moves around the park during the event period. To deliver the video that is interesting for the attendees, we used the movie for moving Nebuta on the last year on the first day and that on the first day on the second day. The duration of the video was 2 minutes and 1 second, and the bit rate is 2Mbps. We set wireless LAN access point. The mobile devices of the attendees can receive the data both from the broadcasting system and the communication system by making them connect to the access point. We got booth, and there, the attendees watched the video using Metreamer. For the performance comparison, we also provide a video-on-demand service using conventional Windows Media System. In the environment for the full experiment, the video soon interrupted when the mobile devices started playing the video using the system. On the other hand, our developed Metreamer realized zero interruption even when about 10 clients connect to the system.

5.2 Experiment at Okayama Castle

With the cooperation of Okayama city, we got a chance to provide a mobile video-on-demand service at Okayama castle. Considering time and scale, the experiment was done on Imagineering OKAYAMA ART PROJECT held on Nov. 22th, 2014 at Torishiro park in Okaya castle. This is the event to demonstrate arts made out and in Okayama. The situations are shown in Figure 8. We set a wireless LAN access point and booth in Okayama castle. Since Okayama castle does not have broadcasting equipment, we used the access point for the communication system and the broadcasting system. Okayama city has a PR of the city and uses it in some events. So, we delivered the video for the experiment. The duration was 3 minutes and 30 seconds and the bit rate is 1 Mbps. Same as the experiment at Nakano central park, the mobile devices of the attendees can receive the data both from the broadcasting system and the communication system by making them connect to the access point. In this environment, the video soon interrupted when the mobile devices could not receive the data from the broadcasting system. On the other hand, our developed Metreamer realized zero interruption even when a few clients connect to the system.

5.3 Evaluation from Experiments

The experiments at Nakano central park was an experiments for a big event. Many people attended to the event during the period. Moreover, we delivered the video of moving Nebuta, which is the main event and many attendees were interested in. So, 16 clients connect to Metreamer at maximum. In the experiment, we made the situation that the clients could receive the data only from the communication system by moving them to the out of the broadcasting area. Even in this case, we confirmed that Metreamer realized zero interruption. Also we made the situation that the clients could receive the data only from the broadcasting system by disabling data transfer via the communication system. Even in this case, we confirmed that Metreamer realized zero interruption.

The experiments at Okayama castle was an experiments for a middle scale event. Only a few people were there sometimes. So, 7 clients connect to Metreamer at maximum. This is smaller than that of Nakano central park. Also in the experiment, we confirmed that Metreamer realized zero interruption even where the clients could receive the data only from the communication system or the broadcasting system.

6 CONCLUSION

Due to the recent proliferation of mobile devices and video-on-demand delivery, mobile video-on-demand delivery gets great attention. In this paper, aiming to zero interruption, we proposed 3 techniques for hybrid broadcasting environments. We developed a mobile video-on-demand system using our proposed techniques called Metreamer. In this paper, we reported the experiments of mobile video-on-demand delivery using our developed system.

In the future, we will again show the effectiveness of our proposed techniques by computer simulation. Also, we will develop the system for multiple streaming servers and live broadcasting.

REFERENCES

[1] YouTube: http://www.youtube.com/.
[2] Ustream.tv:You're On: http://www.ustream.tv/.
[3] niconico: http://www.nicovideo.jp/.
[4] Netflix: http://www.netflix.com/.
[5] Hulu – Watch TV, Original, and Hit Movies: http://www.hulu.com/.
[6] Actvila: http://actvila.jp/.
[7] T. Taleb, N. Kato, and Y. Nemoto: Neighbors-Buffering-Based Video-on-Demand Architecture, Signal Processing: Image Communication, Vol. 18, Issue 7, pp. 515-526 (2003).
[8] J. B. Kwon: Proxy-Assisted Scalable Periodic Broadcasting of Videos for Heterogeneous Clients, Multimedia Tools and Applications, Springer, Vol. 51, No. 3, pp. 1105-1125 (2011).
[9] T. Yoshihisa: A Data Segments Scheduling Method for Streaming Delivery on Hybrid Broadcasting Environments, Proc. of International Workshop on Informatics (IWIN2015), pp. 3-8 (2015)

Application of MongoDB with Transaction Feature in Production Management System

Tsukasa Kudo[†], Yuki Ito[†], and Yuki Serizawa[†]

[†]Faculty of Comprehensive Informatics, Shizuoka Institute of Science and Technology, Japan
kudo@cs.sist.ac.jp

Abstract - We had implemented the transaction processing system for MongoDB, which is a kind of NoSQL database, to maintain the ACID properties in the case of updating plural data in a lump sum. However, at the present time, though many NoSQL databases are used actually, there are few application cases which need such a transaction processing. On the other hand, with spread of the IoT, a variety of sensors have been widely used, such as the surveillance cameras and so on. So, the system operations, in which the sensor data is saved in the database directly to get the necessary information, are spreading. Furthermore, with the spread of the wearable devices, inexpensive sensors have become popular. Therefore, the system operations, in which many users save or query the sensor data concurrently, have become easy. In this situation, it is expected that the transaction feature becomes necessary to perform the concurrency control even for the NoSQL databases. In this paper, we propose the application of the transaction feature of MongoDB for the production management system utilizing such a sensor data.

Keywords: database, transaction processing, MongoDB, IoT, production management system

1 INTRODUCTION

With the development of the Internet business, it has become necessary that the database management systems adopt the feature called 3V, that is, Volume (huge amount), Velocity (speed) and Variety (wide diversity) [8]. Thus, the various database management systems called NoSQL database, which is different from the conventional relational database management systems (RDBMS), have been proposed and put to practical use [12]. On the other hand, in order to deal with the 3V feature, many of the NoSQL databases maintain only the BASE (Basically Available, Soft state and Eventually consistent) property [3]. As for this property, the above-mentioned ACID properties are maintained only in the case of updating a single data. So, in the case of updating the plural data, the data is updated one after another, and finally they become to be updated. That is, there is the anomaly in the midst of this updating, such as one data has been updated and another has not been updated.

To solve this problem, we showed the method of transaction processing for MongoDB, which is a kind of the NoSQL database. By this method, transactions can be performed with the designated individual isolation level according to the business requirement [7]. And, as for the isolation level of REPEATABLE READ, it was possible to perform the concurrency control with maintaining the ACID properties; as for READ UNCOMMITTED, the query process could be performed at the same performance with "findOne" method of MongoDB, which is used to query single data. On the other hand, as for the NoSQL database, since there are few implementation examples of such a strict transaction processing, the verification of the validity of the application in the actual business system became the next challenge.

Here, our laboratory is supporting the introduction of the production management system for the actual company. Especially, as for the inventory management system, which is one of its subsystem, it is important to grasp the actual inventory of parts efficiently. However, since the type of parts are so many in this company, this work load is very large. In particular, such the office work efficiency becomes so low in the field and it cannot be performed so often in order not to disturb the production works. For this reason, the accurate inventory often cannot be grasped, and it has become the factor in causing the shortage of parts.

For this problem, we paid attention to the step of the product assembly preparation, in which the inventory is decreased. And, we have proposed the following: to reduce the inventory management load, the inventory of the parts should be grasped in each product of each order; the pictures or videos in the field should be stored to the database, and the inventory manager should count the actual inventory in the office by querying them. Furthermore, as for the inventory of parts, it is necessary to manage the total quantity in both of the parts shelf and product assembly field. In other words, the requirement of the database has to include the following. First, the database can save not only the characters and numerical data but also the huge data such as images or videos. Second, since the plural workers move the parts between the parts shelf and product assembly field, the transaction processing on the plural *collections* (corresponding the table in RDBMS) should be equipped to perform the concurrency control.

So, we have positioned this inventory management system as the application case of MongoDB equipping the transaction feature. And, we implemented and evaluated a prototype of the production management system, which database saved images and equipped the concurrency control across the two *collections*. As a result, we found the following: the transaction feature across two *collections* could be implemented by applying the above mentioned method; to implement the efficient transaction, it is effective to perform the image manipulation before the start of the transaction. In this case, the concurrency control on the images are performed through the *collection* of the inventory quantity.

The remainder of this paper is organized as follows. Section 2 shows our transaction feature and the problem of the target production management system. And, we propose an inventory management method and show its model in Section 3. Section 4 shows the implementation of this model and the experimental results. We discuss on the result in Section 5, and Section 6 concludes this paper.

2 RELATED WORKS AND PROBLEMS

2.1 Transaction Feature for MongoDB

MongoDB is a document-oriented NoSQL database, and its data is stored as a *document* of the JSON (JavaScript Object Notation) format as shown in Fig. 1 [10], [15]. Its *document* is composed of the fields, and {"_id": Id1 } expresses the field having the identifier "_id " and value "Id1". Here, "_id" indicates the object ID that corresponds to the primary key of the relational databases. In Fig.1, the other fields of the *document* are "name" and "address". Here, field "name" has a nested structure, which is composed of field "first" name and "last" name. Since MongoDB has such a structure, it is not necessary to define the scheme of the database beforehand like RDBMS. That is, the fields of each *document* can be added or removed at any time. Incidentally, the set of *documents* composes the "collection", and each corresponds to the records and the table in a relational database although not strictly. Thus, single *collection* can have *documents* of various structures.

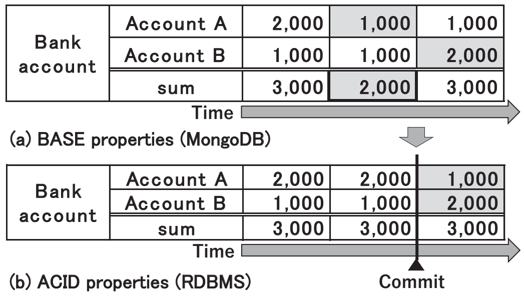
In addition, similar to SQL of the RDBMS, CRUD (Create, Read, Update, Delete) data manipulation is provided. However, since the transaction feature is based on the BASE properties, the ACID property can be maintained as for only single *document*. Therefore, there is the problem that the ACID properties of transactions cannot be maintained in the case of updating multiple *documents* simultaneously. Here, the ACID properties are defined as the following 4 properties: Atomicity means the transaction updates completely or not at all; Consistency means the consistency of database is maintained after it is updated; Isolation means each transaction is executed without effect on the other transactions executing concurrently; Durability means the update results survive the failure [5].

For example, in the case of the bank account transfer from account A to account B, the total amount of the both accounts does not change. However, as shown in (a) of Fig. 2, since the ACID properties are not maintained on the entire updating in MongoDB, there is the problem of the anomaly. That is, the halfway state during the updating is queried by the other transactions: one data has already updated; another data has not updated yet. In this case, when the account A was updated, the anomaly that the sum of the query result was reduced to 2,000 temporarily happened. On the other hand, in the case where the same procedure was performed in the RDBMS, this halfway state can be concealed until the commit as shown in (b).

Furthermore, as for SQL of the RDBMS, the isolation levels of the transactions are defined. That is, corresponding with the business requirement, the suitable isolation level can be selected: it is the efficient execution, or it is the strict concurrency control [5]. So, we implemented and evaluated the method to perform each transaction with the designated isolation level in MongoDB, as well as the RDBMS as shown in Table 1. As a result, we confirmed the following: the isolation levels of Table 1 were achieved; the query transaction performance at READ UNCOMMITTED is same as the query method of MongoDB [7].

```
{ "_id" : Id1, "name" : { "first" : "Tsukasa", "last" : "Kudo" },
"address" : "Hukuroi-shi"}
```

Figure 1: An example of MongoDB document.



Bank account	Account A	2,000	1,000	1,000
	Account B	1,000	1,000	2,000
	sum	3,000	2,000	3,000

Bank account	Account A	2,000	2,000	1,000
	Account B	1,000	1,000	2,000
	sum	3,000	3,000	3,000

Figure 2: Problem of transaction processing in MongoDB

We show the overview of this method below. First, it performs the lock operation during the access to the *document* as well as the RDBMS as shown in Table 1. In Table 1, "2PL" shows the two phase locking protocol [5]. Incidentally, in this method, to prevent the cascade abort of the transaction, the rigorous 2PL is adopted as well as the RDBMS. That is, the lock is held until the commit or rollback.

Second, as shown in Fig. 3, the *document* of the *Data collection*, which saves the business data, has the fields of both the data before and after the update. In addition, it has the fields to save the information of the transactions locking it: for each of the shared lock and the exclusive lock. And, in order to manage the transactions that are locking the *document* of *Data collection*, we implemented *TP (transaction management) collection*. The *document* of *TP collection* saves the corresponding transaction state: before or after the commit. Also, it saves the isolation level of the transaction.

For the case of Fig. 2, we show the procedure to query the *document* of $Datacollection$ when it is updated with the isolation level READ COMMITTED or REPEATABLE READ. In the case of the bank account transfer from the account A to account B, which is shown in (a) of Fig. 2 as the case of MongDB, both of the data before and after update are retained. And, the data before update is queried until the com-

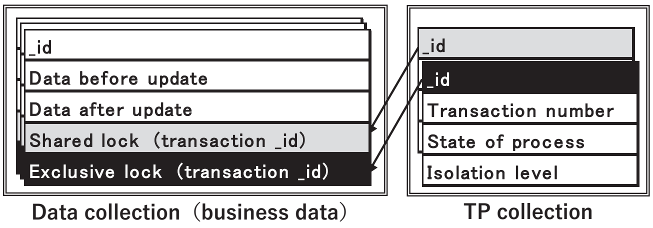


Figure 3: Transaction processing method for MongoDB

Table 1: Locking protocol of each isolation level

Isolation level	Exclusive lock	Shared lock
READ UNCOMMITTED	2PL	(none)
READ COMMITTED	2PL	While query
REPEATABLE READ	2PL	2PL

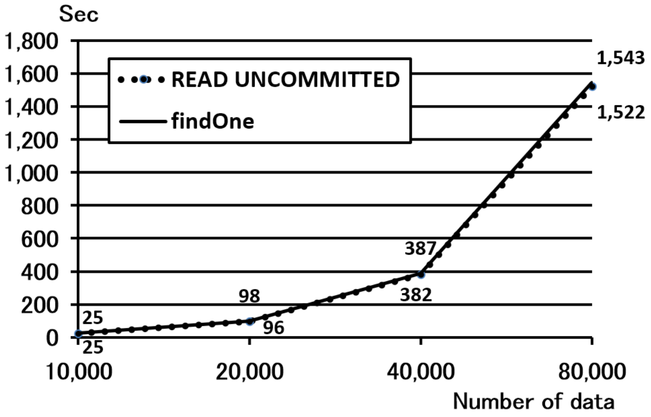


Figure 4: Comparative evaluation with findOne of MongoDB

mit of the transaction; the data after update is queried after the commit. Therefore, in (a), both update results of the account A and B are not queried until the commit; both update results are queried only after the commit; That is, the query results are similar to the RDBMS shown in (b). Thus, the halfway state during the updating is concealed from the other transactions, and the transaction processing maintaining the isolation can be provided.

On the other hand, the Velocity (speed) of the 3V, that is, high efficiency for the query of the NoSQL databases is generally required. So, we performed the comparative evaluation on the query processing between this method and MongoDB. As for the former, the query transaction was performed with the isolation level READ UNCOMMITTED, in which the query was performed efficiently without the shared lock as shown in Table 1. As for the latter, we used "findOne" method, which was the standard query method of MongoDB. As a result, we found that the performance of the both are almost the same as shown in Fig. 4 [7].

Here, it has been shown that if all the transactions are performed with any of the isolation level shown in Table 1, then any transaction can be performed with the designated transaction level [5]. In other words, by using this method, the usual query transactions can be performed efficiently with the isolation level READ UNCOMMITTED as in the conventional data manipulation of MongoDB; only the query and update transactions, which need the strict data manipulations as shown in Fig. 2, can be performed with the isolation level READ COMMITTED or REPEATABLE READ.

2.2 Target Production Management Business and Problem

Our laboratory received the request of the simple production management system from the manufacturing company: the implementation, introduction and support of the system operations. Previously, we had introduced the MPP (Material Requirement Planning) system to automate the calculation of the quantity and cost of the parts, which is necessary to assemble the ordered products. We show the over view of the MPP system in Fig. 5. For example, 2 parts Y is used for product A; 3 for Product B. So, in the case of order 1, which is composed of 10 products A and 20 products B, 80 parts Y is necessary. The cost of the parts, which calculated by this system, is used as the master data of the system in conjunction with the order company by EDI.

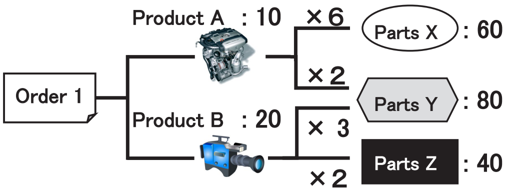


Figure 5: Composition of MRP system

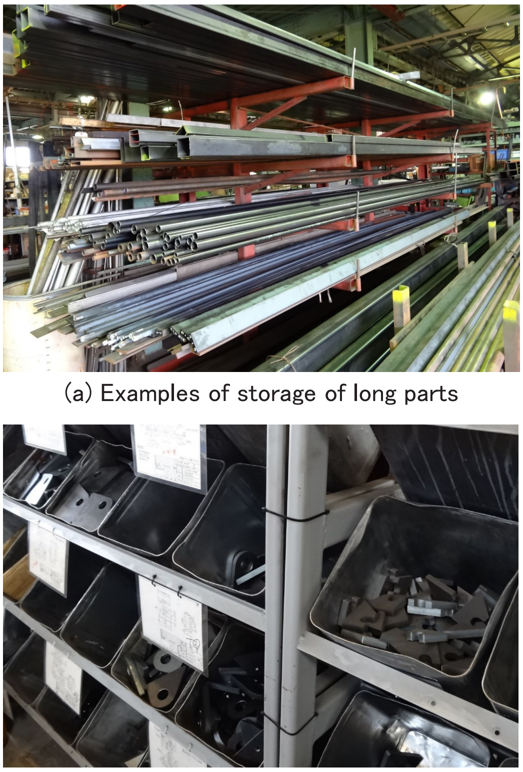
Now, we have been asked to introduce the inventory management system for the purpose. The inventory control is the important function of the production management, and it aims to maintain the inventory quantity at the proper levels. In other words, the inventory levels of all the parts should be controlled such that the following can be achieved: the quantities of each parts are always more than the safety inventory, by which some problems can be dealt with to prevent the parts shortage; on the other hand, there is no excess inventory, which causes the increase of the production cost. So, the production volume is determined based on the production plan as follows.

$$P_i = R_i - I_i + A_i + S_i \tag{1}$$

Here, For each the parts i, P_i is the production volume, R_i is the required amount, I_i is the inventory volume, A_i is the already assigned inventory volume to the other production in I_i, and S_i is the safety inventory volume. In the case of Figure 4, as for parts Y, if R_Y is 80, I_Y is 50, A_Y is 30 and S_Y is 10, then the production volume P_Y becomes 70. Therefore, it is necessary to grasp the accurate inventory volume to determine the production volume. Otherwise, the shortage of parts occur.

However, grasping the accurate inventory volume is not easy in the real factory. The types of the parts are several hundreds, and the parts shelf are dispersed in various places of the factory to adapt to the individual work process. Fig. 6 shows the parts shelf examples as for the long parts and small parts. The long parts have to count from a particular direction. And, the small parts are stored in containers. So, it is necessary to take out them in order to count the exact quantities. In this way, it takes time to move among the parts shelf and to investigate the quantities. Actually, it was estimated to take a few days in the case of one person to inventory. Furthermore, there are the other problems: the office work efficiency is so reduced in the field; the actual inventory is fluctuated during the investigation.

Moreover, the parts are always moving in the factory, and the actual inventory volumes are always changing, too. Fig. 7



(a) Examples of storage of long parts

(b) Examples of storage of small parts

Figure 6: Parts shelf in target factory

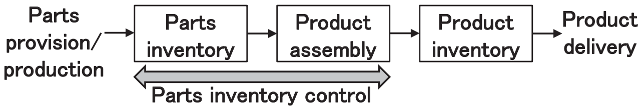


Figure 7: Product manufacturing process

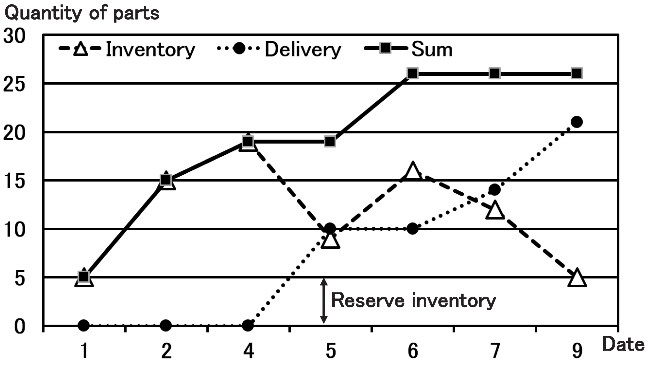


Figure 8: Change of quantity of parts inventory and delivery

shows the production process concerning with the parts. Depending on the parts type, the parts are manufactured in the company or procured from outside, and they are stocked in the parts shelf. The required parts are moved from the shelf to the assembly field at the assembly preparation stage of each product. The assembled products are stocked in the product inventory fields, then they are shipped on the designated date.

Incidentally, in the field of the production management in the large companies, the large scale production management system is introduced, such as the SAP[4], and the production information is managed as the integrated system including the inventory, accounting, order and so on. Also, the inventory volume is sometimes measured by using the RFID (radio frequency identifier) tags in the various field to reduce the data entry workload [2].

However, the target factory is the small and medium-sized company like most companies in Japan, of which proportion is said 99.7% [13], and it is pointed out that the introduction of such a management system is so less than the large company. As for this cause, two factors can be pointed out from the view point of their production scale. First, it is difficult to obtain the effect commensurate with the system investment. Second, it is difficult to reserve the full-time personnel for the system operations, grasping the data in the field and entering the data into the system. On the other hand, with the development of the e-commerce and supply chain management (SCM), it is becoming necessary to introduce the EDI (Electronic Data Interchange) with the large companies. As a result, it is also becoming necessary to introduce the production management system to manage the data for the EDI.

So, the requirement of the target inventory management system is as following. First, based on ”Order 1” in Fig. 5 and equation (1), the production volume must be calculated. For this purpose, the exact inventory volume I_i and assigned inventory volume A_i in equation (1) must be grasped. Second, from the viewpoint of the cost performance, the target system must be implemented without using expensive equipments and devices. Third, the system operations must be performed without increasing the workload of personnel.

3 PROPOSAL OF INVENTRY CONTROL METHOD

3.1 Composition of Proposal System

To achieve the requirements, we proposed the following inventory management system. As for the first requirement, the calculation of the production volume was based on general production management system to grasp each quantity of equation (1). Since the order information is received via EDI as the electronic data, it is possible to calculate the required quantity of the parts in time series by linking the MRP system. Then, the parts are stocked before a certain period of the shipping date. Then, they are moved to the assembly fields.

Fig. 8 shows the change of the quantity of a parts. Here, “Inventory” shows the quantity in the parts shelf; “Delivery” shows the quantity of the parts already moved to the assembly field; “Sum” shows the sum of the both. In this case, the parts are prepared in the parts shelf prior to assembly start 3 days, and the safety inventory volume is 5. For example, 10 parts are prepared (15 including the safety inventory) on second, and they are moved to the assembly field on fifth. Similarly, the parts are prepared 4 on fourth; 6 on sixth, and they are moved on seventh and ninth respectively. So, on second, R_i is 4; I_i is 15; A_i is 15; S_i is 5. Then, the production volume P_i is 4.

The quantity in the parts shelf shown in Fig. 8 is called the theoretical inventory, which is the value in the database.

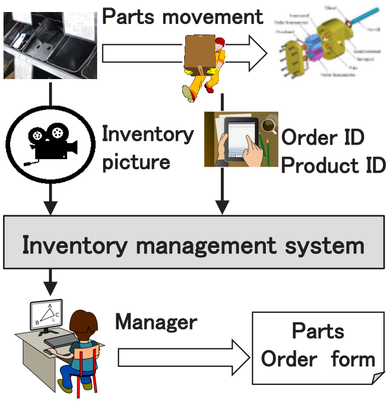


Figure 9: Composition of proposal system

However, in the actual field, since there are manufacturing loss and the process delay, it does not always equal to the actual inventory. Therefore, to perform the inventory control, the following two operations are generally needed.

Firstly, the both quantities of parts have to be grasped: parts in the parts shelf; moved parts to the assembly field. For example, in the case where the movement from the inventory shelf to the assembly on fifth is delayed to ninth in Fig. 8, the quantity in the parts shelf remains 19 as shown by "Sum". In the case where the inventory manager manages only the inventory shelf, he may misunderstand that there are sufficient parts inventory. Then, if he did not supply the parts, the quantity in the parts shelf would be under the safety inventory on ninth.

Secondly, the investigation to the actual inventory has to be performed. In other words, even if the theoretical inventory of the both can be grasped, the quantity in the parts shelf would be also under the safety inventory. For the example, in the case where 10 manufacturing loss happened on sixth. Therefore, there are the challenges to achieve the second and the third requirement.

As for this challenge, the IoT (Internet of Things) is spreading rapidly, and it has become possible to grasp the state of the actual inventory automatically by utilizing the various sensors such as surveillance cameras. That is, by utilizing it, the system can be implemented only by the cheap devices; the work load of the operators can be reduce. So, as shown in Fig. 9, we proposed the function to grasp the actual inventory volume, by which the picture of the parts shelf is saved in the database, and the inventory manager utilize it for the inventory management.

Here, the parts inventory is decreased only when the parts are moved to the assembly field, and this quantity can be calculated by the MRP system using the order information as shown in Fig. 8. And, the status of the parts shelf can be always grasped by querying these images. Particularly, to supply the parts is necessary only in the case where the quantity of the parts shelf is less than the quantity of the next movement except the safety inventory. That is, in the case where the quantity of the parts shelf is enough for the next movement, it is possible to determine the necessity of replenishment only by the image without measuring the exact quantity of the parts.

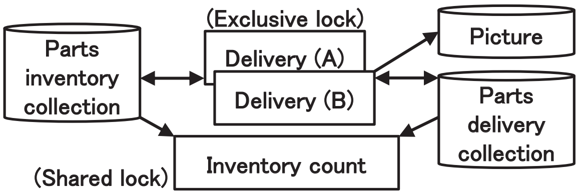


Figure 10: Transaction model for inventory management

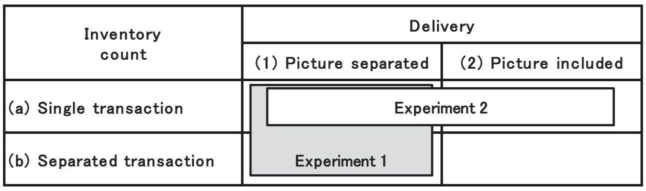
The business procedure by this feature is as follows. The worker takes out the parts from the parts shelf, then take the picture of this shelf by his hand-held camera. And, he enters it into the database with the parts data: the order and product ID. The system calculates the theoretical inventory in the parts shelf and the total quantity that has been moved to the assembly field as shown in Fig. 8. The inventory manager queries the database in the office to determine the necessity of replenishment by using the production plan, theoretical inventory and image data. In the case where the actual inventory is insufficient, he indicates the procurement or the manufacturing of the parts. By this procedure, the inventory manager can perform his business in the office by using the picture without the field work.

3.2 Inventory Mnagement Model

As for this inventory management system, its database has to treat not only the character and numerical data of the inventory information, but also the image or video data in the factory. So, if this system is implemented by using the RDBMS, the significant restriction occurs on the data manipulation. For this reason, we use MongoDB for the database of this system. As for MongoDB, "gridFS" was prescribed, which is the convention to store the huge data, and the official drivers support this. In this convention, the huge data is stored into the *collections* called "bucket", it can be accessed by the file name which was defined individually when it was stored.

We show the transaction model of the proposal inventory management system on MongoDB in Fig. 10. In the following, we simplify this model, that is, only the necessary data fields to implement the transaction are extracted from the actual data fields. In Fig. 10 "Parts inventory" *collection* (below, $Parts\ inventory$) saves the quantity of each parts in the parts shelf; "Parts Delivery" *collection* (below, $Parts\ delivery$) saves {order_ID, product_ID, parts_ID, required_quantity, shortage_quantity, image_name}; "Picture" *collection* (below, $Picture$) saves {image_data}; In addition, as for $Parts\ delivery$ at the planning time, {order_ID, product_ID, parts_ID, required_quantity, shortage_quantity} is saved, and the value of {required_quantity} is also set to {shortage_quantity}. Incidentally, the file name of {image_data} in $Picture$ is saved in {image_name}.

$Delivery$ is the transaction which executes the movement processing of the parts from the parts shelf to the assembly field. And, in Fig. 10, there are transactions $Delivery(A)$ and $Delivery(B)$. They correspond to the product A and B in Fig. 5 respectively. That is, it reduces the parts quantity from $Parts\ inventory$ on the basis of the required amount for product A or B, and it saves the image after parts move-



Inventory count	Delivery	
	(1) Picture separated	(2) Picture included
(a) Single transaction		
(b) Separated transaction		

Figure 11: Program structure and experiment

ment in $Picture$. Also, it updates the {shortage_quantity} of $Parts\ delivery$ according to the movement of parts. Here, in order to process as a transaction, it performs the movement processing for each product unit of the order. For example, in the case of product A in Fig. 5, 60 parts X and 20 parts Y is moved. Then, if any parts is insufficient inventory, no movement is executed as shown in the following equation.

$$shoratage\_quantity = \begin{cases} 0 & (All\ parts\ supplied) \\ at\ plan & (otherwise) \end{cases}$$

Transaction $Inventory\ count$ calculates the total quantity of parts X, Y, Z in two *collections*: $Parts\ inventory$ and $Parts\ delivery$. This process is executed as single transaction for each parts.

The requirements to MongoDB in order to implement this model are as follows. First, the concurrency control of the transactions on two *collections* should be performed. That is, $Parts\ inventory$ and $Parts\ delivery$ are updated simultaneously, also they are queried. Concretely, while the transaction $Delivery$ is updating these *collections*, $Inventory\ count$ is querying these *collections* to calculate the total quantity of the parts as shown in Fig. 8. Here, both of $Delivery$ and $Inventory\ conut$ should be executed as transactions. That is, if the former queries the anomaly state of the parts such that one *collection* has already updated and another *collection* has not updated yet, incorrect quantity of the inventory is calculated. For this requirement, we applied the transaction feature shown in Section 2.2 to this system.

Second, the conflict due to huge data should be suppressed. In this method, the exclusive lock and shared lock are used for the update and query respectively. Therefore, the long latency due to the lock is expected when transactions access the huge image data. Here, the image data in $Picture$ is accessed by the file name saved in the *document* of $Parts\ delivery$. So, we adopt the following method: firstly, the image data is saved; then the transaction that accesses $Parts\ delivery$ is started. In addition, this method is used for the query transaction, too. So, as for the transaction $Inventory\ conut$, we omit the access to the image for the sake of simplicity.

4 IMPLEMENTATIONS AND EVALUATIONS

4.1 Implementation of Inventory Management Model

We implemented the inventory management model on a stand-alone PC. Its implementation environment is as follows: CPU is i7-6700 (3.41GHz); memory is 16GB; disk is SSD memory of 512GB; OS is Windows 10. We adopted MongoDB (Ver. 3.3.6) for the database; Java (Ver. 1.8.0_73) for programming; Mongo DB Java driver (Ver. 2.14.2) to access MongoDB from Java program. The above-mentioned three transaction programs are performed simultaneously using Thread class of Java, and the Java class shown in Section 2.1 was used to provide the transaction feature for their concurrency control. And, we used GridFS class of Mongo DB Java driver to store the image data to MongoDB.

Also, we implemented MongoDB update transaction by the following two methods in order to evaluate the deterioration of conflicts associated with saving the image data. The first method is shown in (2) of Fig. 11, and the image data is saved to MongoDB as a part of the transaction. Its procedure is as following. Firstly, in order to confirm whether there is sufficient stock in the inventory shelf, the transaction $Delivery$ query $Parts\ inventory$ by using the exclusive lock. So, the conflict between other transactions may occur henceforth. Then, it update {shortage_quantity} of the corresponding *document* in $Parts\ delivery$ to 0, and insert the image file name to {image_file}. Next, it reduces {storage_quantity} of $Parts\ inventory$, and save the image data to $Picture$. Finally, it executes the commit. Incidentally, we excluded the case of shortage of inventory in this experiment. In addition, we set the delay between the above mentioned access of the two *collections* in order to confirm the occurrence of the conflicts. Also, in the case where the conflict occurs, the transaction performs the retry after a certain waiting time.

The second method is similar to the first method, except that it saves the image data before updating the *document* in $Parts\ inventory$ and $Parts\ delivery$. In other words, since the image data is saved prior to the start of the transaction, any *document* is not locked by this data manipulation. So, the lock is executed only during the updates of two *collections* except $Picture$, and its time is shortened. We show this method in (1) of Fig. 11. As for the query transaction $Inventory\ count$, we also implemented it by the following two methods, in order to evaluate the difference between the execution as single transaction and as multiple transactions. Here, the latter corresponds to the MongoDB's method such as "findOne".

In the first method, in order to prevent the *collections* to be updated by other transactions during its query, it queries each *collection* by using the shared lock. And, based on the query results, it calculates the sum of the parts. In this way, after it completes the processing, it executes the commit. Then, after waiting for a certain time, the next transaction is started to query another parts.

The second method is similar to the first method, except that it executes the commit after querying $Parts\ inventory$; then it queries $Parts\ delivery$. That is, it separates the query processes into two transactions. We show these two methods in (a) and (b) in Fig. 11 respectively.

4.2 Experiments and Evaluations

We conducted the experiments by the implementation program of the inventory model, and evaluated the methods. The

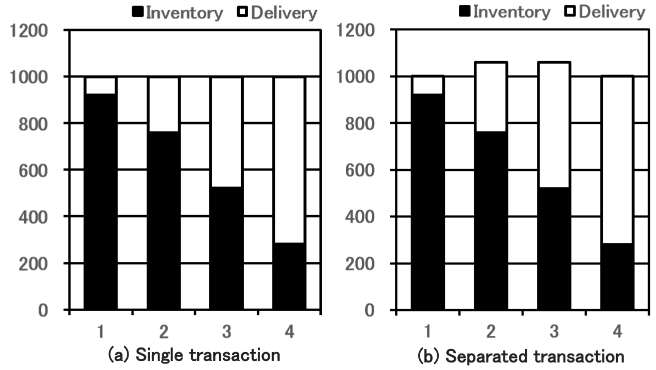


Figure 12: Result of experiment 1

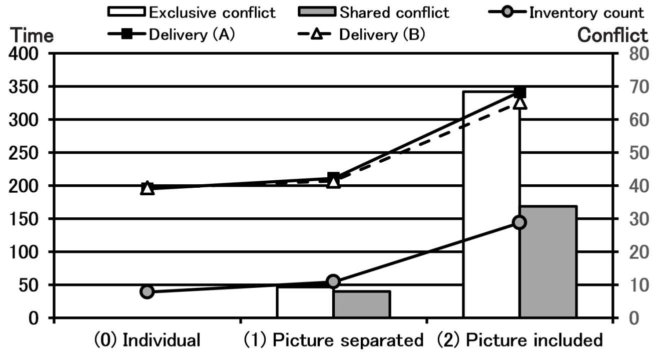


Figure 13: Result of experiment 2

Setting and procedure of the experiments are as follows. We set the sufficient inventory quantity of each parts as the initial value of $Parts\ inventory$. Also, we saved enough order data into $Parts\ delivery$.

Then, we started the three transactions in Fig. 10 at the same time: two update and one query transactions. As for the update transactions, we set 100msec for the delay time between update the two *collections*. we also set 100msec for the delay of next transaction start. As for the query transaction, we also set 250msec for the delay of next transaction start. And, for both transactions, we set 50msec for the delay before the retry when the conflict occurs. In the experiment, we executed each update transaction 12 times, and the query transaction 14 times. We used the same image data for every transaction in order to simplify the evaluation. And, its size is 3.3MB.

In experiment 1, we conducted comparative evaluation between the methods (a) and (b) in (1) of Fig. 11, and we show its results in Fig. 12. As for (a), the total quantity of parts is always constant. On the other hand, as for (b), since the query process was separated into two transactions, the halfway state could be queried. That is, though the query result of $Parts\ delivery$ is as of after the movement, the result of $Parts\ inventory$ can be as of before the movement. So, the anomaly occurred, in which the query result of the total quantity of parts is increased depending on the query timing.

As a result, it was confirmed that the isolation of the transactions is maintained by this method. That is, the halfway state of update is concealed from the other transactions.

In experiment 2, we conducted comparative evaluation between the methods (1) and (2) in (a) of Fig. 11. In addition, we performed experiments three times for each case, and Fig. 13 shows the average of these results. Prior to this experiment, we measure the individual elapsed time of $Delivery$ and $Inventory\ count$. In this case, only one transaction is executed at the same time, and there is no conflict. We show this result in (0) of Fig. 13. In addition, the line graph shows the change of the elapsed time for each transaction; the bar graph shows the number of the conflicts occurred for exclusive lock and shared lock respectively.

(1) of Fig. 13 shows the experimental result of the method of (1) in Fig. 11. That is, the image data is saved before the transaction start. As a result of this case, the number of each conflict was about 10; the increase of elapsed time from (0) was about 10%. On the other hand, as shown in (2), in the case where the image data is stored as a part of the transaction, the number of the exclusive and shared conflict was about 70 and 30 respectively; the elapsed time of $Delivery$ became 1.7 times longer than (0); $Inventory\ count$ became 2.7 times longer than (0).

5 DISCUSSION

As shown in Section 2.1, we had implemented the method to update plural *documents* of “single” *collection* in MongoDB as transaction with maintaining the ACID properties. In this paper, we showed this method is also valid in the case where the *documents* of “plural” *collections* are updated as the single transaction. Incidentally, in this method, each transaction can be performed with the designated isolation levels shown in Table 1. So, as for $Inventory\ count$ in Fig. 10, in the case where the business requirements do not need the accurate quantity, the query can be efficiently executed with the isolation level READ UNCOMMITTED.

On the other hand, as for the NoSQL databases, there is a discussion about the presence or absence of needs of their transaction feature. Therefore, we consider that this model shows one case, of which it is necessary that MongoDB equips the transaction feature. Firstly, the inventory management system proposed in this paper is based on the actual operations of a company. And, the requirements to the database are as follows: the huge data such as videos and pictures can be saved; the plural *collections* can be updated as single transaction with maintaining the ACID properties. The former is a major reason why MongoDB is selected rather than the RDBMS. Secondly, with the development of the IoT, the databases, which can handle the wide diversity of data, is expected to spread to the business systems. Along with this, we consider that the needs of transaction feature for the NoSQL databases as shown in this paper would grow more.

As for the transaction feature using the lock method, the frequency of conflict and the elapsed time of the transaction increase to save the huge data. For this problem, we could improve the efficiency of the transaction by performing the operation of huge data, such as videos and pictures, outside of the transaction. In the experiment, though we confirmed only the case of inserting the picture data, it is possible to perform the update or deletion of it in the same way. We show the

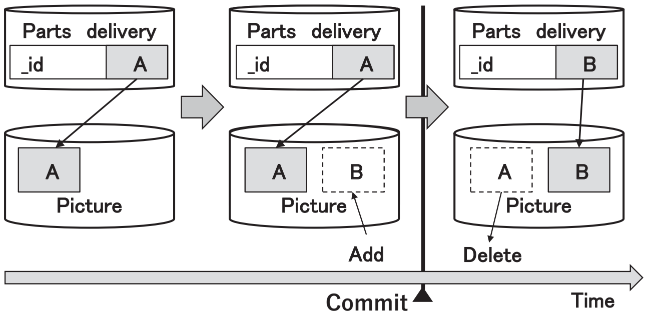


Figure 14: Access method for the huge data

procedure of the update in Fig. 14: firstly the picture data is inserted; then, *Parts delivery* is updated and the inserted picture data can be accessed through the updated data; finally, the previous picture data is removed.

In addition, there was the requirement to introduce this system, which is general for the small and medium-sized company: the system could be implemented at a low cost; the workload of the personnel should not increase. As for the former, there are various kinds of inexpensive input devices such as the wearable or surveillance cameras, and the cellular phones can be also used. And, the software was implemented using the free software, such as MongoDB and Java. As a result, it could be built without a big investment.

As for the latter, since to grasp the actual inventory volume took a few days as shown in Section 2.2, there is the problem that the inventory volume changed during this work. For this problem, we had proposed the following operation. Firstly, the snapshot of the inventory images such as pictures were saved, then the inventory volumes were grasped. As a result, the status of all the inventory at the certain point in time could be recorded efficiently. Therefore, we consider the inventory management utilizing image is effective.

Furthermore, we are planning the verification in the target inventory management business in the actual factory as the next stage. And, we expect that the wide range of the inventory management work shall become to be able to be performed in the office efficiently by utilizing both the production plan information and inventory images.

6 CONCLUSION

We had implemented the transaction feature for MongoDB, and shown the plural *documents* can be updated with maintaining the ACID properties. In this paper, we proposed the application of this feature to the production management system of the actual company, and extracted the requirements to database. As a result, we found that the database should update two *collections* as single transaction; it should manipulate the huge data such as image.

So, we implemented prototype of the inventory management system, and confirmed the following through the experiments. The former problem could be solved by the above mentioned transaction feature. The latter problem could be solved by the following procedure: the huge data is saved before the transaction; then, the update transaction is begun.

The future challenges are the verification of this method through the operations of the inventory management system.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 15K00161.

REFERENCES

[1] K. Banker, “MongoDB in Action,” Manning Pubns Co. (2011).

[2] M. Bertolini, et al., “Reducing out of stock, shrinkage and overstock through RFID in the fresh food supply chain: Evidence from an Italian retail pilot,”International Journal of RF Technologies, Vol. 4, No. 2, pp. 107–125 (2013).

[3] R. Cattell, “Scalable SQL and NoSQL data stores,” ACM SIGMOD Record, Vol. 39, No. 4, pp. 12–27 (2011).

[4] I.J. Chen, “Planning for ERP systems: analysis and future trend, Business process management journal,” Vol. 7, No. 5, pp. 374–386 (2001).

[5] J. Gray, and A. Reuter, “Transaction Processing: Concept and Techniques,” San Francisco: Morgan Kaufmann (1992).

[6] C.L. Iacovou, I. Benbasat, and A.S. Dexter, “Electronic data interchange and small organizations: adoption and impact of technology,” MIS quarterly, Vol. 19, No. 4, pp. 465–485 (1995).

[7] T. Kudo, M. Ishino, K. Saotome, and N. Kataoka, “A Proposal of Transaction Processing Method for MongoDB,” Procedia Computer Science, Vol 96, pp. 801–810 (2016).

[8] D. Laney, “3D Data Management: Controlling Data Volume, Velocity and Variety,” META Group, 2012, http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[9] H. Garcia-Molina, and K. Salem, “SAGAS,” Proc. the 1987 ACM SIGMOD Int. Conf. on Management of data, pp. 249–259 (1987).

[10] MongoDB Inc., “The MongoDB 3.0 Manual,” http://docs.mongodb.org/manual/.

[11] MongoDB Inc., “MongoDB API Documentation for Java,” http://api.mongodb.org/java/.

[12] E. Redmond, and J.R. Wilson, “Seven Databases in Seven Weeks: A guide to Modern Databases and The NoSQL Movement,” Pragmatic Bookshelf (2012).

[13] T. Sakamoto, “Making companies stronger through Financial Management,” Chuokeizai-sha, Inc. (2015) (in Japanese).

[14] K. Seguin, “The Little MongoDB Book” (2011), http://openmymind.net/mongodb.pdf.

[15] S.S. Sriparasa, “JavaScript and JESON Essentials,” Packt Pub. Ltd. (2013).