IWIN2015

International Workshop on Informatics

Proceedings of

International Workshop on Informatics

September 6-9, 2015

Amsterdam, Netherlands

Informatics Society

Sponsored by Informatics Society

IWIN2015

International Workshop on Informatics

Proceedings of

International Workshop on Informatics

September 6-9, 2015

Amsterdam, Netherlands

Informatics Society

Sponsored by Informatics Society

Publication office:
Informatics Laboratory
3-41, Tsujimachi, Kitaku, Nagoya 462-0032, Japan
Publisher:
Tadanori Mizuno, President of Informatics Society
ISBN:
978-4-902523-38-6
Printed in Japan

Table of Contents

Session 1: Network Systems

(Chair: Takuya Yoshihiro) (9:15 - 10:30, Sept. 7)

Session 2: Wireless Networks

(Chair: Tomoki Yoshihisa) (10:45 - 12:00, Sept. 7)

Keynote Speech 1

(13:00-13:45, Sept. 7)

Keynote Speech 2

(13:45-14:30, Sept. 7)

Session 3: Intelligent Transportation Systems

(Chair: Yoshitaka Nakamura) (15:15-16:30, Sept. 7)

Session 4: Data Analytics

(Chair: Tomoo Inoue) (16:45-18:00, Sept. 7)

Session 5: Networks, Applications and Web

(Chair: Yoshia Saito)(9:00-10:15, Sept. 8)

Session 6: Networks, Applications and Web

(Chair: Tomoya Kitani)(10:30-12:10, Sept. 8)

Keynote Speech 3

(14:00-14:45, Sept. 8)

Keynote Speech 4

(14:45-15:30, Sept. 8)

Session 7: Business Systems and Applications

(Chair: Kozo Okano)(16:00-17:40, Sept. 8)

Panel Session: Promising Techniques toward Future Intelligent Transport Technologies

(10:00-12:00, Sept. 9)

Chair

- Prof. Yoshimi Teshigawara, Tokyo Denki University, Japan

Panelists

- Prof. Ryozo Kiyohara, Kanagawa Institute of Technology, Japan
- Prof. Tomoya Kitani, Shizuoka University, Japan
- Prof. Takaaki Umedu, Shiga University, Japan

A Message from the General Chair

It is our great pleasure to welcome all of you to Amsterdam, Netherland, for the Ninth International Workshop on Informatics (IWIN 2015). This workshop has been held annually and sponsored by the Informatics Society. The first, second, third, fourth, fifth, sixth, seventh, and eighth workshops were held in Napoli, Italy, Wien, Austria, Hawaii, USA, Edinburgh, Scotland, Venice, Italy, Chamonix, France, Stockholm, Sweden, and Prague, Czech Republic, respectively. The first workshop was held in 2007. All of workshops were held in September.

In IWIN 2015, 23 papers have been accepted and 13 papers will be further selected as excellent papers that are considered having significant contributions in terms of the quality, significance, current interest among the professionals, and conference scope through the peer reviews by the program committees. Based on the papers, seven 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, data analytics, and groupware and education systems. In addition, IWIN 2015 has four invited sessions from Prof. Yasue Mitsukura of Keio University, Dr. Makoto IMAMURA, Mitsubishi Electric Corp., Mr. Shinichi BABA, Toshiba Europe Research, and Dr. Kazuya KOJIMA, Kanagawa Institute of Technology. We really appreciate the participation of the four invited speakers in this workshop.

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 Amsterdam.

Ryozo Kiyohara
General Chair of The International Workshop on Informatics 2015

Organizing Committee

General Chair

Ryozo Kiyohara (Kanagawa Institute of Technology, 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 (Iwate Prefectural University, Japan)

Ken-ichi Okada (Keio University, Japan)

Norio Shiratori (Tohoku University, Japan)

Osamu Takahashi (Future University-Hakodate, Japan)

Program Chair

Takuya Yoshihiro (Wakayama University, Japan)

Financial Chair

Tomoya Kitani (Shizuoka University, Japan)

Publicity Chair

Yoshitaka Nakamura (Future University-Hakodate, Japan)

Program Committee

Behzad Bordbar (University of Birmingham, UK)
Teruyuki Hasegawa (KDDI R&D Laboratories, Japan)
Tomoo Inoue (University of Tsukuba, Japan)
Yoshinobu Kawabe (Aichi Institute of Technology, Japan)
Tomoya Kitani (Shizuoka University, Japan)
Tsukasa Kudo
(Shizuoka Institute of Science and Technology, Japan)
Hiroshi Mineno (Shizuoka University, Japan)
Jun Munemori (Wakayama University, Japan)
Yoshitaka Nakamura (Future University-Hakodate, Japan)
Masashi Saito (Mitsubishi Electric Corporation, Japan)
Yoshia Saito (Iwate Prefectural University, Japan)
Fumiaki Sato (Toho University, Japan)
Toshifusa Sekizawa (Osaka Gakuin University, Japan
Hiroshi Shigeno (Keio University, Japan)
Yoh Shiraishi (Future University-Hakodate, Japan)
Takaaki Umedu (Osaka University, Japan)
Hirozumi Yamaguchi (Osaka University, Japan)
Tomoki Yoshihisa (Osaka University, Japan)
Takaya Yuizono
(Japan Advanced Institute of Science and Technology)
Tomoyuki Yashiro (Chiba Institute of Technology, Japan)
Kozo Okano (Shinshu University, Japan)
Noriko Otani (Tokyo City University, Japan)
Katsuhiko Kaji (Aichi Institute of Technology, Japan)
Hiroshi Ishikawa (Tokyo Denki University, Japan)
Yuichi Bandai (Kanagawa Institute of Technology, Japan)
Hiroaki Morino (Shibaura Institute of Technology, Japan)
Koji Tsukada (Wakayama University, Japan)
Akira Uchiyama (Osaka University, Japan)
Yoichiro Igarashi (Fujitsu Labolatory, Japan)

Session 1:
Network Systems
(Chair : Takuya Yoshihiro)

A Data Segments Scheduling Method for Streaming Delivery on Hybrid Broadcasting Environments

Tomoki Yoshihisa

Cybermedia Center, Osaka University, Japan
yoshihisa@cmc.osaka-u.ac.jp

Abstract - In hybrid broadcasting environments, the clients play the streaming data such as video or audio while receiving them. When the data reception is later than the time to start playing the data, the interruption of playing the data occurs. Although some methods to reduce the interruption time have been proposed, these methods have a large drawback that the server often broadcasts segments which many clients have already received. Hence, I propose an interruption time reduction method that schedules some segments. By broadcasting segments according to the schedule, the sever broadcasts the segments that many clients have not yet received and reduces the interruption time effectively.

Keywords: Video-on-Demand, Interruption Time, Broadcast Schedule, Continuous Media

1 INTRODUCTION

Due to the recent high interest on hybrid broadcasting environments, streaming delivery on the environments attracts great attention. In streaming delivery on hybrid broadcasting environments, the servers deliver streaming data such as video or audio both from broadcasting systems and communication systems. In broadcasting systems, e.g., TV and radio, the servers broadcast data according to predetermined broadcast schedules and can deliver data to many clients concurrently. In communication systems, e.g., the Internet, the clients can receive their desired data by requiring them to servers directly at arbitrary timings. Hybrid broadcasting environments of these systems are effective for streaming delivery since the clients can receive data both from broadcasting systems and communication systems.

In streaming delivery on hybrid broadcasting environments, the clients play the streaming data while receiving them from both systems. When the data reception is later than the time to start playing the data, the interruption of playing the data occurs. A shorter interruption time is preferable for the viewers to enjoy watching the data. So, some methods to reduce the interruption time for streaming delivery on hybrid broadcasting environments have been proposed ([1]-[6]). Here, interruption time indicates the elapsed time between the time to request playing the data and the time to start playing the data, and also indicates the elapsed time that the playing of the data is interrupted.

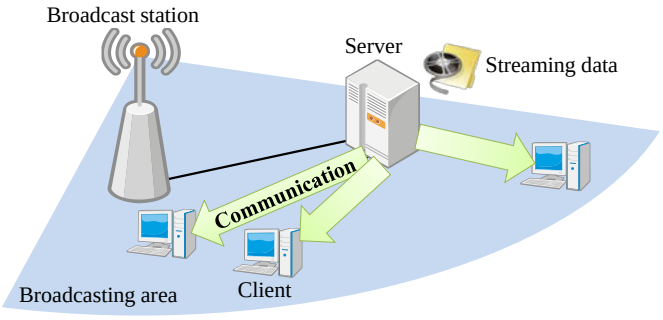


Figure 1: A hybrid broadcasting environment

In previous methods, the data is divided into some segments with fixed data sizes and the server broadcasts the data segment that is required by the client with the shortest *margin time*, i.e., the margin between the current time and the time to occur the next interruption. However, this approach causes the following drawback. The server broadcasts the first segment when a new client requests playing the data since the client with the shortest margin time is the new client and the clients required the first segment to start playing the data. Accordingly, the server often broadcasts segments which many clients have already received since the clients that have started playing the data have always received the first segment (the detail is described in Section 3.2). The server can reduce the interruption time effectively by broadcasting segments which many client have not yet received.

Hence, in this paper, I propose an interruption time reduction method that schedules some segments [1]. By broadcasting segments according to the schedule, the server does not always broadcast only the first segment even when a new client requests playing the data. So, the sever broadcasts the segments that many clients have not yet received and can reduce the average interruption time when there are many clients. The evaluation reveals that the proposed methods can reduce the average interruption time further than conventional methods.

The rest of the paper is organized as follows. Section 2 explains related work. The proposed method is presented in Section 3 and evaluated in Section 4. Finally, I conclude the paper in Section 5.

2 RELATED WORK

Some methods to reduce interruption time have been proposed ([7]-[12]). First, hybrid broadcasting environments

[1] This research was supported in part by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications, Grant-in-Aids for Scientific Research (B) numbered 15H02702, and Grant-in-Aids for Challenging Exploratory Research numbered 26540045.

are explained. Next, some methods for streaming delivery on hybrid broadcasting environments are introduced.

2.1 Hybrid Broadcasting Environments

Figure 1 shows the assumed hybrid broadcasting environment. The clients in the broadcasting area can receive data from the broadcasting system. Also, they can require their desirable 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.

2.2 Methods to Reduce Interruption Time

In UVoD (Unified Video-on-Demand) method proposed in [2], the server broadcasts the streaming data cyclically via each broadcast channel. By delaying the time to start the broadcast cycles for each broadcast channel, the clients can get more opportunities to receive the data. When an interruption will occur, the client tries to receive the data that cause the interruption directly from the server via the communication system.

In SSVoD (Super-Scalar Video-on-Demand) method proposed in [3], the server broadcasts the data in the same way as UVoD. Different from UVoD, the server does not send the required data to the clients until some clients require the same data. After the server receives some requirements, the server multicasts the required data to them.

In NBB VoD (Neighbors-Buffering Based Video-on-Demand) method proposed in [4], the server broadcasts the data in the same way as UVoD, but uses a P2P approach. The clients receive their desired data from other clients that have already received the data. If any clients do not have the data, the server sends them to the clients. In the above methods, however, the server broadcasts the whole data repeatedly although the clients can receive some parts of the data from the communication system.

Hence, FC (First-segment from Communication), MC-LB (Middle-segment from Communication and Last-segment from Broadcast), and MC-LC (Middle-segment from Communication and Last-segment from Communication) methods have been proposed in [5]. These methods predict the data that the clients receive from the communication system and eliminate the predicted data from the broadcast schedule. These three methods are different in the eliminated data. However, the broadcast schedule is static and the methods do not consider the data that the clients have already received.

In SET-C (Shortest Extra Time per Client) method proposed in [6], the server determines the data to broadcast dynamically considering margin time. As explained in Section 1, the margin time is the time between the current time and the time to occur the next interruption, and is calculated from the data that the clients have already received. The proposed method in this paper also uses the margin time. In Figure 2, the vertical red line indicates the current time and the colored area indicates the data that the client has already

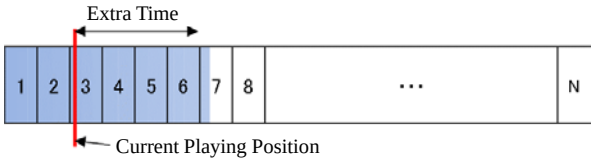


Figure 2: An example for extra time

received. In this case, the client is receiving the 7th segment and the margin time is the time until starting playing the segment. The method can reduce the interruption time since the probability that the client avoid the interruptions increases by broadcasting the segment that the client with the shortest margin time requires. In SET-C method, however, the server broadcasts the first segment when a new client requests playing the data since the server determines the next segment to broadcast every finishing broadcasting blocks. Accordingly, the server often broadcasts segments which many clients have already received since other clients have received the first segment.

3 PROPOSED METHOD

This section explains the proposed interruption time reduction method. First, the assumed system environments are explained. After that, the data delivery for broadcasting systems and for communication systems are respectively explained.

3.1 Assumed System Environments

This research assumes streaming delivery on hybrid broadcasting environments explained in Section 2.1. Since the server has many streaming data and it is difficult to predict which data the clients play, they do not receive data before they request playing data. The clients play streaming data from the beginning to the end continuously without fast-forwarding and rewinding. Their storage capacity is larger than the data size of their requested streaming data. The broadcast station uses one broadcast channel to broadcast one streaming data.

3.2 A Main Problem of Existing Method

In previously proposed methods, the merit of the broadcasting system, i.e., concurrently delivers the same data to multiple clients, works more effectively as the clients of that received data is the same increases. However, as explained in Sections 1 and 2, the server broadcasts the first segment when a new client requests playing the data since the client with the shortest margin time is the new client. Here, the term *new client* means that the client that starts playing the data from the beginning of the streaming data. Accordingly, the server often broadcasts segments which many clients have already received since other clients have the first segment. Figure 3 shows an example. In the figure, Clients 1-3 have received the preceding 7 segments. The black vertical lines indicate the playing positions of each client. This is the situation that the clients request playing the data at different times and the playing positions differ among

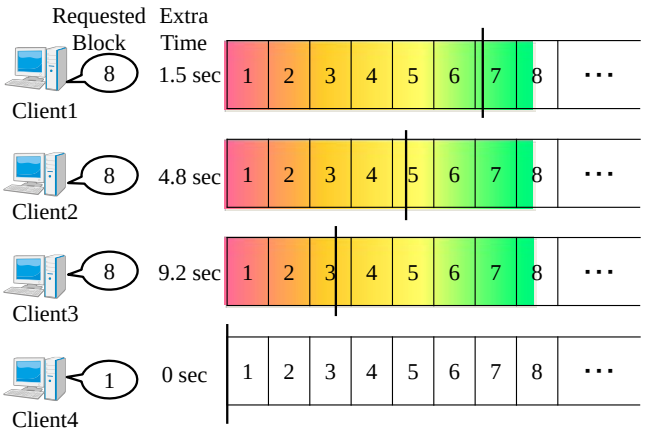


Figure 3: A figure to explain the problem

the clients although the received data is the same. Here, suppose the case when the new client Client 4 requests playing the data. Clients 1-3 are receiving the segment 8 and their margin times are respectively 1.5, 4.8, and 9.2 seconds. The margin time of Client 4 is 0 seconds since the client does not receive any segments. In this case under previously proposed method, the server broadcasts the segment 1 since the margin time of Client 4 is the shortest and Client 4 requires the segment 1. But, other clients have already received segment 1 and the server cannot exploit the merit of the broadcasting system. One of the solutions is that the server does not broadcast the segment that the new client requires, to broadcast the segment that many clients have not yet received.

3.3 G-SET-C (Grouped SET-C) Method

By broadcasting the segment that the new client requires after broadcasting some segments, the probability that the server broadcasts the segment that some clients have not yet received increases. Hence, the proposed method called G-SET-C method schedules some segments.

3.3.1. Delivery on Broadcasting Systems

In G-SET-C method, the server schedules G segments when creating the broadcast schedule. The server creates the next broadcast schedule when finishing broadcasting all scheduled segments. The server creates the broadcast schedule that includes the segments that the clients with a shorter margin time requires. For this, the server predicts the margin times for all clients.

Let $S(g)$ denote the times to start broadcasting the g th segment (g = 1, ..., G) included in the broadcast schedule, and E_i (g) denote the predicted margin time of the client i. The server can get the actual margin time E_i (1) since the server creates the broadcast schedule at S(1) and this is the current time. Let R_i (g) denote the time for the client i to finish receiving the required segment at $S(g)$. The time needed to broadcast one segment is B_b and the duration for playing one segment is P_b.

First, when $S(f+1) < R_i(g)$ (f = 1, ..., G-1), that is, the client i cannot finish receiving the required segment until the time to start broadcasting f + 1 th scheduled segment, the margin time at $S(f+1)$ decreases by the duration for playing one segment. So, $E_i(f+1) = E_i(f) - B_b$. If this is a negative value, $E_i(f+1) = 0$. Next, suppose the case when $R_i(f) < S(f+1)$, that is, the client i can finish receiving the required segment until the time to start broadcasting f th scheduled segment. When an interruption occurs until the time to finish the reception ($E_i(f) < R_i(f) - S(f)$), the client restarts the playing the data after the reception. So, $E_i(f+1) = R_i(f) + P_b - S(f+1)$. This is always positive value since $B_b < P_b$ in this research. Otherwise ($R_i(f) - S(f) < E_i(f)$), the margin time increases by the duration for playing one segment. So, $E_i(f+1) = E_i(f) - B_b + P_b$. Hence,

$$E_i(f+1) = \begin{cases} 0 & (S(f+1) < R_i(f), E_i(f) < B_b) \\ E_i(f) - B_b & (S(f+1) < R_i(f), E_i(f) > B_b) \\ R_i(f) + P_b - S(f+1) & \\ \quad (S(f+1) > R_i(f), E_i(f) < R_i(f) - S(f)) & \\ E_i(f) - B_b + P_b & \\ \quad (S(f+1) > R_i(f), E_i(f) > R_i(f) - S(f)) & \end{cases}$$

In the proposed G-SET-C method, the server schedules the segment that is required by the client j that satisfies the following equation for each g. N is the set of the clients. When some clients have the equivalent margin times, the server schedules the segment that is required by the client of that time to start playing the data is earlier.

$$E_j(g) = \min_{i \in N} E_i(g)$$

3.3.2. Margin Time Prediction

In G-SET-C method, to calculate the predicted margin time E_i (g) (g = 1, ..., G), the server needs to predict the time to finish receiving the segment D_i (g) that the client i requires at $S(g)$.

When the client receives D_i (g) from the broadcasting system, the server can calculate D_i (g) using $S(g)$ since the server grasps the time to start broadcasting each scheduled segment. For example, if D_i (g) is scheduled to the e th segment in the broadcast schedule, $R_i(g) = S(e) + B_b$.

When the client receives D_i (g) from the communication system, the server predicts D_i (g) using the communication bandwidth for the client i at the time to create the broadcast schedule, C_i. C_i is the bandwidth between the server and the client i. First, for D_i (1), the client i may have received a part of D_i (1) at that time. So, R_i (1) is given by the remaining data size divided by C_i. Next, for $D_i(f+1)$ (f = 1, ..., G-1), $R_i(f+1) = R_i(f)$ if $D_i(f+1)$ is the same segment $D_i(f)$. Otherwise, $R_i(f+1)$ is the value adding the segment size divided by C_i to $R_i(f)$.

3.3.3. An Example of Broadcast Schedule Creation

I will show an example of broadcast schedule creation using a simulation result. The simulated situation is shown in Table 1. In this situation, at the time 1982.899 seconds after the beginning of the simulation, the client 395 is playing the segment 22 and the segment that the client does not have and is the closest to the current playing position is the segment 23.

Table 1: Example of scheduling segments

Time [seconds]	Client ID	Playing Segment	Closest Not-Received Segment	Margin Time [seconds]
1982.899	394	22	23	0
	395	22	23	0.445
	396	2	3	0
	397	1	2	0
	398	1	2	0
	399	0	1	0
1984.149	394	24	25	0.250
	395	24	25	0.250
	396	5	7	1.000
	397	3	4	0
	398	3	4	0.125
	399	2	4	0.750
1986.649	394	29	40	5.375
	395	29	30	0.500
	396	8	10	0.750
	397	6	10	1.875
	398	6	10	1.875
	399	6	10	1.875

The margin time is the time between the current time and the time to start playing the segment 23 and is 0.445 seconds. In this example, $G = 2$.

The server schedules the segment 23 as the first scheduled segment since $E_i(1) = 0$ (i = 394, 396, 397, 398, 399) and the server schedules the segment that the client of that time to request playing the data is the earliest when the margin times are equivalent. Here, again, $E_i(g)$ (g = 1, ..., G) is the predicted margin time for g th scheduled segment. Next, the server predicts the margin times of all clients at the time to start broadcasting the second scheduled segment to determine the second segment to be included in the broadcast schedule. The margin times of the clients 394 and 395 are the value of the time to play a segment P_b = 0.469 seconds subtracted by the time to broadcast a segment B_b = 0.125 seconds since they were required the segment 23. So, $E_{394}(2) = 0.344$ seconds and $E_{395}(2) = 0.789$ seconds. In this simulation, the time for the client 396 to finish receiving the segment 3 from the communication system $R_{396}(1) = 1982.993$ seconds and the margin time $E_{396}(2) = 1982.993 + 0.4690 - 1983.024 = 0.438$ seconds. The time for other clients to finish receiving their required segments are later than $S(2)$ and $E_j(2) = 0$ (j = 397, 398, 399). Therefore, the server schedules the segment 2 that is required by the client 397 as the second scheduled segment.

The time 1982.899 seconds is immediately after the client 399 requests playing the data and the client requires the segment 1 at that time. At the time 1984.149 seconds, the segments that the client 399 has received are the same as those of the clients 397 and 398, and they are requiring the segment 4. Also, at the time 1986.649 seconds, the segments those have been received by the client 396-399 are the same and they are requiring the segment 10. In such a situation, the server concurrently satisfies their requirements by broadcasting the segment 10. In this way, the merit of the broadcasting system works well by scheduling segments considering the margin time.

Table 2: Simulation parameter values

Item	Value
Duration for Streaming Data	25 minutes
Bit Rate	2 Mbps
Broadcast Bandwidth	8 Mbps
Clients' Communication Bandwidth	1 Mbps
Server's Communication Bandwidth	30 Mbps
Segment Size	125.012 KBytes
Header Size	12 Bytes

3.3.4. Delivery on Communication Systems

As the same as the previously proposed method, the clients start receiving blocks from the communication system when they request playing the data. The clients receive the block that satisfying the following conditions and they require the next block when they finish receiving each block.

- The block that can cause interruptions if the client wait for the broadcasting of the block.
- The block that can be received faster than the reception from the broadcasting system.
- The block that is closer to the current playing position.

If there are no blocks that satisfying these conditions, the clients do not receive the blocks from the communication system to avoid the redundant communication.

4 EVALUATION

This section show some simulation results to evaluate the proposed G-SET-C method.

4.1 Evaluation Environments

Table 2 shows the evaluation parameter values. The streaming data is assumed to be an MPEG2 encoded (2 Mbps) movie data for 25 minutes. The segments consist of GOPs (Group of Pictures) and the data size is the same as the general GOP data size (0.5 seconds). The broadcast bandwidth is 8 Mbps assuming that the broadcasting system is a terrestrial broadcasting system. The communication bandwidths for all clients are equivalent. If the total communication bandwidth for the clients exceeds the server's communication bandwidth, the server's communication bandwidth is equally divided into each client. The header included the information for the identifiers for the streaming data and segments, and the number of the segments. The data size for each information is 4 bytes and the header size becomes 4 x 3 =12 bytes. G indicates the number of the scheduled segments is shown.

4.2 Comparison Methods

The original SET-C method is similar to G-SET-C method when $G = 1$. Other comparison methods are explained below.

- BCD-BE-AHB (Broadcast and Communication based Delivery-BE-AHB) Method

  This is the method that applies BE-AHB method proposed in [12] to the hybrid broadcast environments. The

data delivery on the broadcasting system is similar to the original one. The data delivery on the communication system under this method uses the same algorithm with that under the proposed method. The broadcast schedule is static in this method though that under the proposed method is dynamic.

・G-MRB (Grouped-Most Requested Block) Method

In this method, the server schedules the top *G* segments required by more clients. When the number of the clients that require the same segment is equivalent, the server schedules the segment that is required by the client that starts playing the data earlier\. When the number of the required segments is less than *G*, the server broadcasts the all required segments. And after that, the server creates the next broadcast schedule.

・G-LTIT-C (Grouped-Longest Total Interruption Time per Client) Method

In this method, the server schedules the segments those are required by the client of that interruption time is the longest at the time to broadcast each scheduled segment. Let I_i (g) denote the predicted interruption time for the client *i* at the time to broadcast the *g* th scheduled segment (g = 1, ..., G), i.e., $S(g)$. The server can get I_i (g) by asking the interruption time to each client. As the same discussion with Section 3.3.1, I_i (f + 1) (f = 1, ..., G-1) is given by the following equation.

$$I_i(f+1) = \begin{cases} I_i(f) + B_b - E_i(f) & (S(f+1) < R_i(f), E_i(f) < B_b) \\ E_i(f) & (S(f+1) < R_i(f), E_i(f) > B_b) \\ I_i(f) + R_i(f) - S(f) - E_i(f) & \\ \qquad (S(f+1) > R_i(f), E_i(f) < R_i(f) - S(f)) & \\ I_i(f) \quad (S(f+1) > R_i(f), E_i(f) > R_i(f) - S(f)) & \end{cases}$$

In G-LTIT-C method, the server schedules the segment that is required by the client *j* that gives the maximum I_i (g)

4.3 Interruption Time

The simulated interruption times for each client are shown in Figure 4. Simulated average arrival intervals are 1, 30, or 60 seconds. The horizontal axis is the client ID that is given along with the request time for playing the data and the vertical axis is the interruption time. We can see that the interruption time has an upper limit though has some dispersion. Hence, I use the average interruption time as an evaluation criteria.

4.4 Influence of the Number of the Scheduled Segments

The time to create broadcast schedule depends on the number of the scheduled segments *G*. the segment 1. Hence, we calculate the average interruption time changing *G*.

Figure 5 shows the average interruption time when the average request arrival interval is 30 seconds. In the figure, the proposed G-SET-C method gives the shortest average interruption time for all cases. This is because the server does not always broadcast the segment that the new client requires by scheduling some segments.

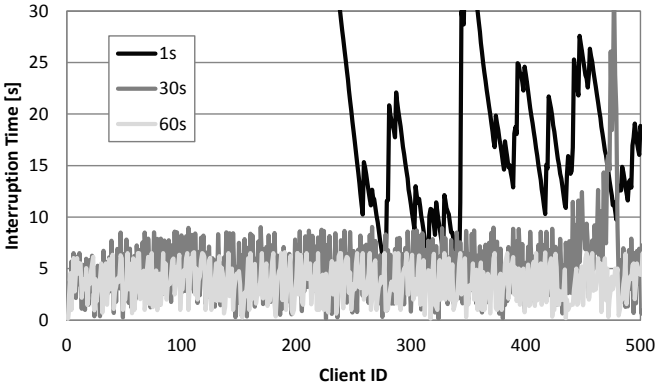


Figure 4: Clients' interruption time

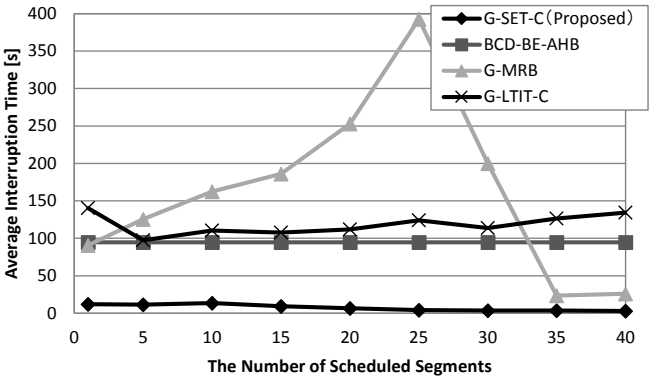


Figure 5: Average interruption time

The average interruption time under G-MRB method differs largely. This is because the probability that some clients require the same segment decreases since the average request arrival interval is long. The influence of the number of the clients that require the same segment is large for G-MRB method compared with other methods since the method considers the number of the required segments. When *G* is less than 25, the average interruption time increase as *G* increases. This is because the interval of creating the broadcast schedule lengthens as *G* increases. However, when *G* is larger than 25, the interruption time decreases since the server can broadcast many segments that are required when the time to create broadcast schedule.

5 CONCLUSION

In this paper, I proposed a segments scheduling method for streaming delivery on hybrid broadcasting environments. In the proposed method, the server schedules some segments considering the margin time to occur the next interruptions. By broadcasting segments according to the schedule, the sever broadcasts the segments that many clients have not yet received and reduces the interruption time when there are many clients. The evaluation revealed that the proposed methods could reduce the average interruption time further than conventional methods in most cases.

In the future, I am planning to propose the method considering the stop of playing the data or apply the P2P technique for the data delivery on the communication system.

REFERENCES

[1] V. Gopalakrishnan, B. Bhattacharjee, K. Ramakrishnan, R. Jana, and M.K. Vernon, CPM: Adaptive Video-on-Demand with Cooperative Peer Assists and Multicast, IEEE INFOCOM 2009, pp. 91-99 (2009).

[2] J.Y.B. Lee, UVoD: An Unified Architecture for Video-on-Demand Services, IEEE Communication Letters, Vol. 3, No. 9, pp. 277-279 (1999).

[3] J.Y.B. Lee and C.H. Lee, Design, Performance Analysis, and Implementation of a Super-Scalar Video-on-Demand System, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, Issue 11, pp. 983-997 (2002).

[4] 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).

[5] M. Umezawa, T. Yoshihisa, T. Hara, and S. Nishio, Interruption Time Reduction Methods by Predicting Data Reception for Streaming Delivery on Hybrid Broadcasting Environments, Proc. IEEE Pacific Rim Conference Communications, Computers and Signal Processing, pp. 185-190 (2011).

[6] T. Yoshihisa, Dynamic Data Broadcasting Methods for Streaming Delivery on Hybrid Broadcasting Environments, Proc. International Workshop on Advances in Data Engineering and Mobile Computing (to appear, 2015).

[7] S.W. Carter, J.F. Paris, S. Mohan, and D.D.E. Long, A Dynamic Heuristic Broadcasting Protocol for Video-on-Demand, Proc. IEEE International Conference on Distributed Computing Systems, pp. 657-664 (2001).

[8] D.L. Eager and M.K. Vernon, Dynamic Skyscraper Broadcast for Video-on-Demand, Proc. of International Workshop on Advances in Multimedia Systems, pp. 18-32 (1998).

[9] H. Kim and H.Y. Yeom, Dynamic Scheme Transition Adaptable to Variable Video Popularity in a Digital Broadcast Network, IEEE Transactions on Multimedia, Vol. 11, No. 3, pp. 486-493 (2009).

[10] J.B. Kwon. and H.Y. Yeom, Adjustable Broadcast Protocol for Large-scale Near Video-on-Demand Systems, Computer Communications, Vol. 28, No. 11, pp. 1303-1316 (2005).

[11] Q. Zhang and J.F. Paris, A Channel-based Heuristic Distribution Protocol for Video-on-Demand, Proc. IEEE International Conference on Multimedia and Expo, Vol. 1, pp. 245-248 (2002).

[12] T. Yoshihisa and S. Nishio, A Division-based Broadcasting Method Considering Channel Bandwidths for NVoD Services, IEEE Transactions on Broadcasting, Vol.59, Issue 1, pp. 62-71 (2013).

Design of an Application Based IP Mobility Scheme on Linux Systems

Kohei Tanaka†, Fumihito Sugihara‡, Katsuhiro Naito†, Hidekazu Suzuki*, Akira Watanabe*

†Faculty of Information Science, Aichi Institute of Technology, Toyota, Aichi 470-0392, Japan
{kohei, naito}@pluslab.org
‡Department of Electrical and Electronic Engineering, Mie University, Tsu, Mie 514-8507, Japan
*Graduate School of Science and Technology, Meijo University, Nagoya, Aichi 468-8502, Japan

Abstract -

Internet of Thing (IoT) systems have been attracting attention as one of the solutions for new services in the Internet. They usually employ a client-server model due to a difficulty of accessibility in practical networks. However, scalability will be a major issue in IoT systems because billion of IoT devices will be installed around the world in recent years. The authors have been proposed a new IP mobility mechanism called NTMobile (Network Traversal with Mobility) to realize end-to-end communication in IoT systems because end-to-end communication can improve system scalability by reducing traffic through servers. Conventional implementation employed a kernel module mechanism for NetFilter because the kernel module implementation is the best way of realizing high throughput performance. On the contrary, the kernel module should be maintained according to changes in NetFilter specifications. This paper designs an application based IP mobility scheme on Linux systems, where the developed IP mobility library can realize the IP mobility function in an application layer on Linux systems. As a result, developers can realize an end-to-end communication model by employing the enhanced IP mobility library. The proposed design ensures compatibility between the development library and the conventional NTMobile. Therefore, developers can select the implementation scheme according to the required performance.

Keywords: IP Mobility, Accessibility, Application library, Linux, NTMobile.

1 INTRODUCTION

Recent microcomputer boards implement some network interfaces to cooperate with another microcomputer boards[1], [2]. Almost all microcomputer boards are usually installed in a private network due to a security policy and limitation of assignable global IP addresses. Typical private networks prohibit accesses from the global Internet to a node in their private networks. Therefore, inter-connectivity is a big issue even if some applications should communicate with each other to realize their specific service. Additionally, operating systems select an interface to access to the Internet according to network condition of each interface and access policies when a device is a mobile node[3]. Therefore, a seamless connectivity scheme is also an important function.

IP mobility protocols are a solution to the requirement for inter-connectivity and seamless connectivity because they can realize continuous communication when an IP address for an interface changes due to switching of access networks[4]–[7]. They are classified into three types: IP mobility schemes for IPv4, IP mobility schemes for IPv6, and IP mobility schemes for IPv4 and IPv6. Mainstream of IP mobility schemes is for IPv6. On the contrary, the number of implementations for IPv4 is quite few though some mechanisms have been proposed [8], [9]. DSMIPv6(Dual Stack Mobile IPv6)[10] supports IPv4 and IPv6 networks. However, it still does not support the inter-connectivity between IPv4 and IPv6.

IPv4 is still the mainstream protocol under the present circumstances of the Internet. Therefore, private networks are usually used in practical IPv4 networks to reduce the number of required global IP addresses. Additionally, some Internet service providers start the service with large scale network address translator (LSN) in order to meet the shortage of IPv4 global addresses[11]. As a result, IP mobility in a private network behind NAT becomes an important issue.

The authors have developed a new IP mobility technology called NTMobile (Network Traversal with Mobility) [12]–[14]. The features of NTMobile are an IP mobility and an accessibility in both IPv4 and IPv6 networks. Therefore, each client of NTMobile can communicate with each other even when they use a different IP protocol version because they can communicate with virtual IP addresses that are independent addresses from physical IP addresses. NTMobile systems have some servers: account server (AS), direction coordinator (DC), notification server (NS), and relay server (RS). DC, RS and NS serve an IP mobility and an accessibility functions for each client, and AS serves an authentication service.

This paper proposes a new design of an application based IP mobility for NTMobile nodes. The original design can provide NTMobile functions by an application library instead of the special kernel module. Therefore, application programmers can obtain IP mobility and accessibility functions by using the proposed application library.

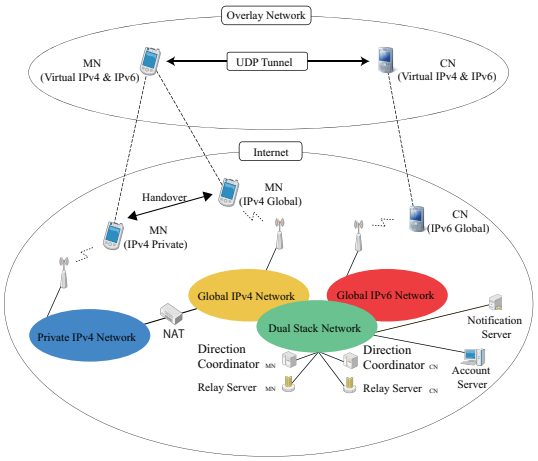


Figure 1: System model of NTMobile

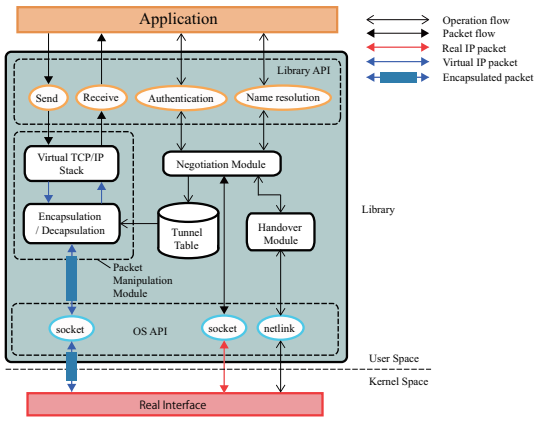


Figure 2: Overview of application library design

2 NTMOBILE

NTMobile can realize IP mobility and accessibility for IPv4 and IPv6 networks. Fig. 1 shows the overview of the NTMobile system model. The NTMobile system consists of an account server (AS), some direction coordinators (DCs) with some relay servers (RSs), notification servers (NSs) and NTMobile nodes. AS serves authentication service to all DCs, and each DC manages their RSs and NTMobile nodes. NTMobile node has IP mobility and accessibility functions by communicating with AS and its DC. Each DC has a virtual IP address pool for its NTMobile nodes, and assigns an address to each NTMobile node. Each NTMobile node constructs a UDP tunnel between NTMobile nodes according to a signaling direction from its DC, and communicate with each other by using their virtual IP addresses. As a result, each NTMobile node can communicate continuously even if real IP addresses at interfaces are changed because the virtual IP addresses are independent from physical IP addresses. The details of the system components are as followings.

2.1 Account Server (AS)

AS is an individual server that manages authentication information. Therefore, AS can distribute node information of each NTMobile node to initialize a setting for NTMobile nodes. Additionally, it bears responsibility for authentication by replying an authentication reply message when a DC makes inquiries about their NTMobile nodes authentication to AS. As a result, NTMobile nodes can get certified by AS through its DC.

2.2 Direction Coordinator (DC)

DC manages location information of each NTMobile node and indicates signaling processes for tunnel construction between NTMobile nodes. Each DC also owns the DNS(Domain Name Server) function. Therefore, DC can easily find another DC by searching a NS (Name Server) record in DNS. In addition, DC manages virtual IP addresses for NTMobile nodes, and assigns them to each NTMobile node without address duplication. In the tunnel construction process, it indicates the tunnel construction processes to both NTMobile nodes.

2.3 Notification Server (NS)

NTMobile typically uses a UDP protocol to communicate between a NTMobile node and its DC. Therefore, the NTMobile node should send keep-alive messages to its DC when it uses a private address under a NAT router. NS can provide a notification service with a TCP connection between NS and the NTMobile node, and can reduce the amount of messages for keep-alive.

2.4 Relay Server (RS)

The relay server function is to relay tunnels between NTMobile nodes when both NTMobile nodes exist under different NAT routers or exist under different version networks such as IPv4 and IPv6 networks. DC manages some relay servers to realize load balancing and to avoid a single point of failure. It also chooses a relay server to activate the relay function for dedicated NTMobile nodes.

2.5 NTMobile Node

Functions of NTMobile nodes are to realize IP mobility and accessibility in IPv4 and IPv6 networks. They obtain their own information from AS in the initialization phase, when they connect to the NTMobile network at first. Then, they inform their own network information to DC because DC should manage network information of each NTMobile node to realize IP mobility and accessibility. They also update the own network information when they switch access networks.

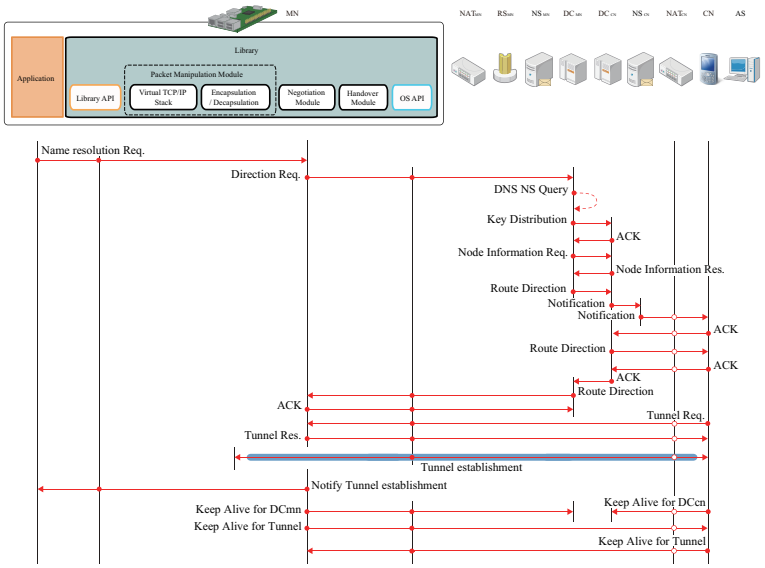


Figure 3: Signaling tunnel creation process of library

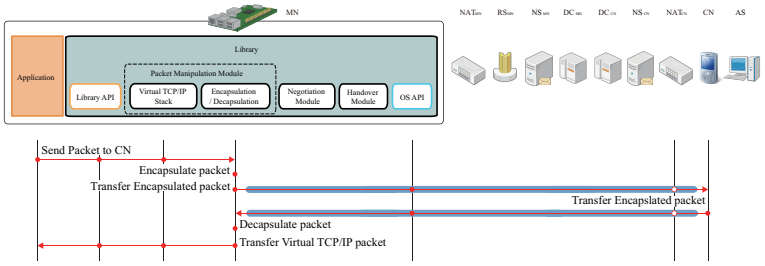


Figure 4: Signaling communication process of library

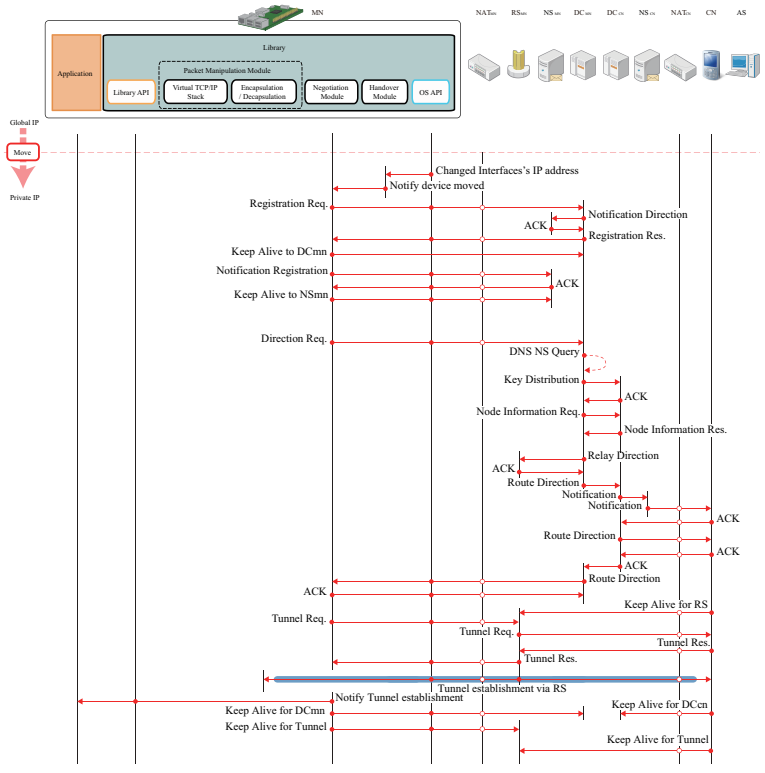


Figure 5: Signaling change network process of library

3 APPLICATION BASED IP MOBILITY

This paper proposes a new design of an application based IP mobility and accessibility for NTMobile nodes. The proposed design provides APIs (Application Programming Interface) for the NTMobile functions: an initialization, an authentication, a tunnel creation request, and network sockets for NTMobile communication. It also provides a virtual IP address for the network sockets. Our API provides BSD compatible network sockets for application developers. Therefore, application developers can easily implement their network application with virtual IP addresses for NTMobile network.

3.1 System Model

Fig. 2 shows the overview design of the proposed application based IP mobility and accessibility system for NTMobile nodes. The proposed application library is inserted between an original network socket in Linux OS and a developers' application. The application can request to the proposed application library for initialization of the NTMobile node functions, an authentication for NTMobile network, a tunnel creation request to a correspondent node. The actual data communication is performed with special network sockets for the NTMobile functions. The application library can perform IP capsulation/decapsulation and encryption/decryption processes between the special network sockets and the original Linux network sockets.

3.2 Library Modules

The proposed library consists of three main modules for negotiation, handover, and packet manipulation. The following is the functions of each module.

- Negotiation Module
  The negotiation module serves as a signaling function for NTMobile communication. Therefore, application developers do not care about the detail process of NTMobile communication because the negotiation module can handle IP mobility and accessibility functions in NTMobile. The negotiation module registers the specific information for packet manipulation into the tunnel table.

- Handover Module
  The handover module should check the network interface status to detect a change in the access network. It also handles reconstruction of a UDP tunnel to a correspondent node when IP address changes due to an access network change.

- Packet Manipulation Module
  The packet manipulation module consists of two main functions: a virtual TCP/IP stack and a capsulation function. The virtual TCP/IP stack provides a transport layer functions such as TCP and UDP to an application. The capsulation function handles the encapsulation and decapsulation for a UDP tunnel according to the information in the tunnel table.

3.3 Signaling Process

Figs. 3, 4 and 5 show the signaling processes for a tunnel creation in NTMobile. These figures are assumed that NTMobile node MN in a global IPv4 network requests to communication with NTMobile node CN in a private IPv4 network. Then, NTMobile node MN changes the network from the global IPv4 networks to a private IPv4 network. Direction coordinators $\mathrm{DC_{MN}}$ and $\mathrm{DC_{CN}}$ are the management servers for NTMobile node MN and CN respectively. The detail signaling process is as follows.

- Tunnel Creation

  Fig. 3 shows the signaling processes of the tunnel creation in NTMobile.

  1. The application calls the name resolution API to construct a UDP tunnel to NTMobile CN.
  2. The negotiation module requests the tunnel construction to $\mathrm{DC_{MN}}$ by transmitting the direction request message. The direction request message contains the FQDN of CN.
  3. $\mathrm{DC_{MN}}$ makes inquiries about NS record for the FQDN of CN since $\mathrm{DC_{CN}}$ is the domain name server and manages the domain for the FQDN of CN.
  4. $\mathrm{DC_{MN}}$ generates a shared key for encryption between $\mathrm{DC_{MN}}$ and $\mathrm{DC_{CN}}$ and distributes the shared key to $\mathrm{DC_{CN}}$.
  5. $\mathrm{DC_{CN}}$ replies the acknowledgement message to $\mathrm{DC_{MN}}$.
  6. $\mathrm{DC_{MN}}$ requests the information for CN to $\mathrm{DC_{CN}}$ by transmitting the NTM information request message.
  7. $\mathrm{DC_{CN}}$ replies the information about CN by replying the NTM information response message.
  8. $\mathrm{DC_{MN}}$ requests the tunnel construction to CN to $\mathrm{DC_{CN}}$ by transmitting the route direction message.
  9. $\mathrm{DC_{CN}}$ requests the notification to CN to $\mathrm{NS_{CN}}$.
  10. $\mathrm{NS_{CN}}$ sends the notification to CN based on the device type of CN.
  11. CN replies the acknowledgement message to $\mathrm{DC_{CN}}$.

12. DC_{CN} forwards the route direction message to CN.
13. CN replies the acknowledgement message to DC_{CN}.
14. DC_{CN} also replies the acknowledgement message to DC_{MN}.
15. DC_{MN} indicates the tunnel construction to CN to MN by transmitting the route direction message.
16. The negotiation module on MN replies the acknowledgement message to DC_{MN}.
17. CN transmits the tunnel request message to MN according to the direction from DC_{MN} because MN has a global IP address in this case.
18. MN replies the tunnel response message to CN to complete the tunnel construction process.
19. The negotiation module notifies the completion of the tunnel construction process to the application.
20. The negotiation module transmits a keep alive message to keep the NAT table on the route.
21. The application starts the communication through the constructed UDP tunnel. Then, it transmits packets by using special NTMobile socket.

- Data Communication

  Fig. 4 shows the signaling processes of data communication in NTMobile.

1. The virtual TCP/IP stack in the packet manipulation module serves the transport layer function to the application.
2. The encapsulation and decapsulation module performs the encapsulation the packet to CN.
3. The encapsulated packet is transmitted to CN.
4. The encapsulated packet is transmitted from CN.
5. The encapsulation and decapsulation module performs the decapsulation the packet from CN.
6. The virtual TCP/IP stack in the packet manipulation module sends the decapsulated packet to the application.

- Switching of Access Network

  Fig. 5 shows the signaling processes for switching the access network of MN.

1. Linux kernel can notify the change of network configuration through netlink socket.
2. The handover module notifies the negotiation module that the access network of the device has changed.
3. The negotiation module registers the new information about the access network to DC_{MN} by transmitting the registration request message.
4. DC_{MN} requests NS_{MN} to provide a notification service using TCP connection for MN.
5. NS_{MN} replies the acknowledgement message to MN.
6. DC_{MN} updates the network information of MN, and replies the registration response message.
7. MN registers the information of MN to NS_{MN}.
8. NS_{MN} replies the acknowledgement message to MN.
9. The negotiation module requests the tunnel construction to DC_{MN} by transmitting the direction request message. The direction request message contains the FQDN of CN.
10. DC_{MN} makes inquiries about NS record for the FQDN of CN.
11. DC_{MN} may generate a shared key for encryption between DC_{MN} and DC_{CN} when the shared key is expired.
12. DC_{MN} requests the information for CN to DC_{CN} by transmitting the NTM information request message.
13. DC_{CN} replies the information about CN by replying the NTM information response message.
14. DC_{MN} requests RS to relay the communication between both nodes.
15. RS prepares the tunnel forwarding between both nodes, and replies the acknowledgement message to MN.
16. DC_{MN} requests the tunnel construction to CN to DC_{CN} by transmitting the route direction message.
17. DC_{CN} requests NS_{CN} to transmit a notification to CN.
18. NS_{CN} sends the notification to CN according to the device type of CN.
19. CN replies the acknowledgement message to DC_{CN}.
20. DC_{CN} forwards the route direction message to CN.
21. CN replies the acknowledgement message to DC_{CN}.
22. DC_{CN} also replies the acknowledgement message to DC_{MN}.
23. DC_{MN} indicates the tunnel construction to CN to MN by transmitting the route direction message.

24. The negotiation module transmits a keep alive message to keep the NAT table on the route.
25. MN transmits the tunnel request message to RS according to the direction from $\mathrm{DC_{MN}}$ because both nodes do not have a global IP address in this case.
26. RS transmits the tunnel request message to CN according to the direction from $\mathrm{DC_{MN}}$.
27. CN replies the tunnel response message to RS to complete the tunnel construction process.
28. RS replies the tunnel response message to MN to complete the tunnel construction process.
29. The negotiation module notifies the completion of the tunnel construction process to the application.
30. The negotiation module transmits a keep alive message to RS to keep the NAT table on the route.

4 CONCLUSION

The authors have been proposed a new IP mobility mechanism called NTMobile (Network Traversal with Mobility) to realize end-to-end communication in IoT systems. This paper designs an application based IP mobility scheme on Linux systems, where the developed IP mobility library can realize the IP mobility function in an application layer on Linux systems. As a result, developers can realize an end-to-end communication model by employing the enhanced IP mobility library.

ACKKNOWLEDGEMENT

This work is supported in part by the Grant-in-Aid for Scientific Research (26330103, 15H02697), Japan Society for the Promotion of Science (JSPS) and the Integration research for agriculture and interdisciplinary fields, Ministry of Agriculture, Forestry and Fisheries, Japan.

REFERENCES

[1] M. Buddhikot, G. Chandranmenon, S. Han, Y. W. Lee, S. Miller and L. Salgarelli, “Integration of 802.11 and third-generation wireless data networks,” Proceedings of the IEEE INFOCOM 2003, Vol. 1, pp. 503-512, 2003.

[2] Q. Zhang, C. Guo, Z. Guo and W. Zhu, “Efficient mobility management for vertical handoff between WWAN and WLAN,” IEEE Communications Magazine, Vol. 41, No. 11, pp. 102-108, 2003.

[3] L. A. Magagula and H. A. Chan, “IEEE802.21-Assisted Cross-Layer Design and PMIPv6 Mobility Management Framework for Next Generation Wireless Networks,” Proc. IEEE WIMOB ’08, pp. 159-164, Oct. 2008.

[4] D. Le, X. Fu and D. Hogrere, “A Review of Mobility Support Paradigms for the Internet,” IEEE Communications surveys, 1st quarter 2006, Volume 8, No. 1, 2006.

[5] A. C. Snoeren and H. Balakrishnan, “An End-to-End Approach to Host Mobility,” Proceedings of the ACM Mobicom, Page(s): 155-166, August 2000.

[6] C. Perkins, “IP Mobility Support for IPv4, Revised,” RFC 5944, IETF (2010).

[7] http://www.mip4.org, retrieved: February , 2012.

[8] S. Salsano, C. Mingardi, S. Niccolini, A. Polidoro and L. Veltri ”SIP-based Mobility Management in Next Generation Networks,” IEEE Wireless Communication, Vol. 15, Issue 2, April 2008.

[9] M. Bonola, S. Salsano and A. Polidoro, “UPMT: universal per-application mobility management using tunnels,” In Proc. of the 28th IEEE conference on Global telecommunications (GLOBECOM’09) 2009.

[10] H. Soliman “Mobile IPv6 Support for Dual Stack Hosts and Routers,” RFC 5555, IETF, 2009.

[11] H. Levkowetz and S. Vaarala, “Mobile IP Traversal of Network Address Translation (NAT) Devices,” RFC 3519, April 2003.

[12] Katsuhiro Naito, Kazuma Kamienoo, Takuya Nishio, Hidekazu Suzuki, Akira Watanabe, Kazuo Mori, Hideo Kobayashi, “Proposal of Seamless IP Mobility Schemes: Network Traversal with Mobility (NTMobile),” IEEE GLOBECOM 2012, December 2012.

[13] Katsuhiro Naito, Kazuma Kamienoo, Hidekazu Suzuki, Akira Watanabe, Kazuo Mori, and Hideo Kobayashi, “End-to-end IP mobility platform in application layer for iOS and Android OS,” IEEE Consumer Communications & Networking Conference (CCNC 2014), January 2014.

[14] http://www.ntmobile.net

A proposal of a countermeasure method against DNS amplification attacks using distributed filtering by traffic route changing

Yuki Katsurai*, Yoshitaka Nakamura** , and Osamu Takahashi**

*Graduate School of Systems Information Science, Future University Hakodate, Japan
**School of Systems Information Science, Future University Hakodate, Japan
{g2114005, y-nakamr, osamu}@fun.ac.jp

Abstract - In recent years, victims of DDoS attacks have been increasing rapidly all over the world, and it has become a very serious problem for network service providers. In particular, DNS amplification attacks have attracted attention. These attacks utilize DNS servers to cause huge damage to services using network systems. There are some methods that network administrators can introduce as countermeasures to DNS amplification attacks. Examples include a method to change the setting of DNS servers and a method to perform packet filtering on a firewall or routers. However, in these methods, it is not possible to suppress the damage to the network due to the large amount of packets passing through the system. Also, in the method of applying filtering, there is the problem that network congestion occurs on the processing terminals. In this paper we propose a countermeasure method against DNS amplification to reduce damage to the network. Our method is implemented on multiple routers on a network and performs distributed filtering using route-changing to prevent attack packets from reaching the target server. We also evaluate the utility of our method from the viewpoint of reducing the number of processes of each filtering terminal and the load on the network.

Keywords: DNS amplification attacks, DDoS, iptables, UDP, filtering

1 INTRODUCTION

In recent years, services that use the Internet have become familiar to people because of the development of the information society. However, damage from cyber-attacks, which are typified by DoS attacks (Denial of Service attacks) and DDoS attacks (Distributed Denial of Service attacks), has also increased. DoS attacks are a kind of cyber-attack which attacks the devices constituting the network, and thereby inhibit the provision of services. DDoS attacks are DoS attacks carried out using several dispersed sources. Among these DDoS attacks, the kind that has been typically exploited for many years is DNS amp attacks (DNS amplification attacks). A DNS amp attack refers to an amplification attack using a DNS server. The server reflexively responds to inquiries from a source, and acts as both a reflector and an amplifier. DNS amp attacks exploit these characteristics. In RFC 5358 / BCP 140, DNS amp attacks have been defined as Reflector Attacks [1], but in this research we unify such attacks under the name which by which they are commonly referred to. Figure 1 shows an overview of a DNS amp attack.

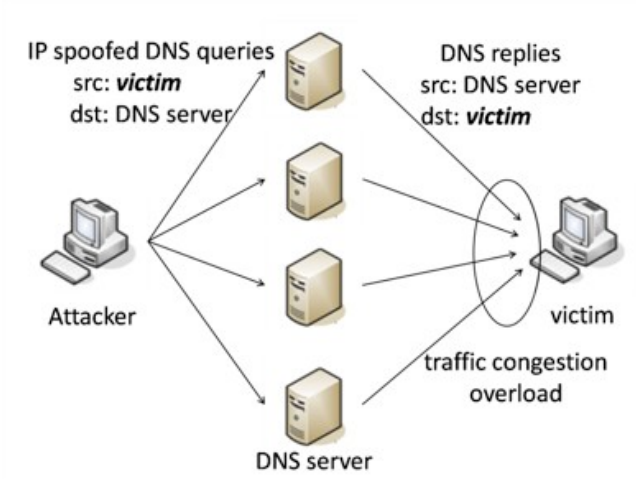


Figure 1: Overview of DNS amp attacks

In DNS amp attacks, name resolution requests of which the source IP address is spoofed are transmitted from an attacker to DNS servers. DNS servers that have received them return a response towards the terminal of which the IP address has been spoofed. Thus, DNS amp attacks lead to network congestion and overload the processing capacity of the victims. The hazards of this attack have been pointed out from 2001 [2].

There are two main types to DNS server: authoritative DNS server and cache DNS server. An authoritative DNS server shares a part of the domain name space it manages with multiple other DNS servers, and manages the distributed data by forming a tree structure. A cache DNS server is also called a resolver. It queries the authoritative DNS server when receiving a name resolution request, and it returns the results to a client. An authoritative DNS server that has a name resolution function enabled which is not inherently necessary, and a cache DNS server that processes a name resolution request sent from outside the network are called open resolvers. Incidentally, a home router can also function as a cache DNS server, so cases in which a home router is used to attack as an open resolver have occurred.

2 RELATED WORK

As a countermeasure to DNS amp attacks, there is a method to prevent DNS servers being used as an amplifier. Also there is a self-defense method that can be used by potential victims. In this chapter, we describe each measure.

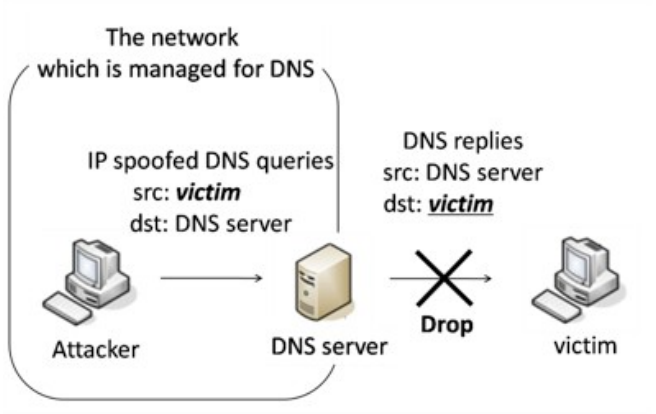


Figure 2: Measure using access control

2.1 Measures to be applied to the DNS servers

In the measure to ensure that DNS servers will not be used to attack others, the server acts as a protection. Regarding this type of countermeasure method, there are separate methods to ensure that cache DNS servers authoritative DNS servers are not used for DNS amp attacks. We describe each method.

2.1.1 Measures for cache DNS servers

There are two main approaches, which are performing access control and performing packet filtering using the router.

In access control, cache DNS servers only permit access to DNS queries from a client that the cache DNS servers regard as a target user, based on IP address. Figure 2 shows a schematic diagram of access control.

Figure 2 represents a situation in which a victim present on the outside of a network managed by the DNS server is set as a target of DNS amp attacks. In this case, the response a cache DNS server sends to a victim who is outside the scope of services is discarded, as dictated by a setting of the server. Therefore, the risk of the server being used as a stepping- stone in DNS amp attacks on the outside of the network is reduced. The details of this countermeasure are set out in the RFC 5358 / BCP 140 [1].

In packet filtering, a setting that prevents the transmission and reception of packets with spoofed source IP addresses is added to network devices such as routers. Spoofed packets do not reach a victim or a DNS server, thus attacks are eventually prevented. The packet filtering method is described under the name of Source Address Validation in RFC 2827 / BCP 38 [3], and also in RFC 3704 / BCP 84 [4]. In addition, this method has been used as a countermeasure against not only DNS amp attacks but various other kinds of cyber-attack.

2.1.2 Measures for authoritative DNS servers

In an authoritative DNS server, the source of a name resolution request is a cache DNS server. Also, authoritative

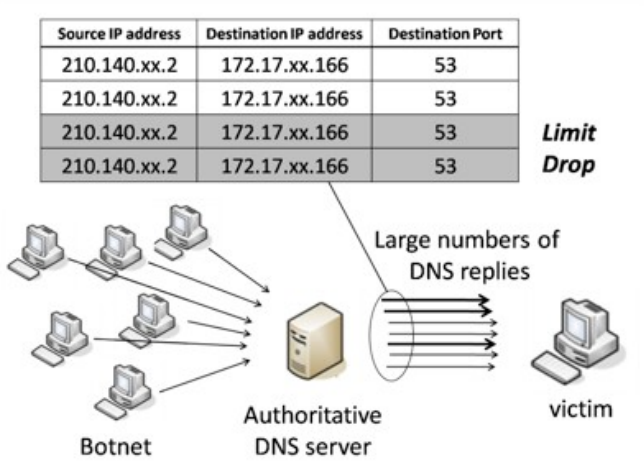


Figure 3: Response limit using DNS RRL

DNS servers are providing services to the entire Internet, so it would be disadvantageous to use the access control method in the same way as cache DNS servers. If authoritative DNS servers receive inquiries from a wide range, sent through technologies such as botnets, this cannot be addressed with access control. Moreover, in DNS amp attacks using the authoritative DNS servers, there is a tendency that the size of response packets of the DNS servers is larger than that of the packets in DNS amp attacks using the cache DNS servers. Thus, further measures have been required. In response to this fact, Paul Vixie et al proposed DNS RRL (DNS Response Rate Limiting) [5]. This method is a countermeasure that utilizes the fact that authoritative DNS servers return the same response at a high frequency to the same destination in a short time during DNS amp attacks. It monitors the response frequency, and if it exceeds a certain percentage, it limits and discards response packets. Also in this case, it is possible to respond to a variety of attacks by flexibly changing the conditions for determining the same responses. In Fig. 3, we show an example of the measures used DNS RRL.

A typical example of the problems of applying the DNS RRL is the occurrence of false detection. Since this determination is made based on statistics, whether it is an attack or not, if there are packets which should not originally have been detected it is determined that these are attack packets. To prevent this false detection, a retransmission request is sent to a cache DNS server using TCP. This action achieves a correct name resolution. In relation to this, Rozekrans et al have shown the results of field trials of DNS RRL [6]. In this reference, a method of giving an evaluation value for each client is applied, and this is called DNS dampening. The author wrote that it is necessary to verify the usefulness of attacks that currently exist, and to perform source verification in order to respond to development attacks in the future.

2.2 Measures that can be applied by victims for self- defense

An ideal countermeasure against DNS amp attacks is the simultaneous application of source verification to all of the

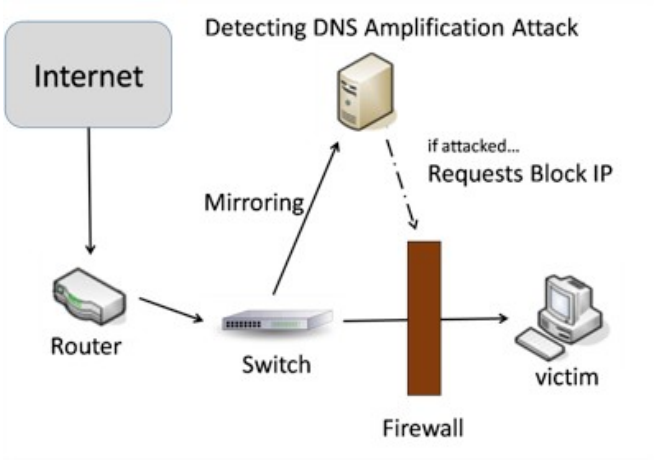


Figure 4: Filtering method using DDAA

network devices around the world, but it is not practical. The next best option is to apply access control and DNS RRL to DNS servers. However, DNS RRL is still under study. Also, these measures are intended to be applied to e each device by administrators of the network and DNS servers. In addition, devices and DNS servers to which measures have not been applied would still be exploited in attacks as stepping-stones. Therefore, a victim requires measures against DNS amp attacks as a means of self-defense. The method used at these times is filtering performed on the victim's side of the network. A method for filtering by attack detection and firewall on the victim's side of the network has been proposed by Ye et al [7]. In this approach, a switch is interposed between the Internet, which is the 'backbone', and the network on the victim's side, and it copies packets. The copied packets are sent to an installed system named DDAA (Detecting DNS Amplification Attacks). This system records the information of the packets, and then blocks packets that are considered attack packets with a firewall. Figure 4 shows this method including DDAA.

This method stores the IP address and destination port of the packets that pass through the switch in the DDAA's internal database. Therefore, the performance of the system is reduced as time elapses. For this reason, if the packet information stored in the database exceeds 10000, it is set to delete all the information which has been in the database more than three seconds. The advantage of this approach is that the information of the filtering can be dynamically updated and saved by the parameter settings of DDAA. In addition, by managing the detection and blocking at the same terminal, using the firewall it can directly drop DNS reply packets that are sent to a victim intermittently. The problem of this approach is that it does not consider the burden on devices and network congestion. Depending on the nature of the firewall, it may not be able to respond to the congestion of the network. Also, saving packets and constantly performing the matching process with the database results in a high load on the firewall and DDAA, which can affect performance. Paola et al proposed a method that maintains low loads on devices [8]. In this approach, packets passing through devices are retrieved efficiently from the database by using a Bloom Filter. Accordingly, the burden on devices is reduced, and it is possible to perform accurate packet filtering. However, regarding this approach, the influence on the network is not taken into consideration, and damage due to congestion in performing filtering is also overlooked.

3 PROPOSED METHOD

3.1 Research task

In this research, we assume a case in which DNS servers, to which countermeasures of the entire network are not applied, to have been used in DNS amp attacks. Our research deals with packet filtering as a means of self-defense means on the victim's side of the network. After attack detection, to ensure that attack packets do not reach the victim's service, we perform filtering along the network path. Further, by performing distributed filtering using multiple routers, the burden of the routers that perform filtering and the networks on either side of them is reduced.

3.2 System configuration

The system of the proposed method consists of the following contents; the Internet, as a backbone, a switch that exists close to the victim in the Internet, multiple routers which perform filtering, and a router which integrates packets that have changed route. In Fig. 5, a schematic view of the system is shown.

We place a switch at the connection point between each network and the Internet. It distributes the packets to an arbitrary number of routers to perform filtering. They operate as filtering routers. After completing the filtering, they send packets to a router that is used for integration of the packets, and it sends packets in a fixed order to the victim server.

3.3 Performance details

In the proposed method, there are four stages. They are: the time until a DNS amp attack is detected, distribution of packets after attack detection, distributed filtering, and

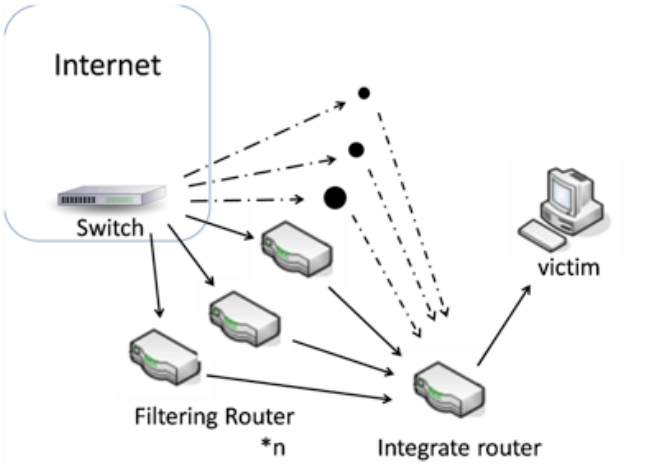


Figure 5: Schematic diagram of the proposed method

integration of packets after filtering. We will describe each of the steps.

3.3.1 Until a DNS amp attack is detected

In the proposed method, all routers and switches perform in the same way as existing devices until a DNS amp attack is detected. In the case of there being several routers, each one functions normally as a router within the network.

3.3.2 Distribution of packets

When a DNS amp attack is detected, a switch sends attack packets which are distributed to filtering routers. Immediately prior to distributing the packets, this switch sends commands ordering the commencement of filtering and stating which router will perform. As an example, router 1, router 2 and router 3 are set as filtering routers. In this case, it is assumed that router 1 is a router which is used as a general path. The switch adds fragment information to the packet head of packets being subjected to encapsulation. This fragment information has a similar meaning to 'flag field' and 'fragment offset field' used in the IP header, so it shows the information of what number this mass of packets is, out of all those that passed through after the detection of a DNS amp attack. Then it changes the route from router 1 to router 2. Similarly to the case of router 1, a switch encapsulates a defined amount of the packets, adds to them the fragment information and then transmits them to router 2. The same is true when the switch sends packets to router 3. Then the object returns to router 1, whereafter the same operation is repeated. Further, if the fragment offset has reached the upper limit, it is set to repeat from 0 again. Figure 6 shows the status of packet distribution.

If the packet distribution is stopped, the switch finally sends the packets to a router that has been used as a path before a DNS amp attack was detected. In this example, after the switch has finally sent the mass of packets to router 1, it also sends a number of masses of packets t to router 1, and then stops the division of the packets to resume communication as usual. This is in order to prevent the communication eventually being affected due to abnormalities in the order of the packets.

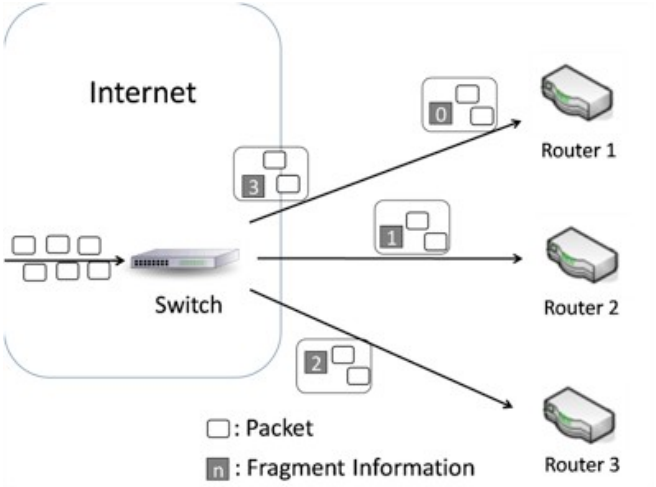


Figure 6: Distribution of packets by the switch

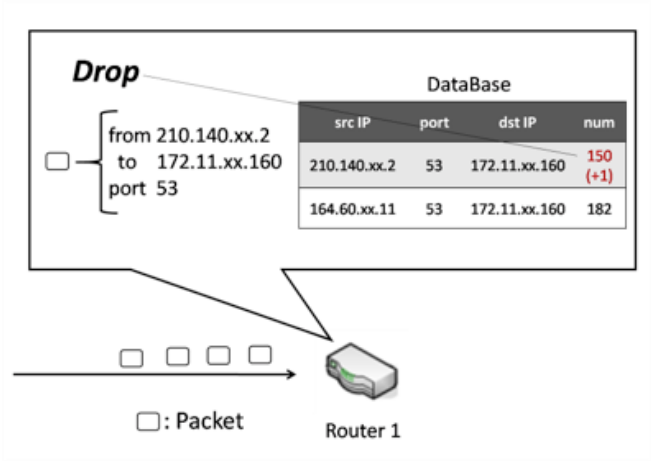


Figure 7: Overview of filtering by database collation

3.3.3 Distributed filtering

Each filtering router starts performing filtering after the switch notified it to do so and sends a mass of packets. First, packets are transmitted to the UDP53 port. If a packet filtered is addressed to the UDP53 port, filtering routers register the source IP address of the packet to their own database. Then the filtering routers repeat the same operation. They recognize a DNS server which has transmitted a certain amount or more packets within a specified time as a server being used as a stepping stone in DNS amp attacks. Subsequently, they share the information of the source IP address with the other filtering routers. Following these operations, they discard all DNS reply packets coming from a DNS server that is considered to be an attack source.

Figure 7 shows the state of filtering. In addition, when performing filtering, excepting information that is shared with other filtering routers, they reset the database at regular intervals. This is a measure to maintain a certain level of search efficiency regarding the information of the packets which are sequentially registered. Filtering routers send packets to an integration router after finishing the filtering for each mass of packets.

3.3.4 Integration of packets

The integration router sorts the mass of packets into the correct order by referring to the fragment offset information that was added by the switch. Then it transmits packets to the server of a victim in ascending order of number.

3.3.5 Relation of the processing

Figure 8 shows the relationship to other devices of each device in the filtering process of the proposed method.

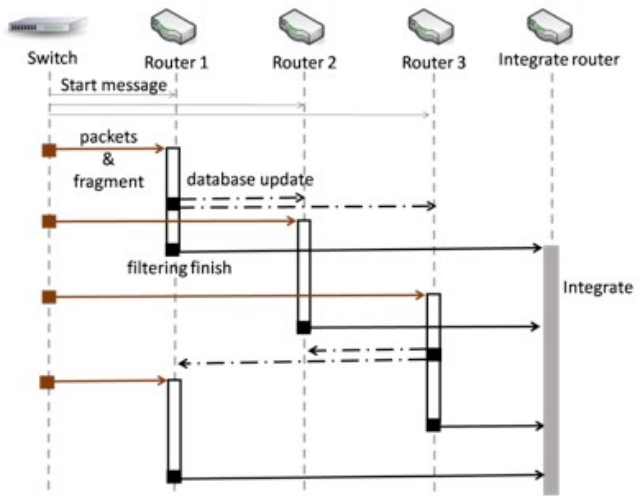


Figure 8: Relationship diagram of the filtering process

4 EXPERIMENTATION

About the proposed method in this research, we reproduced a DNS amp attack in a virtual environment and conducted an evaluative experiment. Hereinafter, we describe the experimental environment, and the details of the experiment contents the evaluation.

4.1 Experimental environment

As an experimental environment, we established virtual machines to act as an attacker and a victim, DNS servers, a switch, filtering routers and an integration router, and we applied the appropriate settings to each one. Thereafter, we created a scenario of a DNS amp attack. Table 1 shows the experimental environment.

4.2 Experiment contents

We gave the role of the devices that are used in the proposed method to virtual machines, and allowed a large number of packets to be sent to a victim using a virtual machine that was configured as a DNS server. After a certain time had elapsed from the start of the DNS amp attack , filtering operations began along the router path. For distributed filtering, filtering routers use the database and iptables to manage packet dropping and communication permission. In this experiment, we performed distributed filtering using two routers.

Table 1: The parameters used in the experiment

Using distribution	Ubuntu 12.04 32bit
Programming language	Python 2.7.3, PHP 5.3.3
Database	My SQL 5.1.73
Number of attack packets / s	1000

4.3 Evaluation contents

We will now evaluate the results of the above experiments.

i. Throughput between the switch and the victim
ii. The number of Processing packets and the percentage of blocking

For these comparisons, we compared single filtering and distributed filtering by calculating each value. The results are shown in the following section.

5 RESULTS AND DISCUSSION

Figure 9 shows throughputs from a switch to a victim. Figure 10 shows the number of packets that filtering routers processed and the percentage of these packets that were blocked.

These results indicate the utility of this research. Comparison of the throughput reveals that the adverse effect on a victim's side of the network is smaller when filtering using the proposed method than when filtering using a single router. It's a measure of the throughput of the normal communication packet with the exception of the attack packets. Furthermore, it is possible that the throughput can be raised by increasing the number of

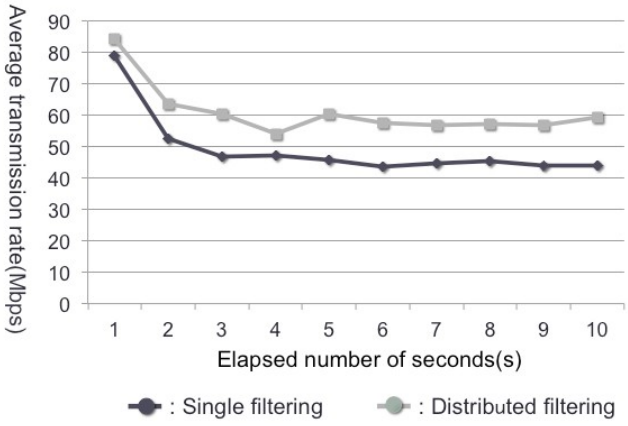


Figure 9: Graph of comparison of throughput from switch to victim

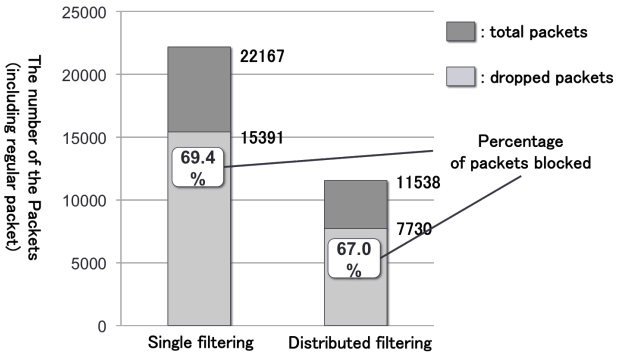


Figure 10: Graph of the number of packets filtering routers processed and the percentage of packets blocked

routers used for distributed filtering. Also, regarding the number of processing packets, the numbers of discarded packets and probability that the attack packet is blocked, the experimental results indicate that using the proposed method decreases the load per single router. By the difference between the time required for the distribution of the packet, the amount of normal communication packets transmitted is increased and block rate has somewhat changed. However, it seems that there is no significant impact on the accuracy of the block. From these results, the usefulness of this research has been proved.

6 CONCLUSION

In this paper, we have described a method of distributed filtering as a counter measure against DNS amp attacks. The purpose of this research is to reduce network congestion and the load on a router. We conducted an experiment to evaluate it, and it indicated the improvement of throughput in the network on the victim's side and decrease in the amount of processing packets per single router. From these results, the purpose of reducing the burden of the network and devices while maintaining the performance can be said to have been achieved. As future challenges, there are a survey of numerical change when the number of routers is increased, and an investigation into the performance of the switch when a large number of attack packets are sent to a victim.

REFERENCES

[1] J. Damas, and F . Neves, Preventing Use of Recursive Nameservers in Reflector Attacks, RFC 5358, BCP 140, https://www.ietf.org/rfc/rfc5358.txt (2008).

[2] V. Paxson, An analysis of using reflectors for distributed denial-of-service attacks, ACM SIGCOMM Computer Communication Review, Vol.31, No.3, pp.38-47 (2001).

[3] P. Ferguson, and D . Senie, Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing, RFC 2827, BCP 38, https://www.ietf.org/rfc/rfc2827.txt (2000).

[4] F. Baker, and P. Savola, Ingress Filtering for Multihomed Networks, RFC 3704, BCP84, https://tools.ietf.org/rfc/rfc3704.txt (2004).

[5] P. Vixie, and V. Schryver, DNS Response Rate Limiting (DNS RRL), ISC-TN-2012-1-Draft1 (2012).

[6] T. Rozekrans, and J. Koning, Defending against DNS reflection amplification attacks, University of Amsterdam System & Network Engineering RP1 (2013).

[7] X. Ye, and Y. Ye, A Practical Mechanism to Counteract DNS Amplification DDoS Attacks, Journal of Computational Information Systems, Vol.9, No.1, pp.265-272 (2013).

[8] S. Paola, and D. Lombardo, Protecting against DNS Reflection Attacks with Bloom Filters, Proceedings of the 8th international conference on Detection of intrusions and malware, and vulnerability assessment (DIMVA'11), pp.1-16 (2011).

Session 2:
Wireless Networks
(Chair : Tomoki Yoshihisa)

Indoor/outdoor determination method using various sensors for the power saving of terminals in Geo-fencing

Yoshitaka Nakamura†, Mifumi Ono†, Masashi Sekiya†, Kazuaki Honda‡, and Osamu Takahashi†

†School of Systems Information Science, Future University Hakodate, Japan
‡IDY Corporation, Japan

Abstract - Recently, various services using the location information are developed with the wide spread of smartphones. Geo-fencing is one of the services. In Geo-fencing services, terminals automatically carry out the some processing associated with the virtual border when they detect that themselves passes the border. However, it is a problem that power saving of Geo-fencing service covering over indoors and outdoors with high accuracy. In this paper, we propose indoor/outdoor estimation method using various types of sensors with small power consumption. The results of the evaluation experiment show that the proposed method can estimate the movement in indoor and outdoor with 82.14% accuracy, and this method is an effective means for indoor and outdoor Geo-fencing service.

Keywords: Indoor/outdoor estimation, Magnetism sensor, Illumination sensor, Temperature sensor, Geo-fencing, M2M

1 INTRODUCTION

By the wide spread of smart devices with GPS (Global Positioning System), the acquisition of the positional information become easier. Therefore the demand for location-based services which provide appropriate service depending on the position of the user increases. Geo-fencing technology[1] is one of the M2M (Machine-to-Machine) type services and attracts attention recently.

Geo-fencing technology sets the virtual border on a map. This technology detects the positional relations between this border and target terminal, and let the terminal perform predetermined processes automatically according to the relations. Location-based service can be easily provided for terminals by using Geo-fencing technology. There are some type of services using Geo-fencing technology such as “Gotouchi information”[2], “Arrived”[3], O2O (Online to Offline) services that let the information of the neighboring store link the positional information of users, and position monitoring services. The monitoring system in prisons and nursing homes are example of position monitoring services using Geo-fencing. It is necessary to manage the target person surely in such facilities, but it is undesirable to restrict the target indoors from a humanitarian point of view. The use of Geo-fencing is considered as a solution for these problems. The monitoring system sets the movement allowable range of the target person outside of facilities as a virtual border on the map.

To realize the service using Geo-fencing technology, it is necessary to acquire positional information with high accuracy, because it is necessary to determine inside/outside of Geo-fence accurately. Most of services using Geo-fencing technology outdoors acquire the positional information of the user terminal from GPS and determine inside and outside of Geo-fence using this information. In addition, there is another the positional information acquisition method using HexRinger[4] using the radio field strength (Received Signal Strength Indicator, RSSI) from base stations. As for the indoor position estimation method, there are techniques to estimate using wireless communication technologies[5], [6], a technique using positional information and RSSI of the wireless LAN base station[7], and techniques using the Fingerprint method[8], [9] . About the use of Geo-fencing service only outdoors or only indoors, acquisition technology of positional information mentioned above can be used depending on use position. However, if a Geo-fence passing over indoor and outdoor exists, it is necessary to use acquisition technologies of positional information indoor and outdoor together.

At the time of using Geo-fencing technology, the point where it is necessary for user terminals to keep acquiring own positional information becomes the big problem. Generally, Geo-fencing technology which turns on GPS function always has large power consumption of the terminal. The state that always activated GPS function consumes power of approximately 120 times in comparison with a standby state[10]. Because the battery capacity of the user terminal is not so large at this time, the problem of this power consumption shortens the available time of Geo-fencing service. Reference[11] tries to reduce power consumption of terminals at the time of the acquisition of the positional information by combining movement detection function, switching function of positioning means, and positioning function with variable interval together. However, an effect of the power reduction becomes small by Geo-fencing service including indoor situation because this method assumes only the outdoor positioning. On the other hand, Ref.[12] reduces power consumption by detecting the situation that GPS positioning is impossible including the indoor from temperature information, and turning off GPS function. For reduction of useless power, it is necessary to switch GPS function and indoor position estimation method when indoor Geo-fencing is assumed. However, there is the situation that effective power reduction is difficult, because the temperature information depends on seasons and weathers and indoor/outdoor estimation with high accuracy is difficult. In this paper, we examine method to improve the power consumption of Geo-fencing service by realizing indoor/outdoor estimation with high accuracy using the intensity of illumination and the quantity of magnetism in addition to temperature information.

2 RELATED WORK

2.1 Power saving method using restraint of GPS positioning

Many existing restraint methods of the large power consumption of GPS positioning use three following functions such as movement detection function, switching function of positioning means, and positioning function with variable interval. The movement detection function is used to control sensing and positioning frequency when a user terminal stands still. When a terminal can determine that the user of the terminal is in a stationary state from positional information acquired from GPS, the terminal stops GPS, starts an acceleration sensor and watch the state of the terminal. When a terminal starts to move and acceleration data are measured, the terminal stops the acceleration sensor and start GPS. Because the user repeats a stationary state and a movement state, this method can control the positioning in a stationary state, and enable effective power saving. However, the movement detection function by the acceleration cannot control power consumption in the state that a terminal continues moving. Therefore Ref. [13] expands the movement detection function by adding a function to change the movement detection timing with the acceleration sensor according to the distance to the virtual border. By this function, GPS is kept in a sleep state for a long time.

The switching function of positioning means is used to switch to the positioning means with smaller power consumption depending on environment. Reference [1], [14] consider a positioning error standard of each positioning means and the distance from the terminal to the virtual border using GPS positioning and radio base station positioning. When the distance to the virtual border is longer enough than the positioning error standard, power saving realizes by stopping GPS positioning with large power consumption, and switching the positioning means to the base station positioning with smaller power consumption. By the space variableness positioning function calculating a virtual border and the distance of the terminal to reduce the positioning number of times, and calculating the distance and the arrival time from expected approach speed to the imagination border, and adjusting a positioning interval.

The positioning function with variable interval reduces the positioning number of times, by calculating the distance between the virtual border and the terminal, and calculating the arrival time from the distance and the expected approach speed to the virtual border., and adjusting the positioning interval[15], [16]. When this function is used, a prediction of the approach speed becomes the problem.

Reference [11] proposes the reduction method of power consumption of terminals in Geo-fencing services by combining these movement detection function, switching function of positioning means, and positioning function with variable interval.

2.2 Power saving method using movement timing of indoors/outdoors

GPS greatly consumes electric power at the time of the detection of the GPS satellite. Therefore, it will use too much useless electricity to keep GPS on in the place where GPS positioning is difficult such as indoor place. Reference[17] aims at the reduction of the consumption electricity by stopping GPS positioning on the place where GPS positioning is difficult, and restarting GPS positioning on movement to the place where GPS positioning is possible (i.e., outdoors). When a terminal moves from indoors to outdoors or from outdoors to indoors, the temperature around the terminal is more likely to greatly change. Reference [17] proposes the estimation method of movement timing between outdoor and indoor using temperature information. This method periodically records the ambient temperature of the terminal using a temperature sensor. And this method estimates the timing that it is moving from indoors to outdoors or from outdoors to indoors when the temperature greatly changed.

Reference [18] investigates the average indoor temperature of each season in Tokyo of 2010. Table 1 shows the comparison result between the average indoor temperature of Ref.[18] and the average outdoor temperature observed by Japan Meteorological Agency in 2010. According to this result, there is

Table 1: The indoor and outdoor temperature difference in Tokyo of 2010

Seasons	Avg. Indoor temp.(°C)	Avg. Outdoor temp.(°C)	Temp. diff.(°C)
Spring	20.2	13.5	6.7
Summer	28.0	27.1	0.9
Fall	24.7	19.2	5.5
Winter	17.6	7.8	9.8

difference of temperature above a certain level between outdoor and indoor in the season except the summer, and the movement estimation to indoor or outdoor by the temperature seems to be possible. However, the accurate movement estimation to indoor or outdoor is difficult only by temperature, because there is little difference of temperature between indoor and outdoor in the summer and it is thought that the difference becomes smaller by the structure of the building. In addition, when the terminal is assumed to be carried by a user, the influence that the heating element except the atmosphere such as the user's body or terminal itself gives to the temperature sensor grows large. Therefore, it is necessary to improve the precision of the estimation of movement timing between outdoor and indoor by using the information except the temperature for the effective reduction of power consumption.

3 PROPOSED METHOD

The existing methods cannot estimate indoor/outdoor with high accuracy. We aim at the improvement of the precision of the indoor/outdoor estimation by using physical quantities except the temperature to the determination.

3.1 Approach

At first, we examine the indoor/outdoor estimation method using the general sensor device except the temperature sensor to achieve our research purpose.

Intensity of illumination The intensity of illumination has the large difference of values between indoor and outdoor in the daytime and also at night. In the daytime, the outdoor illumination is about 10000 - 80000 lux. In contrast, the indoor illumination is about 300 - 800 lux, and there is a difference of about 100 times. In the nighttime, the outdoor streetlight is about 1 - 5 lux, and the brightness of the full moon is about 0.01 - 0.1 lux. The illumination of the general indoor fluorescent lamp is about 300 - 800 lux. From this, it is thought that the illumination is effective for the indoor/outdoor estimation both in the daytime and the nighttime.

Quantity of magnetism The quantity of magnetism is easy to use for the estimation, because there is a large change in the quantity of magnetism indoors. Though the previous setting including the measurement in the building is necessary to use magnetism for indoor positioning estimation, only the observation of the change in the quantity of magnetism is necessary for the indoor/outdoor estimation.

GPS signal We can guess that there is more likely to be the terminal indoors, if the terminal is in a condition not to be able to acquire correct positional information using GPS. In this case, the noise of the GPS is calculated from the signal to noise ratio (S/N ratio, SNR). We can find the quantity of noise from the SNR of the GPS satellite which communicated. The indoor/outdoor estimation is possible using this quantity of noise.

The standard power consumption of sensors measuring these physical quantities is Table 2. According to the table, the

Table 2: Power consumption of each sensor

	Consumption current
GPS module	50 mA
Magnetism sensor	100 μA
Illumination sensor	80 μA
Temperature sensor	6 μA

power consumption of each sensor of physical quantity is much smaller than the GPS module.

3.2 Indoor/outdoor estimation method

We propose the indoor/outdoor estimation method using intensity of illumination, quantity of magnetism, GPS Signal, and temperature conventionally used.

3.2.1 State transition

This method has 3 states, estimating state, indoor state, and outdoor state. Figure 1 is the illustration of the state transition diagram of the proposed method. In the outdoor state, Geo-

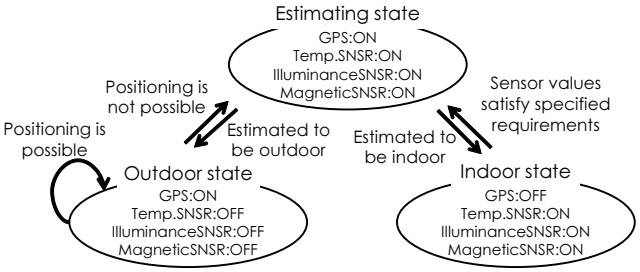


Figure 1: State transition diagram

fencing service uses GPS like existing method. If SNR of all GPS signals are more than 20dB and GPS module of the terminal can communicate with more than 3 satellites, it is judged that positioning is possible and the state remains in the outdoors state. Otherwise, it is judged that positioning is impossible and the state transits to the estimating state.

In the indoor state, GPS positioning is stopped because GPS is not available. Afterwards, if there is some available indoor positioning techniques, GPS positioning is switched to the technique. If some change is seen in one value among the acquired values by sensors (temperature, illumination, and magnetism) in the indoor state, this method causes a transition from the indoor state to the estimating state. It is a condition to estimate to be movement out of the indoor as follows. If each sensor satisfies the following conditions, it is considered that the terminal moved from indoor to outdoor.

1. Intensity of illumination
   The illumination sensor of the terminal measures the intensity of illumination every 1 second and calculate the standard deviation of the illumination every 10 seconds in the daytime. If the illumination sensor sense a change in the variance values more than 35, the state transits to the estimating state. In the night, if there is a change more than 50 lux of the acquired illumination value, the state transits from indoor state to the estimating state.

2. Temperature
   The temperature sensor of the terminal measures temperature every one second. If the temperature sensor sense a change in 1 $^{\circ}$C within 30 seconds

3. Quantity of magnetism
   The magnetism sensor of the terminal measures the quantity of magnetism every 1 second and calculate the variance of the magnetism every 10 seconds. If the magnetism sensor sense a change in the variance values more than 18, the state transits to the estimating state.

The condition to transit from outdoor state or indoor state to an estimated state has described. The transition from an estimated state to outdoor state or indoor state is carried out using the following estimation method.

3.2.2 Estimation of indoor or outdoor

It is necessary to estimate whether the terminal is outdoor or indoor in the estimating state. The indoor/outdoor estimations

by 3 physical quantities mentioned above are carried out in parallel. In addition, the estimation reliability is calculated in each estimation. If each estimated result is different, the estimated result with highest reliability is adopted.

Estimation by illumination Figure 2 shows the flowchart of the indoor/outdoor estimation using intensity of illumination. Outdoor illumination becomes higher than indoor in the

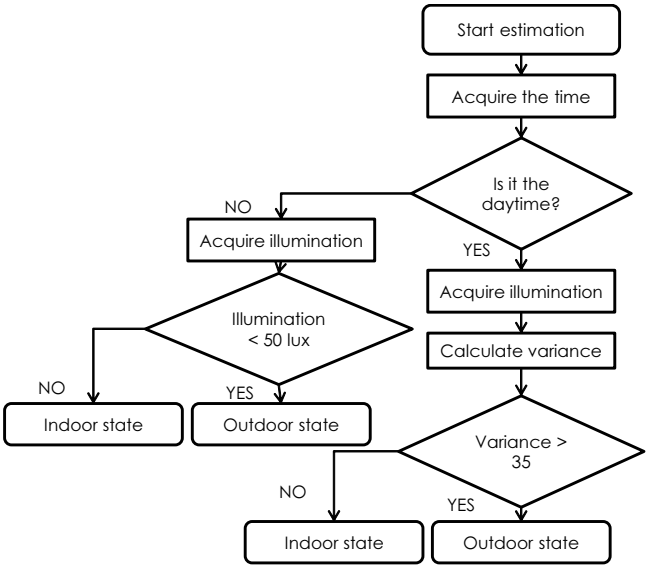


Figure 2: Flowchart of estimation by illumination

day time and indoor illumination becomes higher than outdoor in the night. In the night outdoors, illumination is almost 0 lux. Therefore, it is necessary to switch the estimation techniques according to the time (night or daytime). The local sunrise time and sunset time are calculated from latitude, longitude, the date, and time. Based on sunrise time and sunset time, it is decided whether it is the daytime or night. The outdoor illumination greatly changes from 10000 lux to 80000 lux (measurement limit) under the influence of the shadow of the buildings, the direction of the terminal, and clouds. In most cases, there is a large difference between the outdoor illumination and the indoor illumination. However, the estimation using the absolute values of the illumination is difficult because the illumination value of about 10000 lux may be acquired depending on time, place, and conditions even if it is indoor. On the other hand, the indoor illumination only changes from 0 lux to 10000 lux, so the variances of the illumination have a large difference between outdoor and indoor. The standard deviation of the illumination is calculated every 10 seconds in the daytime. Therefore if the standard deviation of the illumination is more than or equal to 35, it is estimated that it is outdoor and if the variance is less than 35, it is estimated that it is indoor.

We assume 50 lux as the threshold and estimate it by night. It is estimated that it is indoor if the illumination value is more than the threshold, and estimated that it is outdoor if the illumination value is less than the threshold.

The estimation by illumination has high reliability around noon, and low reliability around sunrise and sunset. Therefore we assume the number that the difference between the measurement time and the sunset (or sunrise) time divided by 24 as the illumination reliability. For example, if the measurement date and time are January 24th, 20:00, the illumination reliability becomes $(20 - 16)/24 = 0.1666\cdots$.

Estimation by magnetism Figure 3 shows the flowchart of the indoor/outdoor estimation using a quantity of magnetism. The variance of magnetism greatly changes indoors. In this

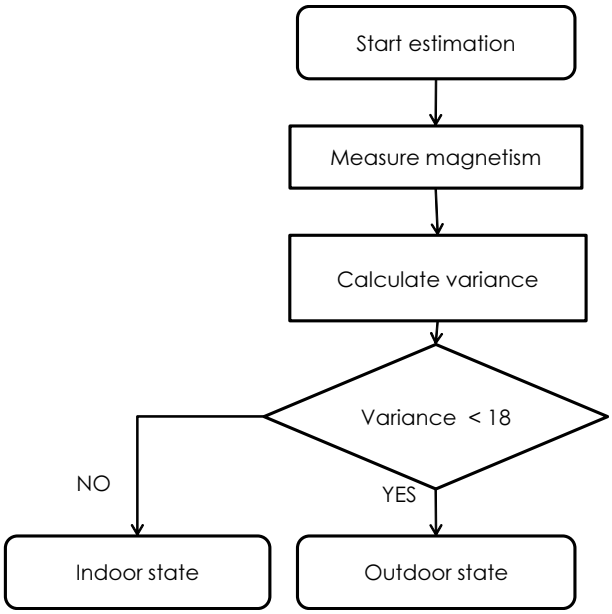


Figure 3: Flowchart of estimation by magnetism

method, the variance of magnetism is calculated every 10 seconds. If the variance of magnetism is less than the threshold, it is estimated to be outdoor. And if more than or equal to the threshold, it is estimated to be indoor. The threshold of the variance is set to 18 based on Ref. [19].

The estimation by magnetism cannot provide accurate result if the terminal is covered in a bag or a pocket. Therefore the covered situation of the terminal is estimated by a proximity sensor, and the value of the proximity sensor at that time is set as the magnetic reliability. Because the proximity sensor acquires 0 if the terminal is covered, the magnetic reliability becomes 0. If the proximity sensor acquires the value except 0 , the magnetic reliability becomes 1.

Estimation by temperature Figure 4 shows the flowchart of the indoor/outdoor estimation using temperature. The relation between outdoor temperature and indoor temperature changes by seasons. Therefore we divided one year into summer from April to September when the indoor is cooler than outdoor, and winter from October to March when outdoor is colder than indoor. Afterwards, this method detects whether the inclination of the temperature change is rise or drop.

The reliability of the estimation by the temperature is high if the season approach to midsummer and the depth of winter. Therefore we assume the number that the difference between the measurement month and the April (or September) divided

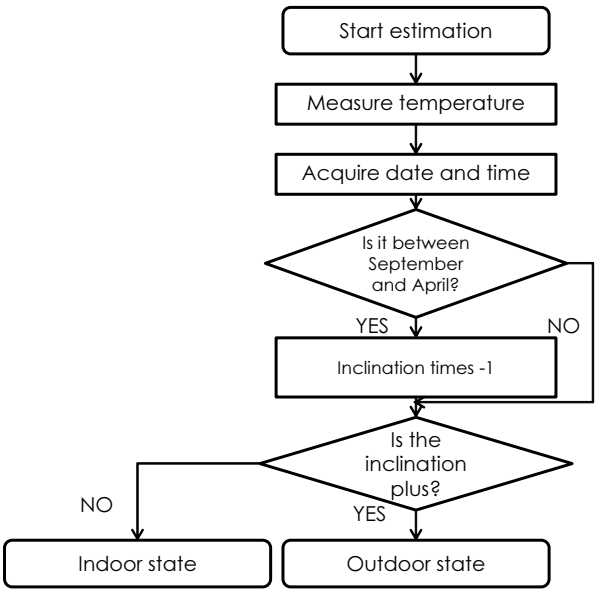


Figure 4: Flowchart of estimation by temperature

by 12 as the temperature reliability. For example, if the measurement date is January 24th, the temperature reliability becomes $(4-1)/12 = 0.25$.

Overall estimation This proposed uses the estimation by illumination, the estimation by magnetism, and the estimation by temperature concurrently and acquires each estimation result and reliability. If 3 estimation results are different, the proposed method considers the result with the highest reliability to be as the estimation result. For example, if this method estimate in the state that the terminal is not covered at 20:00 of January 24th, the illumination reliability is $0.166\cdots$, the magnetic reliability is 0, and the temperature reliability is 0.25. Therefore, the estimation by the temperature is used as the estimated result.

4 EXPERIMENTS

4.1 Overview of the evaluation experiments

We implemented the proposed method to a smartphone (Sony Xperia Z3) as Android application, and evaluated the method by some experiments. For the evaluation of the estimation accuracy, it is necessary to record the correct location of the terminal (correct answer data). At every movement between outdoor and indoor, I assume it correct answer data by pushing the record button, and recording it. Therefore we made the process to record indoor or outdoor, and at every movement between indoor and outdoor, the user pushes the record button and record whether it is indoor or outdoor. This record becomes correct answer data. We compare the estimated result with this correct answer data and calculate the precision ratio.

Figure 5 and 6 are the evaluation results of example scenario. The subject with the smartphone terminal moves indoor $\rightarrow$ outdoors $\rightarrow$ indoor in 10 minutes. Figure 5 shows the estimation result of the proposed method. Similarly, Fig. 6 is the estimated result using only temperature. In this example, the precision ratio of the proposed method is 92%, and the precision ratio of the estimation by temperature is 63%.

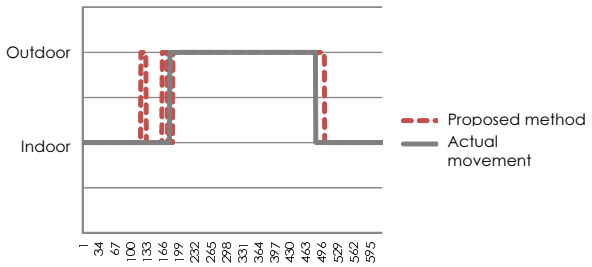


Figure 5: Estimation result using the proposed method

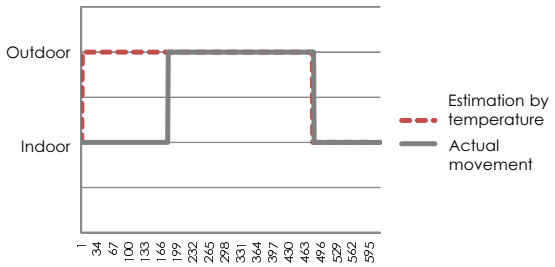


Figure 6: Estimation result using only the estimation by temperature

We evaluate such a precision ratio by comparing the proposed method with 3 kinds of independent indoor/outdoor estimations.

The evaluation experiment is based on 7 scenarios of Table 3. In Ex.1-Ex.4 and Ex.6 scenarios, the subject walks holding the terminal in his hand. And in Ex.5 and Ex.7 scenarios, the subject walks carrying the terminal in a pocket of his jacket.

4.2 Results of the experiments

Table 4 shows the results of experiments. The precision ratio of the proposed method was 82.14% that were higher than other independent estimations. Because the proposed method can use the other estimate technique in the situation which one estimated technique is weak in, the method can keep estimation with high precision. For example, it is the situation such as at the time of the outdoor movement with the large temperature difference in the estimation by the temperature or the situation in the evening in the estimation by the illumination. In Ex.5 and Ex.7, because the terminal is in the subject's pocket and the estimation by magnetism is difficult, the precision ratio of the estimation drops down.

The estimation only by the illumination was able to attain just under 70% of the precision ratio on the average. However, it was not able to be estimated with accuracy at the time without a clear illumination difference between indoor and outdoor such as evening in Ex.4. This is because a false estimate of the indoor occurs by very small difference of the illumination when the subject stopped in the shade for a long

Table 3: Scenario of the evaluation experiments

	Date	Time	Meas.time	Place	Route	Weather
Ex.1	1/28	11:55-12:00	5min.	Within the campus	Indoor → Outdoor → Indoor	Cloudy/Sunny
Ex.2	1/28	12:44-13:16	32min.	Around the campus	Outdoor	Cloudy
Ex.3	1/28	13:26-13:32	7min.	Commercial facility	Outdoor → Indoor	Cloudy
Ex.4	1/28	15:23-16:09	46min.	Commercial facility, around and within the campus	Outdoor → Indoor → Outdoor → Indoor	Snow
Ex.5	1/28	21:44-21:50	6min.	Commercial facility	Outdoor → Indoor	Cloudy
Ex.6	1/28	22:21-22:31	10min.	Home	Outdoor → Indoor	Cloudy
Ex.7	1/28	22:35-22:40	5min.	Home	Indoor	Cloudy

Table 4: Result of the evaluation experiments

	Proposed method	Illumination only	Temperature only	Magnetism only
Ex.1	92%	97%	63%	50%
Ex.2	99%	97%	98%	98%
Ex.3	93%	95%	40%	15%
Ex.4	96%	18%	93%	81%
Ex.5	51%	63%	36%	61%
Ex.6	97%	96%	90%	99%
Ex.7	47%	21%	0%	31%
Avg.	82.14%	69.57%	60%	62.14%

time occurred. By experiment 5 and experiment 7 that a terminal was covered, the estimation by the illumination was not able to achieve the high precision ratio because the illumination became the constant value with approximately 0.

The estimation only by the temperature was able to attain 60% of the precision ratio on the average. The estimation was not able to achieve high precision ratio in the situation that moved indoor from the cold outdoor like Ex.3 and Ex.5. This is caused by the characteristic of the temperature sensor. The temperature sensor is sensitive to cold, and the reaction of the sensor oneself worsens when the sensor observes very low temperature. And it is one reason that there is a time lag for the temperature acquisitions.

The estimation only by the magnetism was able to attain around 60% of the precision ratio on the average. Because it was estimated to be the outdoor by mistake indoor with the magnetic variation, the precision ratio did not rise.

5 CONCLUSION

In this paper, we proposed the indoor/outdoor estimation method to reduce power consumption of Geo-fencing service while maintaining the detection precision of the terminal position in the situation that GPS positioning is impossible such as indoor. In future work, we need a comparative evaluation of the precision and power consumption with conventional Geo-fencing.

REFERENCES

[1] U. Bareth, “Privacy-aware and Energy-efficient Geofencing through Reverse Cellular Positioning,” Proc. of IWCMC2012, pp.153–158 (2012).

[2] NTT TownPage Corporation, “Gotouchi information,” https://play.google.com/store/apps/details?id=jp.co.ntttp.information.

[3] underscore.Inc.,“Arrived!,” http://us.classmethod.jp/apps/arrived.

[4] amay077, “HexRinger,” https://play.google.com/store/apps/details?id=com.amay077.android.hexringer.

[5] T. Fujita, et. al., “Low Complexity TOA Localization Algorithm for NLOS Environments,”IPSJ SIG Technical Reports, Vol.2007, No.74, pp.69–74 (2007).*(in Japanese)*

[6] T. Mogi, et. al., “TOA Localization using RSS Weight with Local Attenuation Constant Estimation in NLOS,” IEICE technical report, Vol.107, No.53, pp.43–48 (2007).*(in Japanese)*

[7] T. Kitasuka, et. al., “An Implementation of Wireless LAN based Indoor Positioning System WiPS,” Proc. of DICOMO2004, pp.349–352 (2004).*(in Japanese)*

[8] H. Konishi, et. al., “A Study on Accuracy of lndoor Fingerprint Localization using WiFi,” Proc. of DICOMO2013, pp.1111–1115 (2013).*(in Japanese)*

[9] Y. Wada, et. al., “A Pedestrian Position Estimation Method using Laser Range Scanner and Wi-Fi Fingerprint,” IPSJ SIG Technical Reports, Vol.2013, No.26, pp.1–7 (2013).*(in Japanese)*

[10] R. Kiyohara, et. al., “Location Detection with Low Power Consumption for Obtaining Context Data for Mobile Devices,” IPSJ SIG technical reports, Vol. 2008, No. 44, pp.33–38 (2008).*(in Japanese)*

[11] T. Nakagawa, et. al., “Evaluation of Variable Interval Positioning Method for Power-saving Geofencing,” Proc. of DICOMO2013, pp.1116–1122 (2013).*(in*

Japanese)

[12] S. Seko, et. al., “An Algorithm to Estimate the Movement Timing to Indoor or Outdoor using Temperature Sensor,” IEICE technical report, Vol. 110, No. 450, pp.131–136 (2011).*(in Japanese)*

[13] C. Lee, et. al, “Energy-efficient Location Logging for Mobile Device,” Proc. of SAINT2010, pp.84–90 (2010).

[14] C. Fritsche, et. al., “Hybrid GPS/GSM Localization of Mobile Terminals using the Extended Kalman Filter,” Proc. of WPNC2009, pp.189–194 (2009).

[15] T. Farrell, et. al., “Energy-efficient tracking of mobile objects with early distance-based reporting,” Proc. of MobiQuitous2007, pp.1–8 (2007).

[16] I. Constandache, et. al., “EnLoc: Energy-Efficient Localization for Mobile Phones,” Proc. of INFOCOM2009, pp.19–25 (2009).

[17] Y. Chon, et. al., “LifeMap: a smartphone- Based context Provider for Location-Based services,” IEEE Pervasive Computing, Vol. 10, No. 2, pp.58–67 (2011).

[18] S. Yoshimura, et al., “Investigation of the comfort temperature and adaptive model in the houses,” Proc. of Architectural Research Meetings, Kanto Chap., AIJ, Vol. 82, No. II, pp.113–116 (2012).*(in Japanese)*

[19] P. Zhou, et al., “IODetector: A Generic Service for Indoor Outdoor Detection,” Proc. of SenSys’12, pp.113–126 (2012).

A Station Assignment Method Considering Applications Being Used in a Mixed Environment of Different Wireless Communication Services

Eiichi Kameda†, Hideo Kobayashi†, and Norihiko Shinomiya†

†Graduate School of Engineering, Soka University, Japan
{e07d5203, e15m5212}@soka-u.jp, shinomi@soka.ac.jp

Abstract - While the LTE has been in widespread use with high-speed data communications, some evolving applications tend to generate the exploding data traffic increasing the utilization factor of public WiFi services. The round trip time (RTT), however, might become longer because of throughput degradation or heavy load in a network. This paper defines a RTT gap as the difference between the required RTT for a user's application and the actual RTT obtained in connecting a mobile terminal to a station. The paper firstly formulates a problem to assign some mobile terminals with the different communications services and proposes a resolution logic to reduce the RTT gap in the whole system based on Hungarian method.

Keywords: WiFi Service, Round-Trip Time, Assignment Problem

1 INTRODUCTION

High-speed wireless communication technology for mobile phones like Long-Term Evolution (LTE) has been spreading gradually with improved capacity and speed. However, the amount of data required by applications on high-end data terminals such as smartphones also has been increasing. This tendency probably make a user feel longer response time depending on how crowded communication environments are with terminals. In order to avoid this surge of data traffic, a telecommunications operator has been implementing data offloading from 3G/LTE to the complementary network technology like WiFi. Thus the utilization of public WiFi service is expected to grow rapidly. If a large number of mobile terminals use the public WiFi service, the overload of WiFi base stations and heavy traffic in a backbone network might cause the degradation of the response time. Hence, this research proposes an evasion scheme of aggravation of the response time provoked by the data traffic surge in a public WiFi base station.

There are various researches about the improvement of response time in public WiFi service. Nevertheless, the data traffic needed by applications is increasing much more and the response time required by users is getting shorter in parallel with development in technology concerned with communication method. Therefore, this paper considers the optimal combination of the base station and terminals in a static communications environment where all mobile services available for a user remain unchanged.

In order to prevent the degradation of response time for users in a certain area as a whole, it is expected to make each terminal connect to a base station, which provides bare essentials of response time required by the terminal, as long as the number of base stations of the wireless-communications service, capacity of terminals, round trip time (RTT) in a base station and available bandwidth remain unchanged in the target area. Some undesired situations such as the following probably happen. A terminal can not connect with an acceptable base station, which meet a RTT requirement of the terminal due to other connected terminals, which can meet RTT requirement even if they connect to other base stations providing slower RTT. The elimination of these mismatches is thought to lead to the improvement of response time for the users in a target area as a whole. Thus, the purpose of this research is set to be the dissolution of such mismatches.

The proposed method focuses on the difference of RTT required by the application currently used on a terminal for mismatch dissolution. Classifying the RTT required by applications on a mobile terminal,the system can determine a base station to which a terminal connects, and reduce the excessive service assignment. As for proposed system, a designated server calculates the optimal base station for each terminal, and transmits control information to the terminal based on the application currently used on the terminal.

2 PROBLEM FORMULATION

2.1 Definition of RTT Gap

The response time when using a mobile terminal is likely to get worse when the RTT in connecting a mobile terminal to a base station becomes longer than the minimal RTT requirement for an application currently used on the terminal. Then, this study defines a RTT gap as the difference between the above two kinds of RTTs (required RTT and RTT for connection) and aims at reducing the RTT gap as a whole in a target area. An RTT gap (G_{RTT}) is denoted by the formula (1) from required RTT (RTT_{need}) and RTT for connection (RTT_{link}).

$$G_{RTT} = RTT_{link} - RTT_{need} \quad (1)$$

2.2 RTT Required by Applications

A required RTT depends strongly on the kind of applications. Thus, we classify applications into the following three kinds of typical applications; a telephone call, a browser and others (non real time).

For a telephone call, the Ministry of Internal Affairs and Communications in Japan has determined the standard of delay of an "050 IP phone" in less than $400ms$ [1]. This value provides an attainment time from an originating terminal to

a recipient one. Therefore, it makes sense that the required RTT from a mobile terminal to an access line is about $200ms$ as a half of $400ms$.

As for a browser, this study assumes a connection to a web site with $1,500kB$ contents, since the top pages of many web sites have $1,500kB$ or less of data. According to the report of Forrester Consulting, more than half of all users expect the response time below $2sec$ [2].

The throughput for displaying the $1,500kB$ contents in $2,000ms$ is $6mbps$ on ground that the throughput $T[mbps]$ for displaying $x[kB]$ of data in $y[ms]$ is denoted by formula 2.

$$T = 8x/y \tag{2}$$

Furthermore, since a required RTT $R[ms]$ can be expressed as formula (3) for $T[mbps]$ of throughput when a window size in TCP is $64kB$, the required RTT for $6mbps$ is estimated at $85ms$.

$$R = 64/8T \tag{3}$$

Then, this paper defines RTT required for telephone call application as $200ms$ and RTT required for a browser as $85ms$. The RTT for other applications is regarded as being unrestricted.

2.3 RTT for Connection

The correct RTT cannot be obtained unless a certain terminal actually connects with a base station. However, it is not realistic for each terminal to try to connect with all the base stations for measuring RTT. Then, this study employs PathQuick as a technique calculating the transmission speed in order to grasp RTT in each base station in real time [3]. Based on the throughput calculated by PathQuick and formula 3, RTT from each base station to an access line can be estimated.

2.4 Minimization of RTT Gap

Suppose x, y and z indicate the number of mobile terminals that base stations sa, sb and sc can accommodate respectively, there would exist access points of sa_1,sa_2,....sa_x,sb_1,sb_2,.... sb_y,sc_1,sc_2,....sc_z. This leads to the total capacity $m = x + y + z$ for the accommodation. If there are n terminals of t_1,t_2,...t_n in the same area, our proposed logic determines the assignment of n terminals to n access points extracted from m ones with minimizing the sum of RTT gaps. The combinatorial number can be expressed as ${}_{(n+p-1)}\mathrm{C}_{(p-1)}$, where p is the number of access points and the capacity for accommodation is not limited. Our method employs the Hungarian method solving the above assignment problem for all possible combinations of n access points that can be drawn from m ones [4][5].

3 RELATED RESEARCH

A previous research in [6] proposes the heuristics logic which determines the access point to be connected by a terminal according to delay of each access point, a throughput and so on. Another research in [7] proposes a system which can choose an access point to be connected in consideration of actual traffic load status embedded in a new field of the beacon information exchanged between a terminal and the access point. However, neither of them are taking into consideration that a required environment depends strongly on an application currently being used. This paper examines an assignment method of access points in consideration of required RTT differing according to not a transport layer protocol but a kind of applications.

As for the WiFi access point selection considering the applications currently being used, a research in [8] proposes a system to choose a suitable route based on radio field strength, available bandwidth, delay and so on. But this proposal does not consider the situations for a large number of mobile terminals. In this paper, we proposes the system to assign a terminal appropriately to a base station by regarding the combination of connection with the base station in the environment which exists many terminals, as a assignment method, and utilizing the Hungarian method.

4 ENVISIONED SYSTEM ENVIRONMENT

This proposed system are processed according to the following procedure.

1. Calculation of RTT for connection by PathQuick
2. Calculation of RTT gap
3. Determination of a base station to be connected
4. Control of a terminal

Also, this system employs the concentrated control type as a system configuration not the distributed control type. This is because the RTT obtained in connecting to each base station changes according to the number of terminals connecting to the base station and cannot be judged correctly from the information on the terminal side. In this system, control server gathers information of each base station, and it determines the base station to which each terminal connects. Moreover, our purpose is suitable assignment to a base station using an exchange of control server and a terminal, without adding change for the base station itself such as LTE and WiFi.

4.1 Calculation of RTT for Connection by PathQuick

In order to grasp RTT for connection when each terminal connects with each base station, it is necessary to grasp RTT in the base station which the terminal has not connected now in real time. In this paper, RTT for connection is grasped in real time by using PathQuick. The system configuration of calculation of the RTT which uses PathQuick is shown in Fig 1. The terminal for use in PathQuick is always connected in each base station. Moreover, control server is installed in an access line side. From the terminal for use in PathQuick linked to each base station, a packet is periodically transmitted to control server. Based on the receiving interval of the received packet, control server grasps the throughput(T_i) from

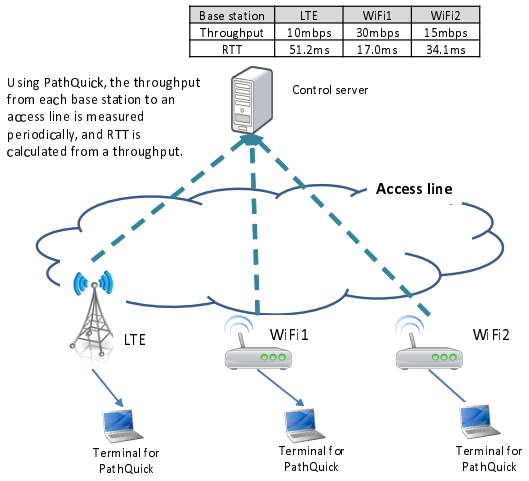


Figure 1: The system configuration of the RTT calculation which uses PathQuick

each base station to the access line in real time, and it calculates RTT for connection on each base station using the throughput and the formula 3.

4.2 Calculation of RTT Gap

As mentioned earlier, RTT for connection with each base station can be calculated. Moreover, required RTT in each terminal is calculated from Table 2.2 based on the application currently used at each terminal. RTT gap is calculated based on RTT for connection and required RTT. In addition, because actual RTT, which is shorter than required RTT should not contribute to improvement of utilization efficiency as a whole system, all negative values on RTT gap are transposed to 0 here.

4.3 Determination of a Base Station to Be Connected

sa, sb, and sc are base stations which is connectable in a certain area. Each number of terminals that each base station can accommodate are x, y, and z respectively. The total capacity of terminals by all stations m is calculated below.

$$m = x + y + z \tag{4}$$

Moreover, the terminals which exist in the area are set to t_1,t_2,.... and t_n. The optimal combination of a base station and a terminal is calculated by solving a assignment method (Hungarian method) to regard the pair of the access point which only n unit extracted from m unit and the terminal of the n unit so that the sum of the RTT gaps is minimized. Same calculation is performed about all the combinations of n access points extracted from m ones. The solution with the minimum sum of RTT gaps is chosen as the global optimum after calculating it in all the combinations. When two or more solutions would be obtained, the solution with the smallest standard deviation could be chosen.

4.4 Assignment Method

This section explains the Hungarian method used by this paper. The Hungarian method is one of the useful techniques for an assignment method. The assignment method can be solved by performing the following procedures to the given cost matrix.

1. To the given matrix, the minimum value of the line is pulled from each element of each line, and the minimum value of the sequence is further pulled from each element of each sequence.

2. All 0 is covered by the vertical or horizontal line as few as possible. When the number of the line drawn at this time is the same as the size of a matrix or larger, processing is ended because one 0 can be chosen from each line and each sequence.

3. In 2), when the number of the drawn line is smaller than the size of a matrix, the minimum value of the element in which it is not being underlined with the line is pulled from the element in which it is not being underlined with the line. Moreover, the minimum of the element in which it is not being underlined with the line is added to the element with which the line has overlapped.

2) and 3) are repeated until it ends. Combination whose cost is the minimum can be derived by the above operation.

In this paper, the RTT gap to each base station of each terminal is considered as a cost matrix, and the combination which RTT gap minimizes is derived using the Hungarian method.

4.5 Control of a Terminal

Based on the combination computed for 4.3, the control information is transmitted to each terminal from control server. Each terminal changes an access point according to the received control information.

5 EVALUATION BY A SIMULATION EXPERIMENT

5.1 The Outline of an Experiment

We developed a simulation program for ”Calculation of RTT gap” and ”Determination of a base station to be connected” based on the logic described chapter 4. The number of terminals, the number of base stations, and the capacity of each base station are described in Table 5.1. RTT for connection of base station1, base station2 and base station3 are respectively $250ms$, $150ms$ and $100ms$. These values were obtained by PathQuick. ”Telephone call application”, ”Browser”, or ”other” are assigned to application currently used at each terminal at random. Moreover, required RTT for each application used the value of Table 2.2.

In the simulation program, the RTT gap in each terminal and each base station are calculated based on these values. Calculation by the Hungarian method is performed to all the

Case	Number of Terminals	Number of Base Stations	Number of Base Station accommodation		
			Base Station1	Base Station2	Base Station3
Case1	7	3	4	3	3
Case2	100	3	100	30	30
Case3	200	3	200	30	30
Case4	300	3	300	30	30
Case5	400	3	400	30	30
Case6	500	3	500	30	30
Case7	600	3	600	30	30
Case8	100	3	No Limit	No Limit	No Limit
Case9	200	3	No Limit	No Limit	No Limit
Case10	300	3	No Limit	No Limit	No Limit
Case11	400	3	No Limit	No Limit	No Limit
Case12	500	3	No Limit	No Limit	No Limit
Case13	600	3	No Limit	No Limit	No Limit

Table 1: The outline of an experiment

combination of selection of the number of terminals from the sum of the capacity of all base station. The combination of which the sum total of the RTT gap and standard deviation is the minimum is outputted. Moreover, in this experiment, calculation in each case was performed also with the following two logic other than the calculation by the proposal system (Solution 1).

1. First, each terminal are assigned to the base station 3 in numerical sequence of terminal until the number of terminal exceed the number of accommodation of the base station 3. Then, each terminal are assigned to the base station 2 in numerical sequence of terminal until the number of terminal exceed the number of accommodation of the base station 2. Finally, each terminal are assigned to the base station 1. (Solution 2)
2. Each terminal are assigned to each base station at random. (Solution 3)

In addition, the Solution 2 assumes the system where the terminals connect to WiFi preferentially regardless of the application currently used.

5.2 Evaluation of an Experiment

Based on the experimental result, we evaluate by two viewpoints that is 1) the average value of the RTT gap after assignment to base station and 2) the processing time required to assign.

5.2.1 Average Value of the RTT Gap after Assignment to Base Station

In all the cases, the average value and the standard deviation of the RTT gap of the solution 1 (this proposal system) are the both smallest among three solutions. It indicates that this proposal system are effective in minimization of the RTT gap.

5.2.2 Processing Time Required to Assign

The number of times of execution and the processing time of the Hungarian method in each case are shown in Table 5.2.2.

If the Hungarian method is not used, the calculating in a huge number of terminals substantially is difficult. By using the Hungarian method, the number of calculation can be reduced.

Case	Terminals	Limit of accommodation	Number of execution for the Hungarian method	Processing time (second)	Number of calculation when not using the Hungarian method
1	7	Yes	10	0	1,330
2	100	Yes	961	0	${}_{100}C_{30} *_{70} C_{30} = 1.6 * 10^{45}$
3	200	Yes	961	1	${}_{200}C_{30} *_{170} C_{30} = 8.3 * 10^{68}$
4	300	Yes	961	1	${}_{300}C_{30} *_{270} C_{30} = 1.1 * 10^{81}$
5	400	Yes	961	2	${}_{400}C_{30} *_{370} C_{30} = 1.8 * 10^{89}$
6	500	Yes	961	3	${}_{500}C_{30} *_{470} C_{30} = 3.1 * 10^{95}$
7	600	Yes	961	5	${}_{600}C_{30} *_{570} C_{30} = 3.3 * 10^{100}$
8	100	No	5,151	1	$3^{100} = 5.2 * 10^{47}$
9	200	No	20,301	11	$3^{200} = 2.7 * 10^{95}$
10	300	No	45,451	52	$3^{300} = 1.4 * 10^{143}$
11	400	No	80,601	170	$3^{400} = 7.1 * 10^{190}$
12	500	No	125,751	459	$3^{500} = 3.6 * 10^{238}$
13	600	No	180,901	1030	$3^{600} = 1.9 * 10^{286}$

Table 2: Processing time of each case

6 CONCLUTION

We defined the RTT gap in consideration of the application which the user is using, and formulized the connection base station determination logic for reducing the RTT gap. Moreover, we proposed the connection base station determination system using the Hungarian method, and conducted the experiment by a simulation program and showed the validity from two viewpoints, minimization of the RTT gap and processing time. From now on, we will also run a simulation about ”calculation of RTT by PathQuick”, ”control of a terminal”, and will evaluate a whole system.

REFERENCES

[1] Ministry of Internal Affairs and Communications, http://www.soumu.go.jp/main_content/000158162.pdf.

[2] Forrester Consulting : eCommerce Web Site Performance Today, http://www.damcogroup.com/white-papers/ecommerce_website_perf_wp.pdf.

[3] T.Oshiba and K.Nakajima, Quick and simultaneous estimation of available bandwidth and effective UDP throughput for real-time communication, Computers and Communications (ISCC), 2011 IEEE Symposium on, pp.1123-1130(2011).

[4] Alan Doran,Joan Aldous : Networks and Algorithms: An Introductory Approach, Wiley(1994).

[5] Ravindra K. Ahuja,Thomas L. Magnanti,James B. Orlin : NETWORK FLOWS: Theory, algorithms, and Applications, Prentice Hall (1993).

[6] Gaurav S. Kasbekar,Pavan Nuggehalli,Joy Kuri : Online Client-AP Association in WLANs, Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006 4th International Symposium on, pp.1-8 (2006).

[7] Huazhi Gong,Nahm, K.,JongWon Kim : Distributed Fair Access Point Selection for Multi-Rate IEEE 802.11 WLANs, Consumer Communications and Networking Conference,5th IEEE, pp.528-532 (2008).

[8] Ryuichi Takechi,Koji Ogawa,Masato Okuda: Use of Self-Organizing Networks to Optimize Radio Access Networks, http://www.fujitsu.com/downloads/MAG/vol48-1/paper15.pdf

An Evaluation of the Influence of Communication Metrics in Realizing Multi-path Routing in Consideration of the Communication Situation on an Ad hoc Network

Yuki Asanuma[*], Yoshitaka Nakamura[**] , and Osamu Takahashi[**]

[*]Graduate School of Systems Information Science, Future University Hakodate, Japan
[**]School of Systems Information Science, Future University Hakodate, Japan
{g2114001, y-nakamr, osamu}@fun.ac.jp

Abstract - In recent years, with the development of wireless technologies, ad hoc networks, which are formed without any central administration and consist of mobile terminals, have begun to be used in a variety of applications. Clustering is one of the techniques useful for ad hoc network routing, because clustering techniques can manage wireless communication terminals hierarchically in an ad hoc network. The technique of routing using clustering can create a communication path which realizes stable communications. Furthermore, routing using Cluster-by-Cluster technique, which is an extension of Clustering, can improve packet arrival rate and reduce routing overheads. However, Cluster-by-Cluster routing does not respond to changes in the communication state. Therefore, there is a problem in that overall network throughput decreases in the network environment where there are multiple communications. In this paper, we propose a flexible route selection method to improve the overall network throughput. We also evaluate the influence of various network metrics based on traffic amount in our method.

Keywords: Ad hoc network, Clustering, Traffic. Path switching, multiple communications

1 INTRODUCTION

In recent years, the use of ad hoc networks has spread due to the progression of wireless communication and increasingly high performance of devices. Ad hoc networks do not need infrastructures such as base stations. Additionally, an ad hoc network will autonomously form a wireless terminal network. The reactive protocols DSR (Dynamic Source Routing) [2] and AODV (Ad hoc On-Demand Distance Vector) [3] are well-known routing protocols used on ad hoc networks. With these routing protocols, as the data communication path becomes longer, there is a decrease in the packet arrival rate and an increase in the routing overhead. Therefore, long communication paths cause congestion on ad hoc networks when there are multiple communications. As a result, communication reliability decreases. Thus, in order to realize stable communications, it is necessary to suppress the occurrence of long-path communication. Routing using clustering is an effective method in this case [4]. Clustering is used to form a group of physically close nodes on the network. This group is called a cluster. Clustering can streamline processes, such as the building of routes at the time of communication, more effectively than Non-cluster-based routing. In respect to this cluster-based routing, Cluster-by-Cluster routing [5] has further improved the suppression of long communication paths. Cluster-by-Cluster routing is able to generate and combine multiple fresh short paths by utilizing the mechanism of the route cache. Thereby, the method has solved the issues of reduction in packet arrival rate and increase of the overhead caused by long-path communication. However, Cluster-by-Cluster routing continues to use one path from the start of communication until the end and has no mechanism by which to switch paths. The communication status of nodes will change at any time on a network on which multiple communications exist. Consequently, the throughput is reduced due to communication congestion and increase in traffic.

In this paper, a path-switching mechanism is added to Cluster-by-Cluster routing in order to maintain the throughput of the entire network, for networks on which multiple communications exist. We selected effective metrics for the path-switching mechanism and evaluated their effect on these networks using the network simulator ns-2.

2 RELATED WORK

2.1 Cluster-by-Cluster Routing

Cluster-by-Cluster routing builds an overlay network on a cluster in addition to the recorded path relay node present in a normal ad hoc network. It performs routing between clusters on the overlay network. Fig. 1 shows an example of Cluster-by-Cluster routing.

Cluster-by-Cluster routing forms a cluster of physically close nodes around a central node called the CH (Cluster Head), in the same manner as typical cluster-based routing. Every cluster has one CH which centralizes the information of the nodes in the cluster. Also, when communicating with different clusters, the CH will designate the node located at the border between clusters as a CG (Cluster Gateway), which is used to bridge the communications between clusters. Nodes other than the CH and CG are set as MN (Member Node).

When carrying out data communication between nodes, the communication travels cluster-by-cluster from the source node cluster to the destination node cluster (Fig. 1). Thereby, Cluster-by-Cluster splits the long path from the source node to the destination node into multiple short paths at a unit of one cluster (Fig.2 and 3).

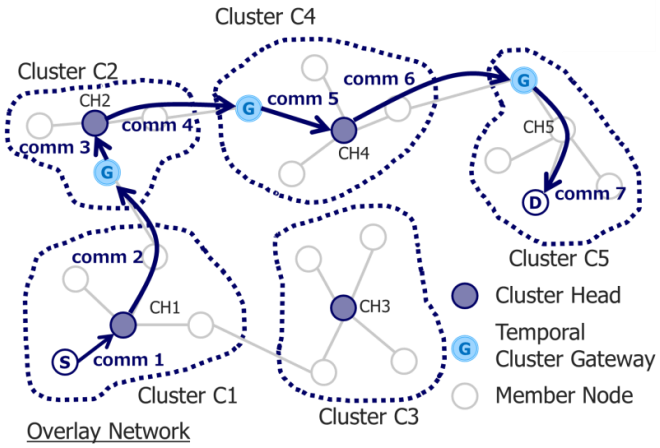


Figure 1: Cluster-by-Cluster routing

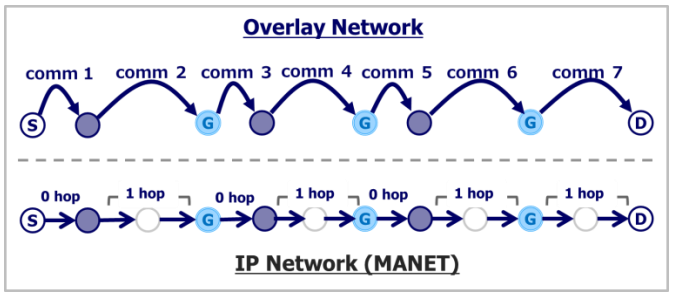


Figure 2: The communication division in Cluster-by-Cluster routing

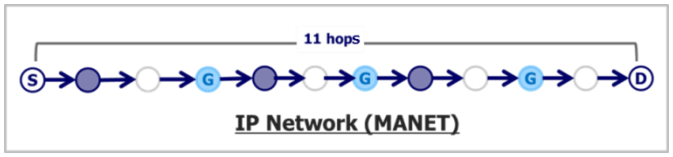


Figure 3: Communication path of traditional routing

2.2 Cluster-by-Cluster Routing Considering Adjacent Terminals

Cluster-by-Cluster routing is assumed to be used on a network which has a large number of nodes. However there is a case in which multiple communications exist, creating a risk that one communication may decrease the performance of other communications. To solve this problem, a route-switching function that makes use of certain network metrics was applied to Cluster-by-Cluster routing, thus creating Cluster-by-Cluster routing that considers the adjacent terminals[6]. In this approach, the number of adjacent terminals $\boldsymbol{N}$, average link cutting estimated time level $\boldsymbol{LLT}$, battery power $\boldsymbol{B}$, and the behavior of the node $\boldsymbol{T}$, are used in the following formula, to calculate the priority-level of a path.

$$\boldsymbol{W} = \boldsymbol{\omega_1} \times \frac{\boldsymbol{1}}{\boldsymbol{N}} + \boldsymbol{\omega_2} \times \frac{\boldsymbol{1}}{\boldsymbol{LLT_{average}}} + \boldsymbol{\omega_3} \times \frac{\boldsymbol{1}}{\boldsymbol{B}} + \boldsymbol{\omega_4} \times \frac{\boldsymbol{1}}{\boldsymbol{T_{role}}} \quad (1)$$

The priority $\boldsymbol{W}$ figures obtained for each path are compared and the route is switched dynamically to the path with the largest value of $\boldsymbol{W}$. As a result, the overall load of the network is dispersed, thereby improving throughput.

3 PROPOSED METHOD

3.1 Positioning of This Research

In this paper on a path-switching function using multiple metrics for Cluster-by-Cluster routing, we will evaluate the effect of each metric on ad hoc networks that do not form clusters. The proposed method was created with the original premise of forming clusters, but in this paper, as befits the circumstances of the experiment, cluster-formation is not covered.

3.2 Cluster-by-Cluster Routing for Performing Path Switching by Using the Traffic Situation in the Metrics

The Cluster-by-Cluster routing [7] proposed in previous studies by the authors of this paper (hereinafter, the proposed method), performs path-switching using the following three steps:

(1) Collection and measurement of metrics
(2) Calculation of priority of paths
(3) Switching of path

Cluster-by-Cluster routing is built based on the DSR, and can only build and maintain a single path. The proposed method extends the Cluster-by-Cluster routing by creating one primary path and two alternative paths. Generally, DSR only holds the path which RREP arrived at earliest, but the proposed method expands DSR and allows to hold paths which RREP at after the second. We assume the path which RREP arrived at earliest as the primary path and assume the path which RREP arrived after the second as the alternate paths. By switching pathways in response to the situation, this system stabilizes routing. For path-switching, the metrics needed to maintain the throughput are essential. The proposed method uses four metrics (the number of hops, the amount of traffic, link status and battery level) for path-switching.

After measuring the metrics at each node, the Neighbor Feedback with which Cluster-by-Cluster routing is equipped is used to pass metrics information to the source node. The proposed method uses Neighbor Feedback to transmit a beacon, to which metrics information has been added, from each node to the source node.

The metrics of number of hops are calculated using the 'hop count' function of the RREQ (Route Request). If movement of the nodes has caused the MN and the CG to move to a place where communication is impossible, the number of hops will change dynamically in Cluster-by-Cluster routing. However, this paper does not take into account the change in the number of hops, because the nodes are fixed. As for traffic, each node refers to its own

queue to examine the amount of packet being held, and then saves that amount as ***traffic***. Generally, the link state refers to the communication delay time between adjacent nodes. $\boldsymbol{Link}$ is the average delay time per hop, calculated by dividing the communication delay time by the number of hops of a communication path. Each node saves its own battery power as $\boldsymbol{B}$.

The source node calculates the priority $\boldsymbol{W}$ for each path based on the collected metrics, and uses this data as path selection criteria. $\boldsymbol{W}$ *is* calculated using the following formula:

$$W = \omega_1 \times \frac{1}{N} + \omega_2 \times \frac{1}{traffic} + \omega_3 \times \frac{1}{Link} + \omega_4 \times B \quad (2)$$

Here, the number of hops $\boldsymbol{N}$ is from the source node to the destination node, ***traffic*** is the sum of the traffic of each node in the corresponding path, and $\boldsymbol{Link}$ is the average delay time per hop between nodes in the corresponding path. $\boldsymbol{B}$ represents the average amount of remaining battery of the node in the corresponding path. In addition, each of the metrics is weighted by parameters from $\boldsymbol{\omega_1}$ to $\boldsymbol{\omega_4}$. By using these metrics, it is possible to consider the communication situation for a certain network easily, and to decide which metrics are most important for the network. Path switching is performed based on the flowchart depicted in FIG4. In the proposed method two types of path exist. The first is the primary path which communicates the packet to the destination node in response to the transmission request of the source node. The second is the alternative path, of which one or more will be constructed to act as potential candidates for the path-switching. At fixed times, the source node calculates the priority $\boldsymbol{W}$ of all the paths, primary and alternative. After the calculation, it sets the path with the largest value of $\boldsymbol{W}$ as the primary path and uses this path at the time of communication. If none of the primary and alternate paths is available, the source node once again sends a transmission request, to reconstruct the communication path (Fig. 5).

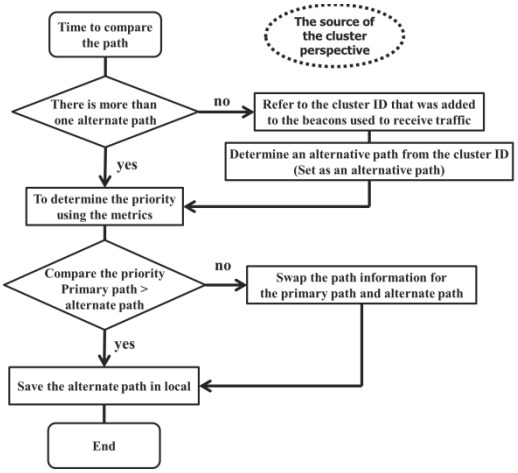


Figure 4: The path switching algorithm by priority

3.1 Metrics Based Strategy

In general, the number of hops from the source node to the destination node influences the throughput. However, the number of hops alone cannot respond to a case in which multiple communications are present, resulting in a decrease in throughput due to communication congestion caused by the concentration of traffic on a specific path. Consequently, it is believed possible to prevent a decrease in throughput by including the traffic amount in the metrics. This is because even if multiple communications are present, the concentration of communication on a specific path is prevented. Furthermore, the throughput is lowered when a communication delay is caused by various problems with the nodes. In this case, it is possible to prevent a decrease in throughput by creating a communication path that does not pass through any nodes in which communication delay is occurring. This communication delay is referred to as a link state and used as metrics. Nodes consume battery during communication and therefore is possible that the continuation of communication will become impossible if the residual battery amount is significantly reduced. In this case, the communication is disconnected to cause a reduction in rapid throughput. By considering metrics of remaining battery amount, it is possible to avoid using nodes with low remaining battery capacity, and thus long-term communication becomes possible [8].

4 SIMULATION RESULTS

4.1 Experiment

We conducted evaluative experiments to investigate the effect of routing using the metrics, in the network simulator ns-2(Network Simulator version 2) [9]. The simulation parameters are shown in Table 1.

This experiment was carried out on an ad hoc network with fixed nodes without forming clusters, in order to investigate the effect of the metrics on the network. However, in this experiment the destination node was configured to transmit a beacon message to convey each type of metrics information to the source node, in the same way as Cluster-by-Cluster routing. If the remaining battery power of the source and destination nodes is zero,

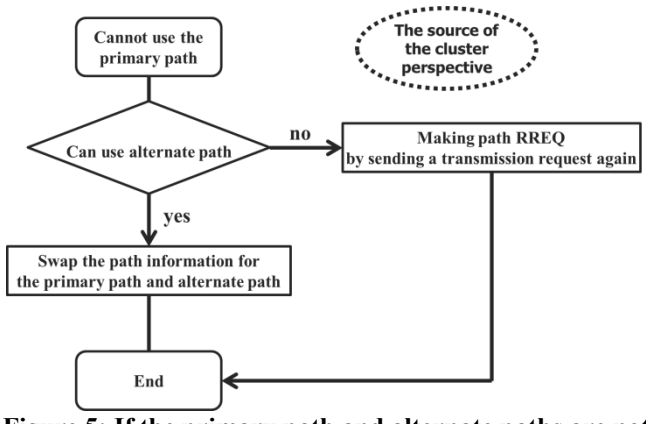


Figure 5: If the primary path and alternate paths are not available

communication becomes impossible regardless of the communication path. Therefore, the battery capacity of the source and destination nodes was twice that of the other nodes, to ensure that communication would not be interrupted during the experiment. The topology of the experiment is as shown in Fig. 6, with communications A, B and C shifting in time on the topology. Thereafter, it calculates priority $\boldsymbol{W}$ and switches to a path based on this. Priority $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ and $\boldsymbol{W_3}$ and $\boldsymbol{W_4}$ were determined in Experiments 1, 2, 3 and 4 as follows:

Experiment 1 : $$\boldsymbol{W_1} = \boldsymbol{\omega_1} \times \frac{\boldsymbol{1}}{\boldsymbol{N}} \quad (3)$$

Experiment 2 : $$\boldsymbol{W_2} = \boldsymbol{\omega_1} \times \frac{\boldsymbol{1}}{\boldsymbol{N}} + \boldsymbol{\omega_2} \times \frac{\boldsymbol{1}}{\boldsymbol{traffic}} \quad (4)$$

Experiment 3 : $$\boldsymbol{W_3} = \boldsymbol{\omega_1} \times \frac{\boldsymbol{1}}{\boldsymbol{N}} + \boldsymbol{\omega_2} \times \frac{\boldsymbol{1}}{\boldsymbol{traffic}} + \boldsymbol{\omega_3} \times \frac{\boldsymbol{1}}{\boldsymbol{Link}} \quad (5)$$

Experiment 4 : $$\boldsymbol{W_4} = \boldsymbol{\omega_1} \times \frac{\boldsymbol{1}}{\boldsymbol{N}} + \boldsymbol{\omega_2} \times \frac{\boldsymbol{1}}{\boldsymbol{traffic}} + \boldsymbol{\omega_3} \times \frac{\boldsymbol{1}}{\boldsymbol{Link}} + \boldsymbol{\omega_4} \times \boldsymbol{B} \quad (6)$$

Communication A communicates from 100s to 1200s. Communication B communicates from 500s to 1500s. Communication C communicates from 700s to 1400s. In all the experiments, calculation of priority and switching of path occurs at 750s. Table 2 shows the parameters relating to communication.

Table 1: Simulation parameters

Measurement time (seconds)	1500s
Network size	1200m × 1200m
Number of nodes	36
Communication protocol	UDP
Routing protocol	DSR

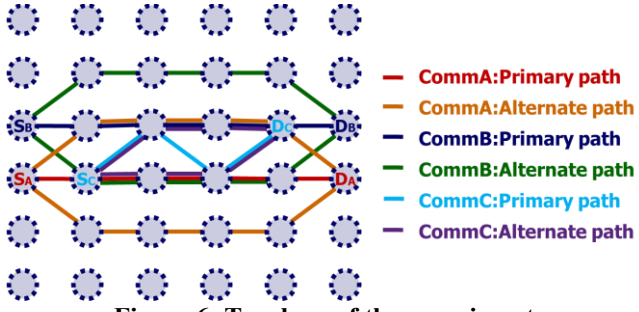


Figure 6: Topology of the experiment

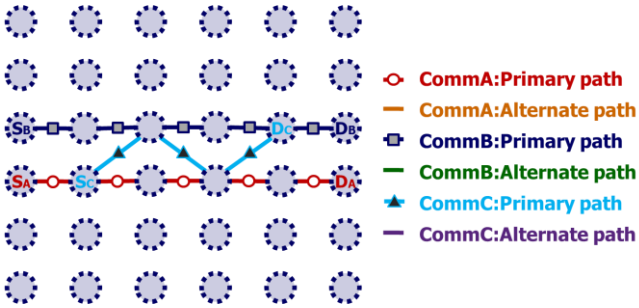


Figure 7: Experiment 1: Communication path

4.2 Experiment Results and Discussion

Results of path-switching in Experiment 1 and the communication paths are shown in Fig.7, and the throughput fluctuations of the entire network are depicted in Fig. 8. Priority $\boldsymbol{W_1}$ for path switching does not consider the number of hops. Therefore, it exhibited the same behavior as regular DSR in this experimental environment where the number of hops does not change. As a result, the path could not be switched, the throughput of the entire network was greatly reduced and communication became congested after 750s. Communication became impossible at 1000s because the remaining battery amount of the nodes used by communication A and communication C became zero. Therefore, communication B recovered throughput in 1000s and maintained a substantially constant throughput up to 1500s.

$\boldsymbol{W_2}$ has traffic added to the number of hops as metrics. We were using $\boldsymbol{W_2}$ as the priority value of Experiment 2 for path-switching. Figures 9 and 10 show the results of Experiment 2. The results reveal that path-switching was possible when the number of hops and traffic amount were considered. This is because communication could use a well-balanced selection of nodes, rather than being concentrated on fixed nodes. However, communication was temporarily disconnected for a moment when path switching and the throughput was drastically reduced. Communication A in Experiment 1 was disconnected due to low battery. However, in Experiment 2 communication became possible at 750s because the communication path switched to the alternative path. From this result, it can be seen that it is possible to carry out prolonged communication even without using remaining battery amount metrics, by using path-switching to communicate across multiple different nodes, in particular environment.

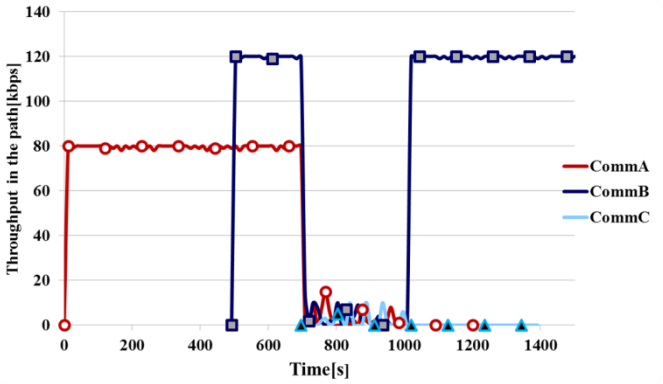


Figure 8: Experiment 1s: Throughput on path

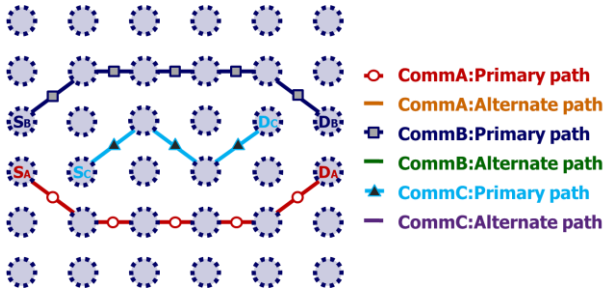


Figure 9: Experiment 2: Communication path

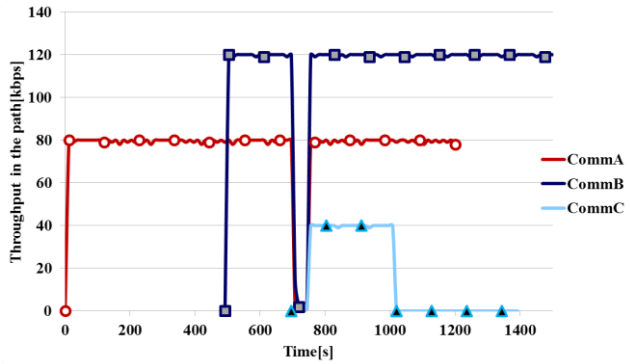


Figure 10: Experiment 2: Throughput on path

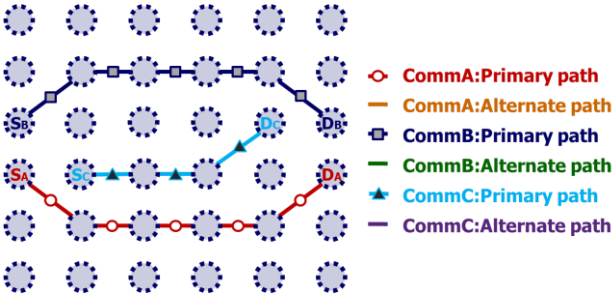


Figure 11: Experiment 3: Communication path

W_3 had link state added to the pre-existing metrics of number of hops and traffic volume. We used W_3 as the priority value of path-switching in Experiment 3. Fig. 11 and Fig. 12 show the results of Experiment 3. The results show that communication A and communication B did not change, but the communication path of communication C did change. The notes being set in a grid formation, DSR control messages arrived at vertically or horizontally adjacent nodes faster than diagonally adjacent nodes, due to the slightly longer distance in this direction. Despite this, looking the results in Fig. 12, it can be understood that there was no change in throughput when compared to Experiment 2. This is most likely because this experiment did not take into account the performance differences of each node and UDP communications were carried out in environments where there were no obstacles such as interference. Thus, the link state can be expected to be effective in environments where there exists a shield between the nodes, or physical performance of each node is different.

W_4 had remaining battery power added to the three aforementioned metrics. We used W_4 as the priority value in Experiment 4. Fig. 14 and Fig. 15 show the results of this experiment. Similar to Experiment 3, both the the

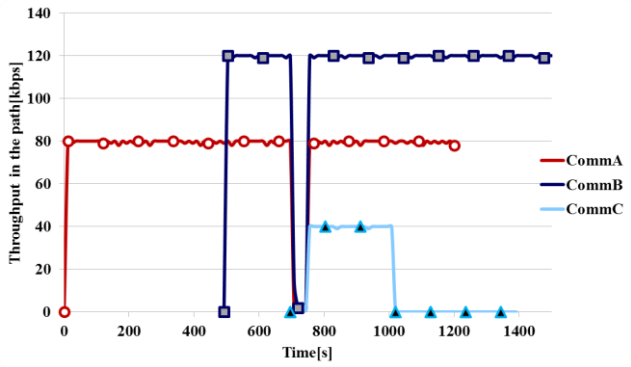


Figure 12: Experiment 3: Throughput on path

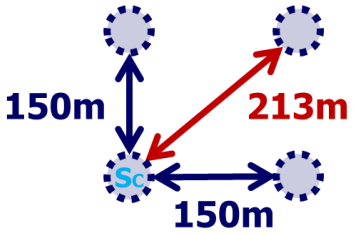


Figure 13: Distance between nodes

communication path and throughput of communication C changed. The communication path was been switched to a path with greater remaining battery power. As a result, the throughput of communication C did not decrease even after 1000 seconds, unlike in Experiment 3. The relay node used by communication C in Experiment 3 was used by communication A after 10 seconds, meaning that the remaining battery amount was lower than that of other nodes. It is expected that were communication C to continue to use nodes with little remaining battery capacity, the battery power would eventually become zero, causing the throughput to rapidly decrease. Using metrics of remaining battery capacity worked effectively to alter the route of communication C to pass through nodes with a large amount of remaining battery. In this way, reduction of communication throughput was prevented and the throughput of communication C is considered to have improved after 1000s.

Finally, the graph in Fig. 16 summarizes the throughput of the entire network in Experiments 1, 2, 3, and 4. The results confirm that when the amount of traffic is included in the metrics, the entire network throughput is improved and benefits from the effect. Furthermore, the remaining battery amount was also found to be necessary metrics in the case of long-term communication. However, link-state metrics did not work effectively and were not suitable for the assumed network environment. It is considered that these metrics would be effective in situations where there is an obstacle between nodes and a difference in the performance of each node, or where, as in MANET [10], the nodes move and the strength of the radio waves that nodes can receive is constantly changing. Through all of the experiments, the overall throughput is seen to fluctuate slightly. This is a result of the effect of the metrics collection beacon, and it is possible that the throughput may decrease further in a large network with many nodes.

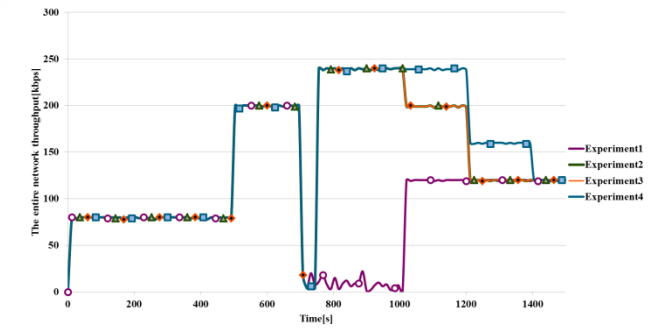


Figure 16: Overall network throughputs of Experiment 1-4

5 CONCLUSION

In this paper, we used the network simulator ns-2 to evaluate the influence on the network of adding a path-switching function to Cluster-by-Cluster routing. We confirmed that improvement of throughout of the entire network can be achieved by combining four metrics (number of hops, amount of traffic, link state and remaining amount of battery). In the future, it is necessary to evaluate whether the proposed method is effective in a network in which clusters are formed.

REFERENCES

[1] M. Kumar, R. Rishi, and D.K. Madan, Comparative Analysis of CBRP, DSR, AODV Routing Protocol in MANET, International Journal on Computer Science and Engineering, Vol.2, No.9, pp.2853-2858 (2010).

[2] D. Johnson, Y. Hu, and D. Maltz, The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4: RFC4728, http://www.ietf.org/rfc/rfc4728.txt.

[3] C. Perkins, E. Belding-Royer, and S. Das, Ad hoc On-Demand Distance Vector (AODV) Routing: RFC3561, http://www.ietf.org/rfc/rfc3561.txt.

[4] J. Y. Yu, and P. H. J. Chong, A survey of clustering schemes for mobile ad hoc networks, IEEE Communications Surveys & Tutorials, Vol.7, No.1, pp.32-48 (2005).

[5] H. Narumi, Y. Shiraishi, and O. Takahashi, A Reliable Cluster-based Routing Algorithm for MANET, Proceeding of the International workshop on Informatics (IWIN2009), pp.44-51 (2009).

[6] S. Suzuki, A Cluster-by-Cluster Routing Protocol Considered Information of Neighbor Nodes for Mobile Ad Hoc Networks, Graduate Research annual report, No.40 (2010).

[7] Y. Asanuma, Y. Nakamura, and O. Takahashi, A proposal of cluster-based routing algorithm considering traffic condition on MANET, Proceeding of the Multimedia, Distributed, Cooperative, and Mobile Symposium (DICOMO2014), pp.809 – 814 (2014) (*in Japanese*).

[8] N. Iguchi et al., Ad Hoc Communication Method for Farm Networks Considering Priority and Battery Remainder by IEEE802.11e, Agricultural Information Research, Vol.16, No.3, pp.81-90 (2007).

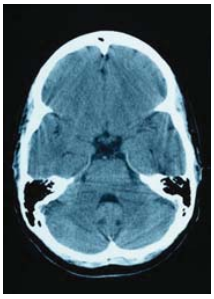
[9] The Network Simulator version 2 (ns-2), http://www.isi.edu/nsnam/ns/.

[10] M. Tauchi et al., Proposal of Routing Method with Route Break Avoidance and its Evaluation in MANET, IPSJ SIG Technical Report, Vol.2006, No.14, pp.25-30 (2006)(*in Japanese*).

Keynote Speech 1:
Prof. Yasue Mitsukura
(Keio University)

Keio University
1858
CALAMVS
GLADIO
FORTIOR

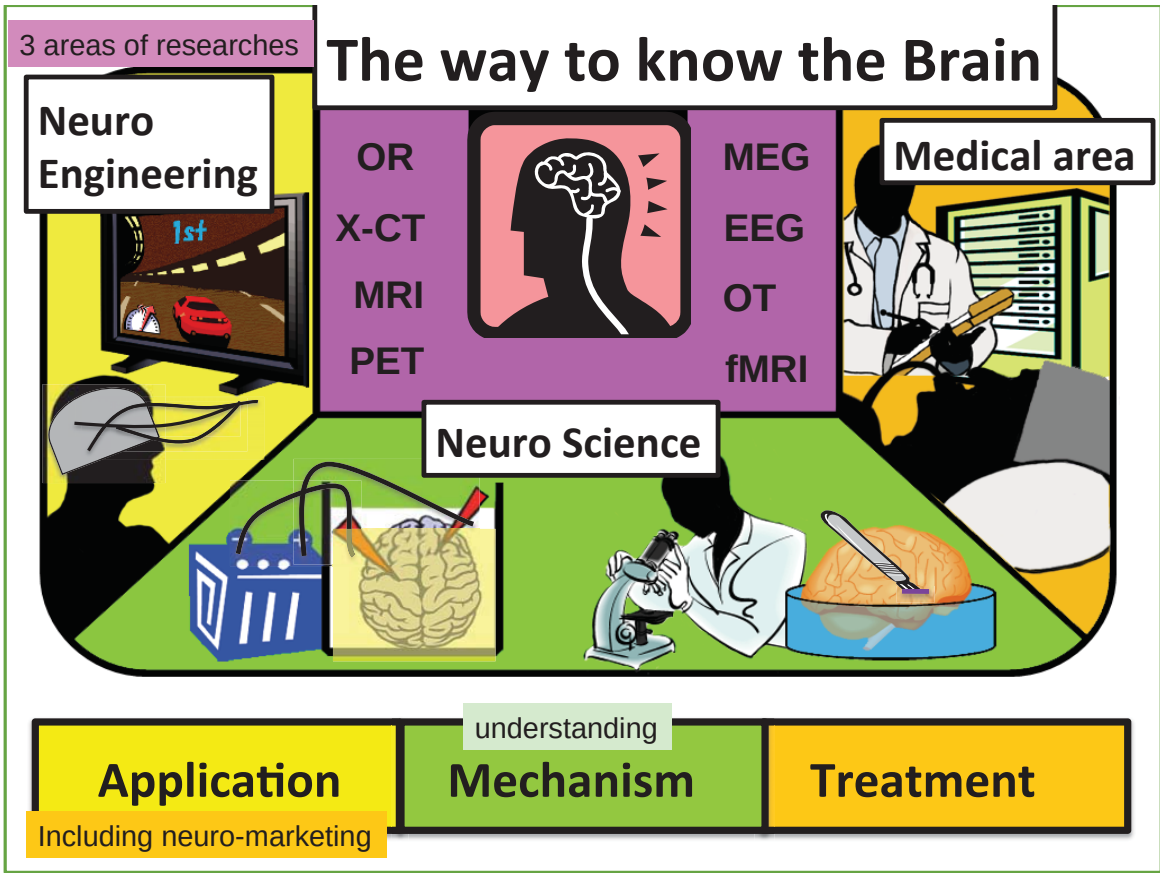
KANSEI detection and Its application using simple Device



Yasue MITSUKURA
KEIO University
mitsukura@sd.keio.ac.jp

Contents

- How to know the Brain
- Conventional device for getting the EEG (Brain wave) and application using the EEG.
- Our research and novelty of our research
  - using simple device and the strict signal processing
- Strict signal processing
- Apply to the real product and show the result
- Conclusions



Motivation

Different from sensibility or emotion. KANSEI alike "how to feel"

Our research

- getting the KANSEI and human situation using the EEG (brain wave).

(haptic, taste, olfactory, visual and acoustic sense, sleepiness, concentration, want, like/dislike, interest, stress)

- smart and simple BCI using the brain information.

f-MRI* *functional magnetic resonance imaging

✓Large-scale
✓Expensive
✓Binding

unusefulness

EEG** **electroencephalogram

✓Small-scale
✓Inexpensive
✓Nonbinding

useful

5

EEG(Electroencephalogram)

The action potential measured from scalp [μ V]

Detected data is translate to the frequency domain

frequency	band name	Each wave has a meaning (situation)
0.4～4	Delta wave	deep sleep
4～6	Theta wave	light sleep
7～8	Slow-alpha wave	resting/dazed
9～11	Mid-alpha wave	relaxed/concentrated/work well
12～14	Fast-alpha wave	nervous/unrelaxed
14～26	Beta wave	alert/working/active thinking

Usually, our human EEG is up to around 30 Hz

6

Research example using EEG

- Neuro-science (Science・Medicine area): Brain understanding
- Neuro-engineering (Engineering area): Meaning and intention detection (including the BCI)
- Neuro-marketing: EEG is used for obtaining the KANSEI.

<Neuro-engineering area >
Brain-controlled wheel chair with EEG
(Riken-Toyota project)

Problems : Device is heavy burden. No one can control without him.

Signal analysis is not enough.

Training system Using EEG

MENTAL Training by Brain Function Research center

This is just only threshold method.
People who can do it and can't.

Signal analysis is not enough.

EEG Mouse

Detected data is not EEG but EMG

All goods: EEG signal analysis is not enough.

Problems of conventional method

9

- Device problem.
- EEG signal analysis is not enough.
- Threshold method for frequency band is used for various situation.

- NEED an easy device for using everywhere
- NEED strict signal processing
- NEED signal processing for getting the meaning

We realize these NEEDS in my research

10

For NEED1 (device problem)

Simple EEG device is produced

- Reduction of the number of electrodes
  - ⇒ make the hair-band type device

  Needless of gel for the wearing
- Easy to wear, it takes 10 second.

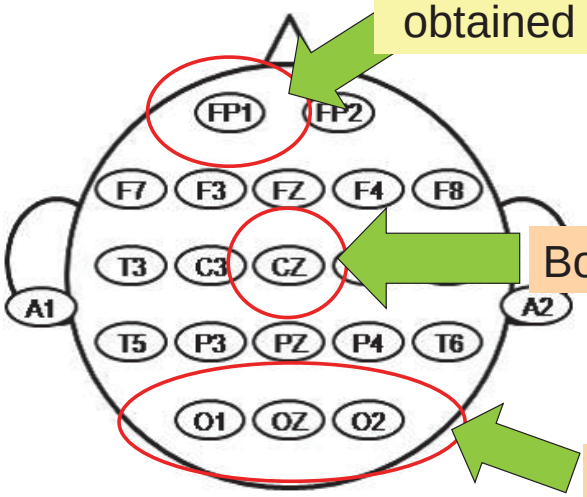
How to reduce the electrode

11

- EEG electrodes

KANSEI (How to feel) can be obtained around these area.

Our former research



Body movements

Visual information

Measurement point : International 10-20 system

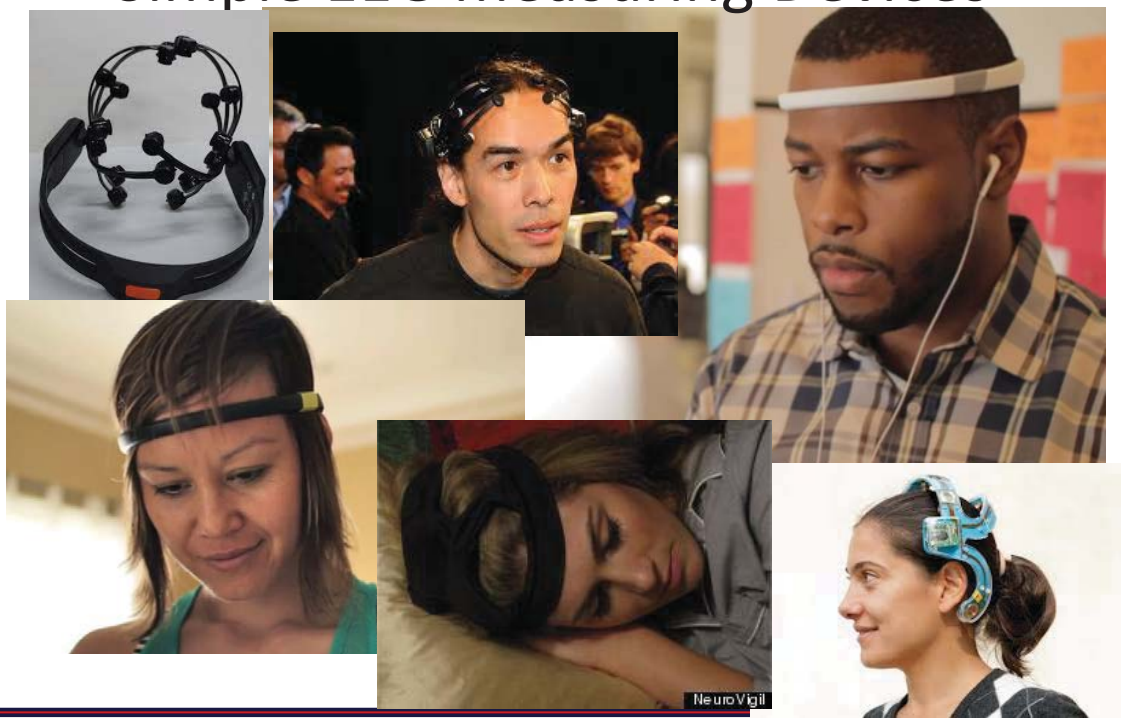
EEG measuring Device

12

- Conventional EEG systems
  - – Expensive, Large size, Stressful
    - 40 minuets for wearing
    - Intrusive gel
  - → Ineffective in real environments

- Simple EEG systems
  - – Easy & Compact
    - 120×135×35mm, 500g
  - – Measurement point : Fp
    - Left frontal cortex

Simple EEG measuring Devices



What device is the best ?

- I'm often asked to various person.
  - What device is the best ????

- My answer : Anything.
  - Our skill is signal processing.
  - If the raw data can be obtained from the device, no matter what device it may use, the same result can be shown.
  - Our novelty is robust signal processing.
  - Our system is structured by 17 years data par one situation.

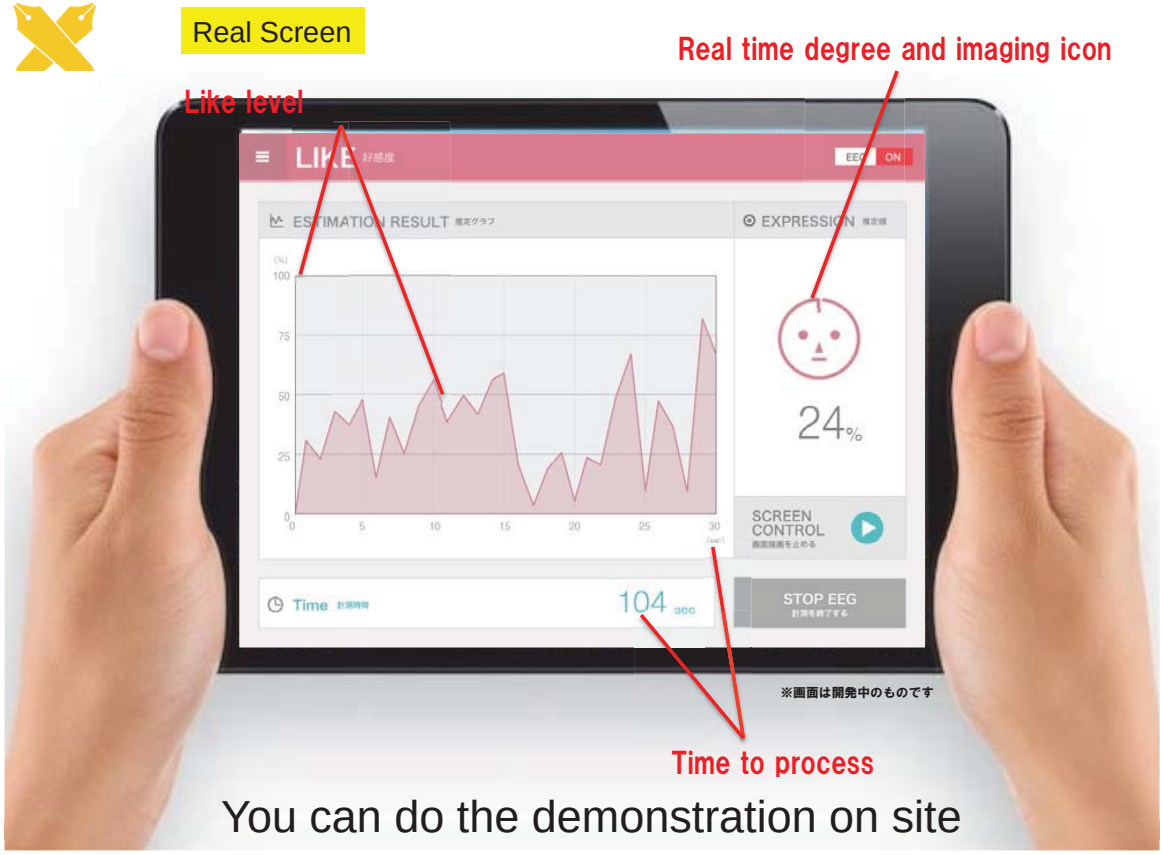
Noise Removable

- We must know
  - What is the noise ?
  - Noise is changeable in the same environment.
  - Most product currently called the simple EEG device are EMG (electromyograph) device.
  - If noises can be removed truly, no matter what device it may use, the same result is obtained.

we can remove the noise using the various device

Real Product : KANSEI Analyzer

Real Screen



You can do the demonstration on site

How to decide the KANSEI

- Concentration :

<conventional> used only alpha band

<proposed> rule extraction for 12 years data while “Azuki eperiment”

Neurocam

RECORDING
(UPPER LEVEL 60)

REMEMBER YOUR MOMENTS.
INTERESTS 0-100

NEUROCAM

by neurowear

Real Application

- NHK world news

Research example using EEG

23

- Neuro-science (Science・Medicine area): Brain understanding
- Neuro-engineering (Engineering area): Meaning and intention detection (including the BCI)
- Neuro-marketing: EEG is used for obtaining the KANSEI.

Real Application example

TV commercials

Important advertisement

- Many number of contacts between consumers and TV commercials
- Giving audio and visual information

➡Need producing effective TV commercials

Interesting for consumers

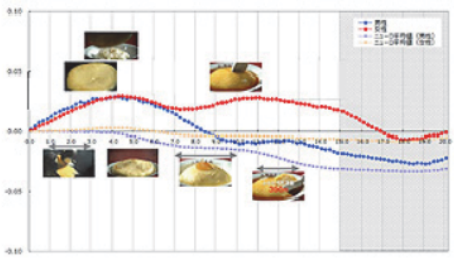
CM Evaluation using NIRS

日立ハイテクノロジーズ　http://www.hitachi-hightech.com/jp/product_detail/?pn=ot_007

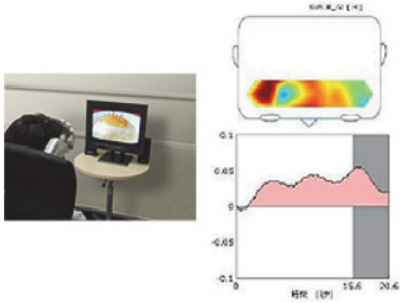
活用事例　～光トポグラフィ計測による広告評価の一例～

テレビコマーシャル　「大阪王将　ふわとろ天津飯」

株式会社インテージとの共同調査



濃活動の変化を男女別、時系列に表示した例。



脳活動の変化を部分的、時系列に表示した例。

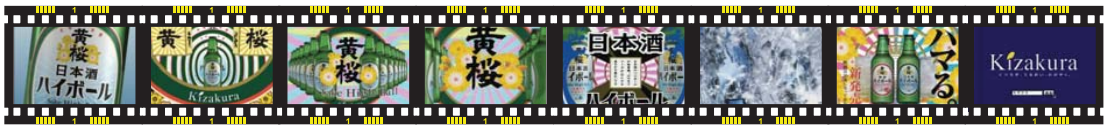
活用事例　～光トポグラフィ計測による広告評価の一例～

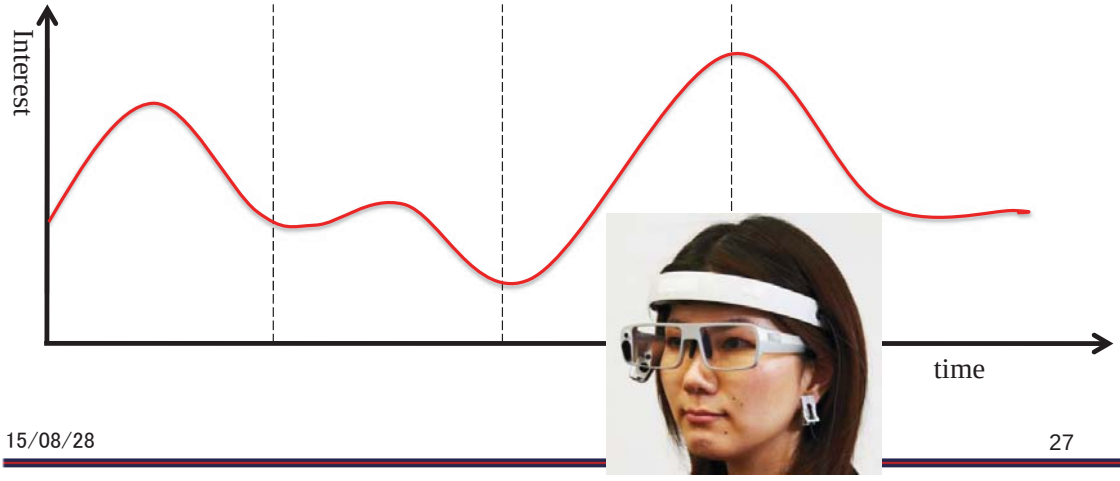
- 学習環境に適した知的照明
- デザイン評価
- 快・不快などの感性評価

Reference

http://www.hitachi-hightech.com/jp/product_detail/?pn=ot_007

Proposed method





15/08/28

27

Change of Interest

video

Conclusions & Future Works

29

< Our contributions >

- Simple EEG device is proposed.
- Strict signal processing can be done.
- Noise can be removed in the various devices.

Future Works

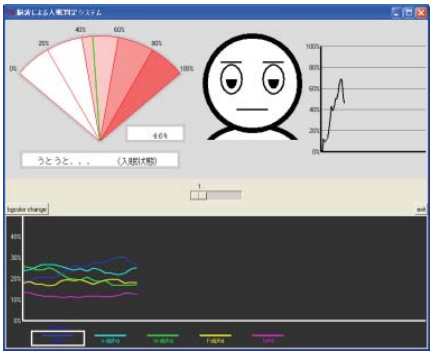
- Communication of only thinking

Keio University
1858
CALAMVS
GLADIO
FORTIOR

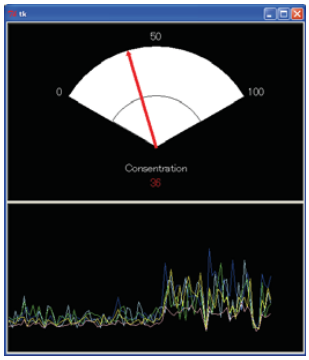
Thank you very much !

31

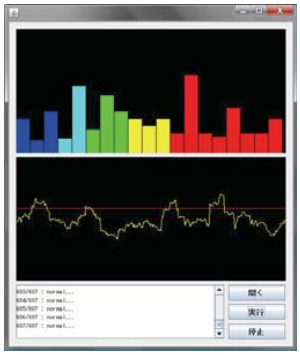
Visualization of Mental State Using EEG



Visualization of sleepiness



Visualization of degree of concentration



Visualization of stress level

Determine mode of analysis from the past research results

Sleepiness	→ Sleepiness in the 0 to 100 range
Degree of concentration	→ Degree of concentration in the 0 to 100 range
Stress level	→ Increase and decrease of stress level compared to criterion

32

Ubiquitous Information sensing

- Application implementation
  - Drowsiness
  - Concentration
  - Stress
- Demo. on the simulator
  - Image input
  - Touch input
  - Voice output
  - File input/output
- Demo. on the real devices
  - Touch input
  - Voice output
  - File input/output

33

Visualization of Sleepiness

Original application

DEMO

Recorded output

You can do the demonstration on site

Keynote Speech 2:
Dr. Makoto Imamura
(Mitsubishi Electric Corp.)

Data Analytics for Equipment Condition Monitoring

Makoto Imamura*, Takaaki Nakamura*, Michael Jones**, and Daniel Nikovski **

* Information Technology R&D Center, Mitsubishi Electric Corporation, Japan
** Mitsubishi Electric Research, USA
{Imamura.Makoto@bx, Takaaki.Nakamura@dy}.MitsubishiElectric.co.jp
{mjones, nikovski}@merl.com

Abstract - We have been developing a framework for equipment condition monitoring in order to deal with the growing need for preventive maintenance enabled by the rapid spread of the Internet of Things. This paper describes our framework for equipment condition monitoring. Our framework comprises three modules: feature generation, anomaly detection, and fault knowledge-base. The feature generation module has two characteristic functions: first, leg analysis that can extract the global trend pattern in time series with local fluctuations, so that it can capture the vibration depending on a given amplitude and a given window size; and second, statistical and smoothed trajectory (SST) features that can capture the shape and the stochastic behavior of the time series within the window, so that various types of sensor data can be processed. The anomaly detection module has two characteristic functions: first, Trend Pattern Query Language (TPQL) that can describe the anomaly detection rules that experts might have; and second, exemplar learning that can summarize the training time series with a small set of exemplars, so that it can perform fast anomaly detection. The fault knowledge base describes causal knowledge, so that the anomaly detection module can estimate failure factors by comparing statistical results with causal relations.

Keywords: Equipment Condition Monitoring, Feature Generation, Anomaly Detection, Fault Knowledge Base

1 INTRODUCTION

As the Internet of Things (IOT) [1] has been emerging and growing, sensor big data that is streamed from various equipment in power plant, industrial facilities, and buildings can be made available for monitoring, diagnosis, energy-saving, productivity improvement, quality management, and marketing. As a result, industry has paid much attention to the use of big sensor data generated from equipment or facility in order to create a smarter society.

Equipment Condition Monitoring (ECM) is a typical service that uses big sensor data, and machine learning techniques for big sensor data are key technologies to make ECM smarter[2].

We have been developing a data analytic framework for equipment condition monitoring. As shown in Figure 1, our framework comprises three modules: feature generation, anomaly detection, and domain knowledge representation. The feature generation module has two functions: first, leg analysis that calculates the frequency of variations; and second, statistical and smoothed trajectory (SST) features that represent the high and low frequency information in a window from a time series. The anomaly detection module has two functions: first, a rule-based fault detection based on Trend Pattern Query Language (TPQL); and second, a machine learning based fault detection module that provides efficient nearest neighbor outlier search with the help of

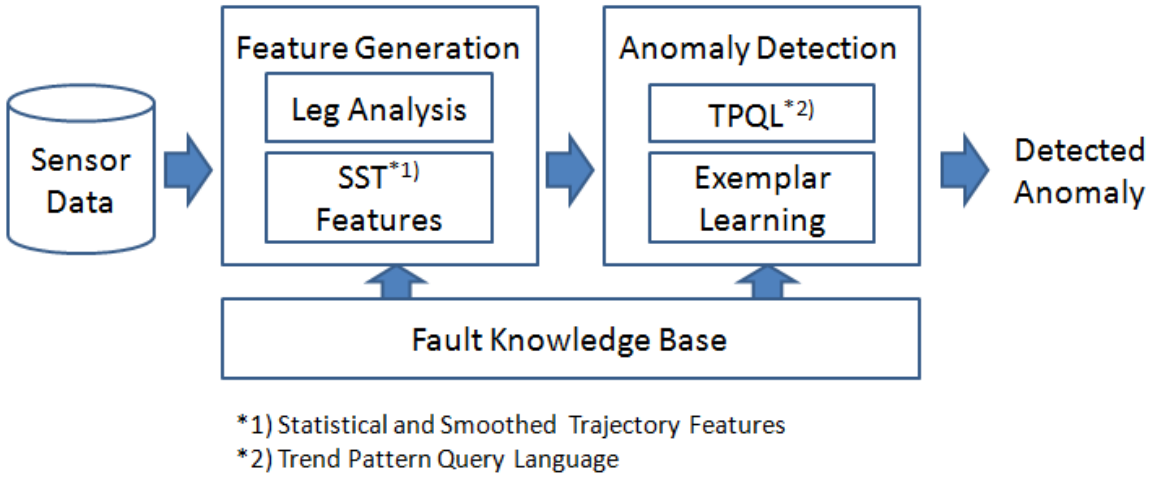


Figure 1: The structure of our framework for equipment condition monitoring.

exemplars. Fault knowledge is represented as a chain of causes and their effects that explain the detected anomaly.

The following part of this article has been organized as follows. Section 2 describes feature generation functions. Section 3 describes anomaly detection functions. Section 4 describes the fault knowledge base.

2 FEATURE GENERATION

2.1 Leg Analysis

Facility maintenance in a building, a plant, or a factory needs to calculate the frequency of variations in sensor data in order to detect a sign of failure or deterioration. Fink et. all proposed a leg search method [3] to find a global trend in a time-series including small variations such as noise. The dotted lines in figure 2 are examples of a leg. Both lines show the global upward trend that includes local up-down segments.

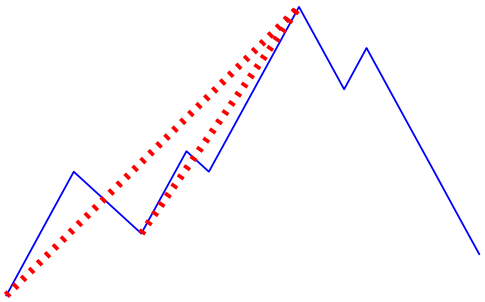


Figure 2: Leg

However, their method treats only single legs so that it can find an upward or downward trend, but can't calculate the frequency of variations. We developed leg variation analysis that can calculate the frequency of variations in time-series that includes upward trends and downward trends that can appear alternately and iteratively. We showed an algorithm whose calculation order is linear in the window size[4]. In contrast, the computational order of a naive algorithm is factorial in the square of window size.

(1) Leg Frequency

Definition: time series X, subsequence Y

A *Time Series* $X=[x_1,\cdots,x_m]$ is a sequence of real values. The value of the i-th time point is denoted by $X[i] = x_i$.

A Time Series *subsequence* $Y = [x_p, x_{p+1},...,x_q] = X[p:q]$ is a continuous subsequence of X starting at position p and ending at position q. We denote the starting time point, the ending time point, and the length of a subsequence by the following, respectively:

start(Y) = p
end(Y) = q
length(Y) = q-p+1

Definition: Leg

A *leg* is a subsequence X that satisfies the following conditions.

$\forall i.\ p \leqq i \leqq q \quad X[p] \leqq X[i] \leqq X[q] \quad \cdots\cdots(1)$
$\forall i.\ p \leqq i \leqq q \quad X[p] \geqq X[i] \geqq X[q] \quad \cdots\cdots(2)$

We define and denote the amplitude and sign of a leg X by the following, respectively:

amp(X[p:q]) = abs(X[q] - X[p])
sign(X[p:q]) = sign(X[q] - X[p])

Definition: Maximal Leg

A *maximal leg* is a leg X that satisfies the following conditions. If X[p-1] or X[q] does not exist, the corresponding conditions (3) or (4) are ignored:

$\forall i.\ p < i \leqq q \quad X[p] < X[i] \quad \cdots(1)$
$\forall i.\ p \leqq i < q \quad X[i] < X[q] \quad \cdots(2)$
$X[p-1] \geqq X[p] \quad \cdots(3)$
$X[q] \geqq X[q+1] \quad \cdots(4)$

Definition: Leg Vibration Sequence

Let $X_1, X_2,..,X_n$ be maximal legs, and A a positive real number. A *leg vibration sequence with amplitude A* is a leg sequence $s =[X_1, X_2,..,X_n]$ that satisfies the following conditions.

$\text{For } 1 \leqq i \leqq n-1 \quad end(X_i) \leqq start(X_{i+1}) \quad \cdots(1)$
$amp(X_i) \geqq a \quad \cdots(2)$
$sign(X_i) \times sign(X_{i+1}) < 0 \quad \cdots(3)$

We define and denote the length, the first leg, the last leg, the sign, the starting time point, and the ending time point of a leg variation sequence s by the following, respectively:

length(s) = n
first(s) = X_1
last(s) = $X_{length(s)} = X_n$
sign(s) = sign(first(s))
start(s) = start(first(s))
end(s) = end(last(s))

Definition: Leg Variation Sequence Set S(X,A,W,t)

Let X, A, W, and t be a time series, a positive real number, a positive integer, and a point of X, respectively. A *Leg Variation Sequence Set with amplitude A and window size W* is defined as follows:

$S(X,A,W,t) = \{\ s \mid s \text{ is a leg variation sequence with amplitude A and } t \leqq start(s) \text{ and } end(s) \leqq t + W - 1\ \}$

Lemma: The signs of leg variation sequences that have maximal length in S(X,A,W,t) are equal.

Definition: Leg Frequency F(X,A,W, t)

Let S(X,A,W,t) be a leg variation sequence set.

A *Leg Frequency of time series X at t with amplitude A and window size W,* F(X,A,W,t), is defined as follows:

$F(X,A,W,t) = sign(s_{max}) \times length(s_{max})$
such that
$s_{max} = \text{argmax}_{s \in S(X,A,w,t)}\ length(s)$

The lemma makes the above definition of leg frequency well defined.

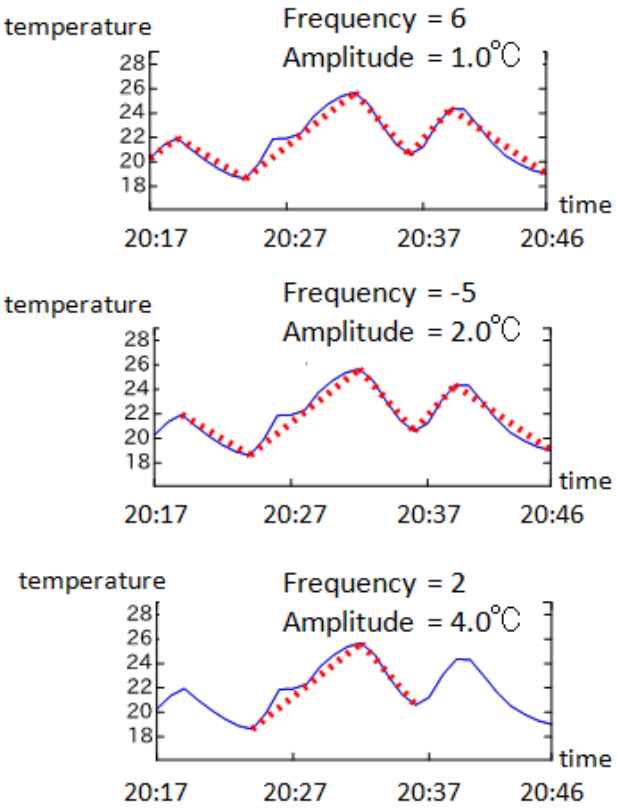


Figure 3: Trend Graphs of Experimental Data

Leg frequency qualifies the vibration of a time series subsequence, when amplitude and window size are given. Larger amplitude means larger width of variation. The sign of frequency is the sign of the first leg of the longest leg sequence. That is, a positive frequency at time point t with window size W means that the first leg has an upward trend in the subsequence X[t:t+w-1]. Similarly, a negative frequency means that the first leg is downward. A frequency of 2 means that the subsequence has one convex pattern whose amplitude is larger than A, while a frequency of -2 means that it has one concave pattern. If the absolute value of the frequency for a subsequence is larger than 4, we know that the subsequence has at least 4 consecutive up-down trends with the specified amplitude within the specified window size. This rule is often used for detecting vibration over specified amplitude.

Figure 3 shows the leg sequences in the same subsequence for several different amplitudes. The amplitudes are 1 for the top, 2 for the middle, and 4 for the bottom, respectively. The frequencies are 6 for the top, - 5 for the middle and 2 for the bottom respectively.

A naïve algorithm derived directly from the definition of the leg variation number is too slow for the algorithm to be applied to real applications. Because its computational order is O(n(W!), due to the combinatorial explosion resulting when selecting leg variation sequences. However, we can obtain a fast algorithm whose computational order is O(nw) by using the below leftmost leg sequence, and the below theorem that shows that a leftmost leg sequence has the longest length in a given leg variation sequence set.

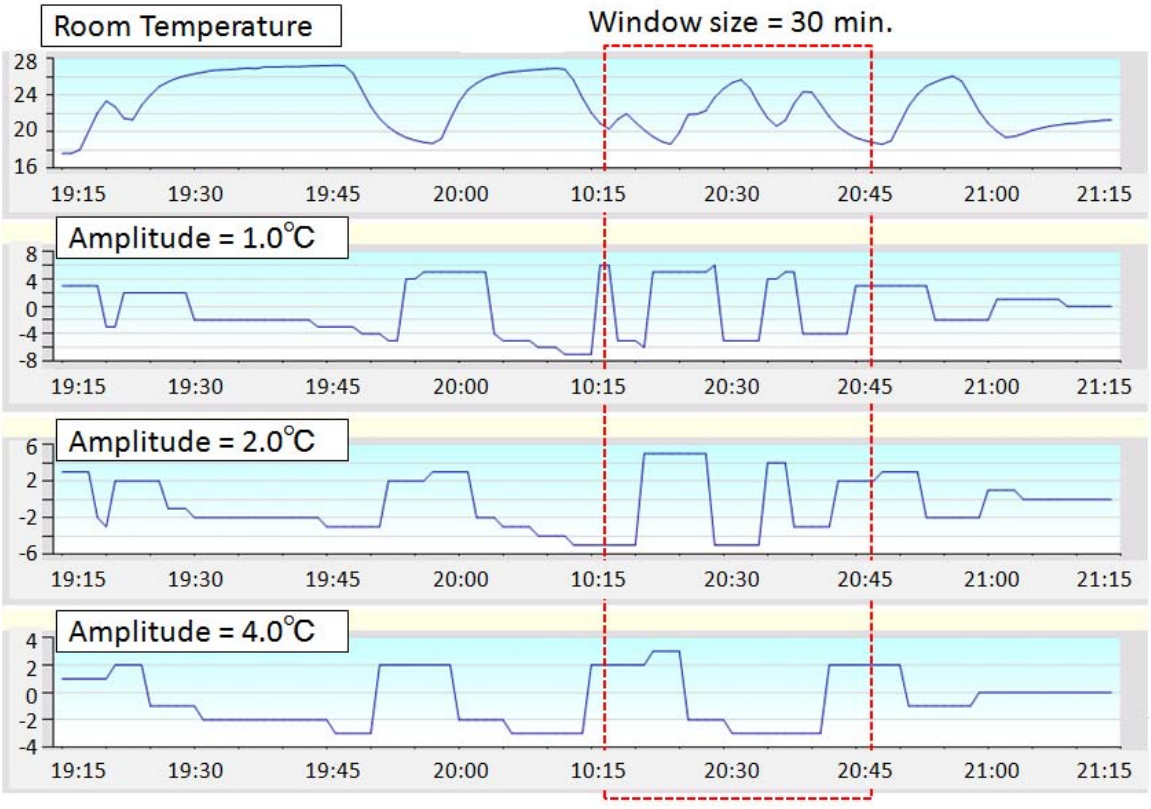


Figure 4: Trend Graphs of Experimental Data

Definition: Leftmost Vibration Sequence
The *leftmost vibration sequence* in S(X,A,W,t) is a leg sequence $[X_1,.., X_i,\ldots, X_n]$ that satisfies the following conditions:

Let Xi be defined by the following procedures:

if i = 1 $X_1 = \text{argmin}_{X \in S(X,A,W,t)}\, \text{end}(X)$

if i $\neq 1$ $X_i = \text{argmin}_{l \in Li}\, \text{end}(X)$

such that

$Li = \{ X \in S(X,A,W,t)\} \mid$

$\text{start}(X) \geqq \text{end}(X_{i-1})$ and

$\text{sign}(X) \times \text{sign}(X_{i-1}) < 0 \}$

Theorem: The leftmost sequence is a leg sequence that has the maximal length in S(X,A,W,t).

(2) Evaluation

We applied leg analysis to anomaly detection for HVAC (Heating Ventilation, and Air Conditioning) systems. It is known that a significant variation of room temperature during operation often shows an anomaly in the control and/or sensor system.

The data for the experiment consisted of room temperature readings with sampling period one minute, for a total duration of three years. The total number of time points is thus 1,578,239. We obtain leg frequencies with window size 30 minutes and with amplitude 1.0℃, 2.0℃, and 4.0℃, respectively. A higher amplitude means a higher warning level. Leg analysis can enable adaptive monitoring by selecting an appropriate window size and amplitude.

Our leg analysis software detected 1901 points (0.12%), 454points (0.029%), and 69 points (0.0044%) for amplitudes of 1.0℃, 2.0℃, and 4.0℃, respectively. Figure 4 shows a snapshot of leg frequencies as a function of time for each amplitude. The top, the middle, and the bottom graphs correspond to amplitudes of 1.0℃, 2.0℃, and 4.0℃, respectively. Fig. 3 above shows the leg sequences for the subsequences that are surrounded by the rectangular area in Figure 4.

The processing times for amplitudes 1.0℃, 2.0℃, and 4.0℃ were 0.612 sec., 0.554sec., and 0.489 sec., respectively. We set the threshold on the absolute value of leg frequency to 4, in order to detect anomalies for each temperature. All computational times satisfy the requirements of our application.

2.2 Satistical and Smoothed Trajectory (SST) Features

We proposed statistical and smoothed trajectory (SST) features [5] that can capture the shape and the stochastic behavior of the time series within the window so that it can handle various types of sensor data.

To detect anomalies in a time series, we first learn a model of the time series given normal time series data. To learn a model we use a fixed-size sliding window over the training time series and compute a feature vector representing each window. Our model consists of a set of exemplars representing the variety of feature vectors that exist over all windows in the training time series. The feature vector that is computed for each window consists of a trajectory component that captures the shape of the time series within the window and a statistical component that captures the stochastic component. The trajectory component is designed to capture the low frequency information in the time series window. It is computed using a simple fixed window running average of the raw time series to yield a smoothed time series after subtracting the mean of the window. Because of smoothing, half of the values in the smoothed time series can be discarded without losing important information. Thus, the trajectory component has w/2 elements where w is the number of time steps in the window. Figure 5a) shows a noisy sine wave time series and the corresponding smoothed time series with half the values discarded is shown in Figure 5b). The statistical component is a small set of statistics computed over time series values in the window which are mainly designed to characterize the high frequency information in the raw time series window. There are many possible choices for a set of statistics that well characterize the high frequency information in a time series window. The statistics used for experiments in this paper are mean, standard deviation, mean of the absolute difference ($|z(t) - z(t+1)|$), number of mean crossings divided by window length, percentage of

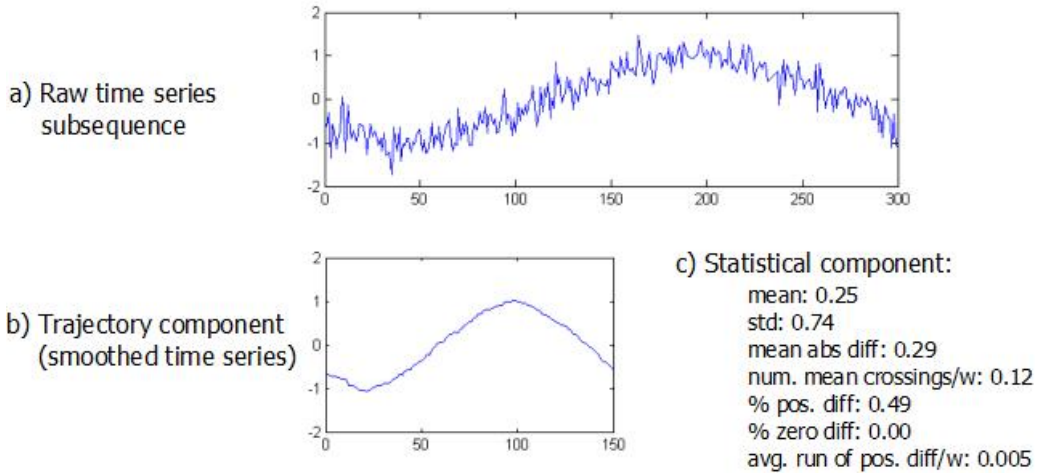


Figure 5: Example time series window (a) along with its trajectory (b) and statistical (c) components

positive differences, percentage of zero differences, and the average length of a run of positive differences divided by window length. Here, $z(t)$ is the value of the raw time series at time t. Figure 2c shows the vector of statistics for an example window. This choice of statistics has worked well in practice across a variety of different time series, but as mentioned before other statistics would likely also work well. The trajectory component is half the length of the window (w/2 time steps), and the statistical component is 7 real numbers for a total of w/2+7 real values. We call this novel representation Statistical and Smoothed Trajectory (SST) features.

3 ANOMALY DETECTION

3.1 Trend Pattern Query Language

In facility management for plants and buildings, the need for facility diagnosis to save energy or facility management cost by analyzing time series data from sensors of equipments in facilities has been increasing. We have proposed a relation-based stream query language TPQL (Trend Pattern Query Language) [6] [7] [8] to express constraints in time series data for anomaly detection in facilities. We implemented an anomaly detection system based on TPQL with Java and evaluated the expression ability to describe the anomaly condition and whether its processing speed is adequate for real applications. The features of TPQL are the following. (1) TPQL introduces a convolution operator into SQL (Structured Query Language) to describe contextual anomaly conditions over window sequences. (2) TPQL introduces time-interval based join into SQL to allow join operation on time series data with different sampling rates.

(1) Convolution Operator

A convolution operator is an operator over window sequences. Window functions in a convolution operator correspond to aggregate functions over window sequences in existing query languages with sliding window operations such as CQL (Continuous Query Language) [9]. So, we can easily describe anomaly conditions with SQL-like language by translating anomaly conditions described with convolution operator in control theory or signal processing to constraints in SQL-like language. In other words, this correspondence suggests the kind of aggregate functions that are useful for anomaly detection in a facility.

We define built-in window functions that describe anomaly conditions based on the requirement analysis of anomaly detection in a facility. TPQL has the following built-in aggregate functions:

1. Equality or inequality constraint with universally or existentially quantified operators.

This constraint is used to describe a query such as “Find the time t when the value f(t) is Y degrees higher than a signal g during X minutes”;

2. Residual function, the value of which is the difference between the real value and the value estimated by an auto-regressive model.

This constraint is used to describe a query such as “Find the time when the difference between the real value and the estimated value is larger than C”.

3. Hunting constraint over window sequences with “wave_count” operator

This constraint is used to describe a query such as “Find the time when the number of waves whose amplitude of signal f(t) is above Y degrees during X minutes window ”. Hunting constraint can be described by leg analysis in section 2.1.

(2) Time-interval based join

We define a time-series data table whose key is a column

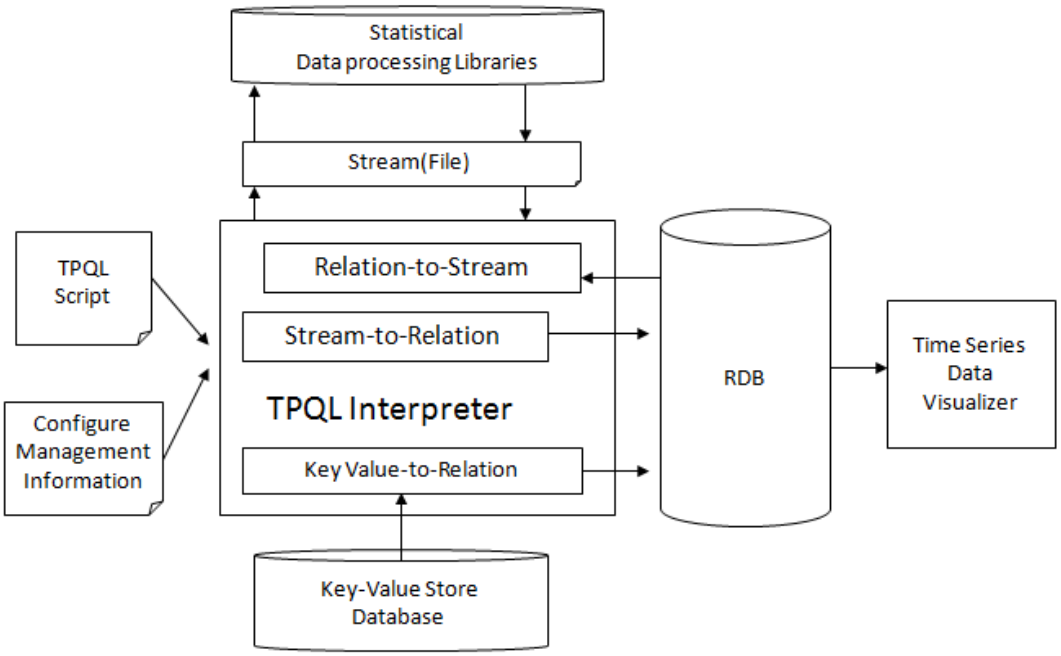


Fig. 6. Anomaly Detection System based on TPQL

pair “time and time interval”. The column “time interval” means a sampling period. The columns of the time-series data table can be interpreted as functions on time intervals. We use a construction method of a step function on a subdivision, used in the definition of the Stieltjes integral, to join tables with different time intervals.

Existing standard temporal query language TSQL [10] supports operation over time intervals, such as intersection and inclusion and so on, but does not support time-interval join. In TPQL, the mathematical expression over functions on time is important, so time-interval based joins are newly introduced as the basic operation which enables arithmetic operations over functions with different sampling periods.

(3) Anomaly detection system based on TPQL

We implemented an anomaly detection system based on TPQL, the structure of which is shown in Fig. 6. We confirmed that the queries from our requirement analysis on anomaly detection in facilities can be realized by the implemented system.

TPQL interpreter executes TPQL script and calls MySQL as an SQL engine and calls statistical data processing libraries which are implemented in R language or MATLAB. TPQL script can describe anomaly conditions with configure management information that describe domain dependent knowledge. We use KeyValue store database and introduce KeyValue-to-Relation transformation in addition to Stream-to-Relation and Relation-to-Stream transformations introduced by CQL.

3.2 Exemplar Learning

One possible model for a time series is simply the set of all SST features that are computed from all overlapping windows of the time series. This model would be an inefficient representation because the overlapping windows would produce many very similar feature vectors. A much more efficient model is created by finding a small set of exemplars that compactly represent the set of all SST features from the time series. An exemplar in this context is a representation of the SST features of a group of similar windows (overlapping or not) from the training time series. We use an agglomerative clustering algorithm to select SST exemplars from the set of all SST features for a time series.

The agglomerative clustering algorithm works as follows. After computing SST features for every window of the training time series, a set of exemplars is learned by initially assigning each SST feature as its own exemplar and then iteratively combining the two nearest exemplars until the minimum distance between nearest exemplars is above a threshold. This is illustrated in Figure 7.

We use Euclidean distance to measure the distance between two exemplars:

$$dist(f_1,f_2) = \sum_{i=1}^{\frac{w}{2}}\left(f_1.t(i) - f_2.t(i)\right)^2 + \frac{w}{14}\sum_{i=1}^{7}\left(f_1.s(i) - f_2.s(i)\right)^2 \quad \text{(eq. 1)}$$

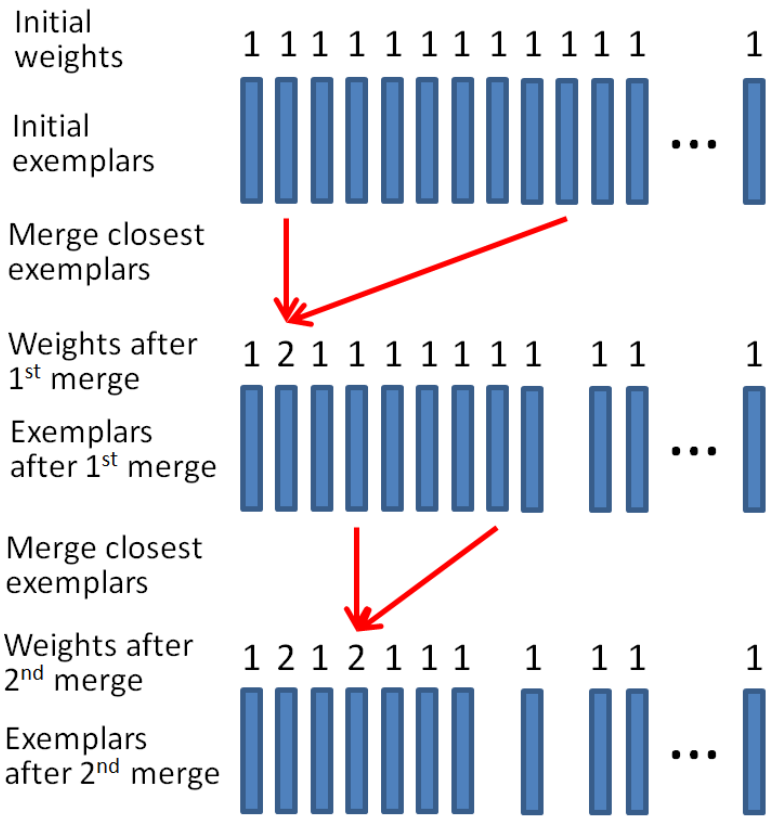


Figure 7: Illustration of agglomerative clustering for learning exemplars. The exemplars (which are SST feature vectors) are represented by blue rectangles. At each iteration the exemplars with minimum distance between them are averaged together using a weighted average. This process is repeated until the minimum distance is above a threshold.

where f_1 and f_2 are two feature vectors, $f_j.t$ is the length w/2 trajectory component of f_j, and $f_j.s$ is the length 7 statistical component of f_j. The w/14 coefficient causes the statistical and trajectory components to be weighted equally.

Two exemplars are combined by a weighted average of the corresponding elements. The weight is the count of the number of feature vectors that have already been averaged into each exemplar divided by the total count. Each resulting exemplar is thus simply the overall average of the feature vectors that went into it. The threshold that determines when to stop combining exemplars is set to $\mu + 3\sigma$ where μ is the mean of the Euclidean distances ($dist(f_1,f_2)$) between each initial SST feature vector and its nearest neighbor among the initial SST feature vectors and σ is the sample standard deviation of these distances. The running time of this exemplar selection algorithm is $O(n^2w)$ (where n is the length of the training time series and w is the chosen window size).

After exemplar selection, each exemplar is associated with a set of original SST features that were averaged together to form the exemplar. The standard deviation of each element of the w/2+7 length feature vector is then computed and stored with each exemplar. These standard deviations are computed over the set of SST feature vectors associated with a particular exemplar. An exemplar is thus represented by w/2 + 7 mean elements and w/2 + 7 standard deviation elements. In our experiments, the final exemplar set is typically between 1% and 5% of the total number of features (windows).

After the model is learned, anomalies are found in a testing time series as follows. For each window of the testing time series, an anomaly score is computed. This is done by first computing the SST feature of the window. Then the nearest neighbor exemplar to the SST feature is found. The distance function used is

$$d(f,e) = \sum_{i=1}^{\frac{w}{2}} \max\left(0, \frac{|f.t(i)-e.t(i)|}{e.\sigma(i)} - 3\right) + \frac{w}{14}\sum_{i=1}^{7} \max\left(0, \frac{|f.s(i)-e.s(i)|}{e.\varepsilon(i)} - 3\right) \quad \text{(eq. 2)}$$

where f is the SST feature vector for the current window consisting of a trajectory vector, $f.t$ and a statistical vector $f.s$, e is an exemplar for the current dimension consisting of trajectory ($e.t$) and statistical ($e.s$) vectors as well as the corresponding standard deviation vectors, $e.\sigma$ for the trajectory component and $e.\varepsilon$ for the statistical component.

This distance corresponds to assigning 0 distance for each element of the trajectory or statistical component that is less than 3 standard deviations from the mean and otherwise assigning the absolute value of the difference divided by the standard deviation for each element that is more than 3 standard deviations from the mean. In equation 2 and in our experiments, the statistical component is given equal weighting to the trajectory component, although this weighting can be changed based on the application.

4 FAULT KNOWLEDGE BASE

Following growing public interest in product recalls as a social issue, improving the product quality is becoming a very important concern for industries. We have developed and operated a design defect prevention system [11] based on SSM (Stress-Strength-Model)[12], which is a method to structuralize design knowledge using a chain of causes and their effects. How to let busy designers build and utilize structured knowledge is a very important issue in operating a structured knowledge system. The design defect prevention system has been in operation since 2006 for controlling circuit design and structural design in developing air-conditioning equipment. The system achieved a 59 percent reduction of the number of design changes noticed, and additional 12 percent of check items that manual design

definition attribute	controlling attribute	strength	stress	failure mode
Negative terminal post			Liquid spill	leakage current
			Causal Chain	
Electrolytic capacitor	No oxide film on negative terminal post	Low tolerance for alkali inside a condenser	Occurrence of strong alkali electrolytic solution	liquid spill
			Causal Chain	
Electrolytic capacitor	Electrolytic solution is quaternary salt		Passage of electric current	Occurrence of strong alkali electrolytic solution

Fig. 8. An example of a causal chain of SSMs

review could not point out. Furthermore, we confirmed the improvement of the design ability to prevent design defects through a competence-based questionnaire for ten designers.

A fault knowledge base comprises three kinds of knowledge: a dictionary that describes class knowledge about is-a relation and has-part relation of parts and the attributes of parts; configuration information that describes instance knowledge about the structure of a specific machinery or system; and SSM knowledge about the cause of failure.

SSM comprises 5 items: a definition attribute that means a part or region where a fault occurs, a strength that denotes the tolerance that the definition attribute should have for preventing a failure; a controlling attribute that denotes a design parameter that decides the strength of the definition attribute; a stress that denotes a condition or an input that causes a failure; a failure mode that denotes the manner in which the failure of the definition attribute occurs. Figure 8 shows an example of a causal chain of SSMs.

5 CONCLUSIONS

This paper describes our framework for equipment condition monitoring, whose purpose is to address the growing needs for preventive maintenance enabled by the rapid spread of the Internet of Things.

Our development efforts have been mainly focused on statistical feature generation and machine learning based anomaly detection, that is, a data-driven approach. An important problem of a data-driven approach is that it can only detect the differences from the ordinary behavior of sensor data, but it cannot distinguish the symptoms of failures from only unusual behavior. Our future direction is to develop a hybrid method to combine a data-driven approach with fault training data and a model-based approach with domain knowledge of equipment in order to detect the symptoms of failures from sensor data.

REFERENCES

[1] J. Zheng, D. Simplot-Ryl, C. Bisdikian, H.T.Mouftah: “The Internet of Things [Guest Editorial] ”, Communications Magazine, IEEE , vol.49, no.11, pp.30-31 (2011).

[2] M. Imamura, D. Nikovski, Z. Sahinoglu, M. Jones: A Survey on Machine Learning for Equipment Condition Monitoring Using Sensor Big Data, IIEEJ Transactions on Image Electronics and Visual Computing Vol.2 No.2, pp. 112-121 (2014).

[3] E. Fin, B. P. Kevin: Indexing of Compressed Time series, DATA MINING IN TIME SERIES DATABASES, World Scientific, pp. 43-65 (2004).

[4] M. Imamura, T. Nakamura, H. Shibata, N. Hirai, S. Kitagami, T. Munaka: Leg Vibration Analysis for Time Series (in submitting)

[5] M. Jones and D. Nikovski and M. Imamura and T. Hirata: “Anomaly Detection in Real-Valued Multidimensional Time Series”, Proceedings of the 2nd International ASE Conference on Big Data Science and Computing (2014).

[6] Makoto Imamura, Shigenobu Takayama, and Tatsuji Munaka: A stream query language TPQL for anomaly detection in facility management. In Proceedings of the 16th International Database Engineering & Applications Sysmposium (IDEAS '12). ACM, New York, NY, USA, 235-238. (2012).

[7] M. Imamura,T. Takeuchi, S. Kitagami,M. Kanno, T. Munaka: Time Series Data Query Language TPQL for anomaly detection in facility, Journal C of Electronics and Communications in Japan,Vol.134, No.1, pp. 156-167 (2014). (in Japanese)

[8] M. Imamura,T. Nakamura: Window Size dependency on Anomaly Detection Based on Discord for Time Series, Journal C of Electronics and Communications in Japan, Vol. 135, No. 10, (2015). (in Japanese)(in printing)

[9] Arvind Arasu, Shivnath Babu, Jennifer Widom: The CQL continuous query language: semantic foundations and query execution, The International Journal on Very Large Data Bases archive, vol. 15, no. 2, (2006).

[10] Richard T. Snodgrass (Ed.): The TSQL2 Temporal Query Language. Kluwer (1995).

[11] T. Kashima, H. Kimura, H. Koizumi, M. Imamura: Operation and Evaluation for Design Defect Prevention System Based on Structured Knowledge, Journal of the Japanese Society for Artificial Intelligence, Vol.26, No.5, pp. 607-620. (2011). (in Japanese)

[12] Y. Tamura, Y. Iizuka: A Study on the Method to Manage Design Knowledge on Failures : Construction of the Knowledge Structure of a Causal Chain of Failures, Journal of the Japanese Society for Quality Control, Vol. 32.1, pp. 122-135 (2002). (in Japanese)

Session 3:
Intelligent Transportation Systems
(Chair : Yoshitaka Nakamura)

Evaluation Platform for Driver-Distraction Problem Using On-vehicle Information Devices

Yutaka Onuma*, Suguru Nakazawa* , Daishi Shimizu* , Ryozo Kiyohara*

*Dept. of Information and Computer Science, Kanagawa Institute of Technology, Japan
{{s1221026, s1221066, s1221054}@ccy, kiyohara@ic}.kanagawa-it.ac.jp

Abstract - Recently, automotive navigation systems and smartphone navigation applications have become increasingly popular. However, there are many driver distraction (DD) problems associated with such systems that may result in serious traffic accidents. Therefore, it is essential to evaluate the user interfaces of such types of systems from a display audio (DA) perspective. Driving simulators (DSs) are considered to be suitable evaluation platforms for navigation devices. However, simple DSs in which participants cannot perceive acceleration are inadequate. Moreover, it is necessary to naturally simulate the approach of other vehicles inform the participants' perspective. In this paper, we propose the integration of a DS and a traffic simulator as the evaluation platform for on-vehicle information devices. Further, we propose a method to share the map information between two simulators and discuss integration issues.

Keywords: driving simulator; traffic simulator; car navigation system; evaluation platform

1 INTRODUCTION

On-vehicle information systems such as car navigation devices are increasingly common, and many drivers actively focus on the device screens and operate the touch screen interfaces. Moreover, instrument panels that have LCDs are expected to become commercially available. Therefore, there will be a lot of information in close proximity to the drivers. Moreover, there are also head-up-displays (HUDs), rearview monitors, etc.

However, these types of car navigation systems and instrument panels may cause driver distraction (DD) problems. This is mainly because of the amount of information that is difficult to readily comprehend.

There are actually many traffic accidents that have been caused by DD. Therefore, the design and evaluation of these devices' user interfaces are very important.

In order to evaluate DD problems, driving simulators (DS) are widely considered to be indispensable platforms. While user interfaces should be evaluated by real vehicles on actual roads, this approach is dangerous and the evaluations should not be performed in the earlier stages.

Therefore, interfaces should be evaluated on a DS. Figure 1 shows a picture of a typical DS. However, the use of evaluations that involve DSs in which participants cannot perceive realistic acceleration scenarios results in ineffective experiences for drivers.

On the other hand, in many cases, vehicles that are used with DSs can only simulate fixed scenarios. Therefore, the behavior of surrounding vehicles in a DS is expected to be different from the behavior of actual vehicles. As a result, it is difficult to precisely evaluate DD using a DS. In order to solve this problem, we introduce a network-type DS, as shown in Figure 1. In this system, there is a primary DS and several secondary DS units, which can be independently operated by participants [1].

In this system, there are no fixed scenarios because each vehicle can be independently operated by its participant. Therefore, in the main DS, real world experiences can be realized for participants. Moreover, a sufficiently accurate evaluation can be performed in this environment. If the number of participants in a network-type DS increases, it is expected that the accuracy also increases. [1]

In this paper, we propose a new platform that integrates a DS and a multi-agent-type traffic simulator (TS). In the TS, one agent is controlled by one participant; in the DS, the vehicles around the main DS are multi-agents in the TS, and several vehicles are operated by other participants.

2 ON-VEHICLE INFORMATION DEVICES

There are many kinds of on-vehicle information devices, e.g., car navigation systems and different kinds of monitors

Figure 1: Driving simulator

Table 1 Position and operation

Position \ Operation	Operation	No operation
Center under dashboard	Navigation, music player	
Center on the dashboard	Navigation, music player	Velocity meter, Odometer
Center on the rearview mirror		Rearview monitor, Turn by turn navigation
Front of the driver on dashboard	Odometer	Velocity meter, Turn by turn monitor
Head-up display in front of the driver		Real monitor navigation

(velocity monitors, odometers, energy monitors, and rearview monitor).

Moreover, many displays are positioned in the center of the dashboard, in front of the driver, and on the rearview mirror, etc. Table 1 shows the various types of displays.

It is believed that there is an excessive number of sources of information that may cause DD problems. A lot of manufacturers have therefore developed new products that are aimed at reducing DD problems. However, it is not safe to experiment with these new products on actual roads. Therefore, there is a need for a DS that is used as an evaluation platform for user interfaces.

3 NETWORK TYPE DS

Figure 1 shows an image of a typical DS, in which the participant can observe all around the vehicle using monitors. In many cases, persons may have had some experience at a driving school or driver's license center, where the driver operates the steering, acceleration, and braking pedals.

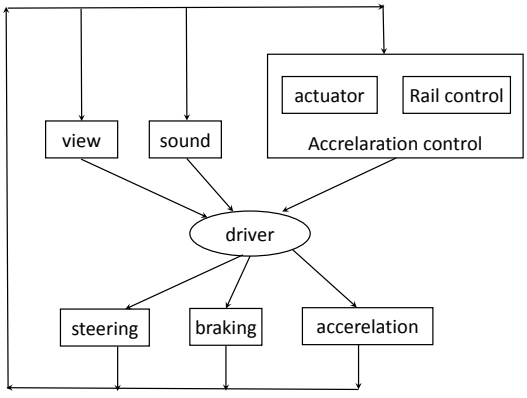


Figure 2: System structure of typical DS

The typical system structure is shown in Figure 2. Depending on the situation when the participant watches the screens, hears sounds, and feels the acceleration, he/she operates the steering, acceleration pedal, or braking pedal. These sequences are repeated in 10—30 ms intervals, and are implemented on a typical DS.

However, this type of DS results in fixed-scenario problems. Actually, the behavior of the vehicle on the opposite side or of other vehicles in the same lane should vary according to the behavior of the vehicle, which is driven by a participant. However, this is difficult to simulate in a DS.

Therefore, network-type DSs have been developed. In a network-type DS, there is a main simulator, which is used for evaluation, and sub-simulators, which affect the main simulator through dangerous driving or driving that follows the main driver. The simulator share common maps, signals, and other information. Moreover, during each simulation period, the simulators exchange information regarding location, direction, and velocity, and there are displays of other vehicles on each vehicle's screen. Natural images last approximately 30 ms in each simulation period. These items of information are logged for analysis in the event of accidents or unexpected driver operation [2].

Figure 3 shows our proposed driving simulator [2]. The DS on the left side of Figure 3 is the main simulator that is

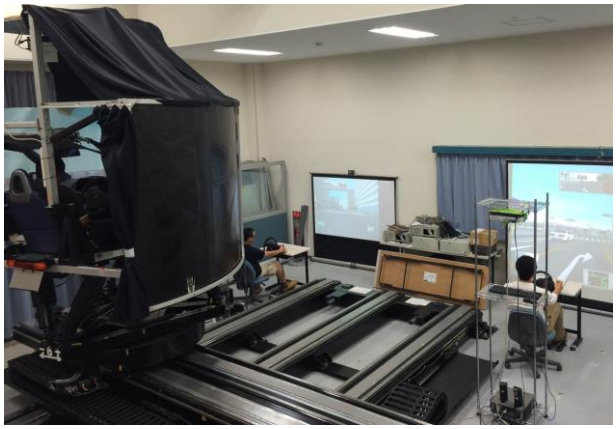


Figure 3: Network type DS.

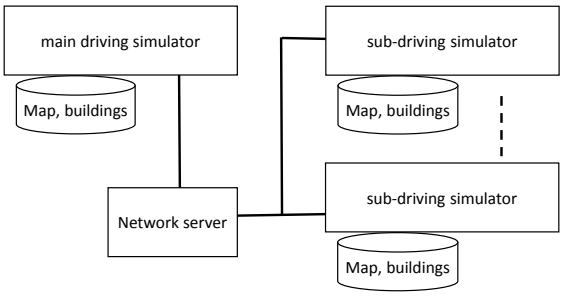


Figure 4: Structure of network type DS

driven by the participant. The screens on the center and right side of Figure 3 are sub-simulators that are connected by a UDP/IP network.

Figure 4 shows the structure of a network-type DS. Currently, we have a main simulator and five sub-driving simulators, which are connected by a UDP/IP network.

If the number of participants in a network-type DS increases, it is known that the effect also increases.

On the other hand, it is difficult to prepare several DSs, which are very expensive, and to gather many participants. Therefore, we believe that an equivalent effect may be realized by substituting a multi-agent-type TS for vehicles such as a sub-DS.

4 COMBINED DS

In order to communicate by connecting a TS with a DS, as shown in Figure 5, each vehicle (agent) state is treated separately in a network-type simulator.

This integration imitates a network-type simulator as though many participants are operating many vehicles. Two simulators exchange information about the location, direction, and velocity as the vehicle status, and they also exchange signal status. The map information is considered to be common data.

However, there are several problems when combining the two types of simulators. These are as follows:

1) Map formats
2) Simulation interval time
3) Multi-agent models for driver and vehicle behavior

In this paper, we focus on the map-related problems. DSs and TSs require different sets of information for the MAP. The DS requires an image of the road, while the TS requires information related to the load length and previous vehicle information. Then, we proposed a map-conversion method to solve this problem.

5 RELATED WORKS

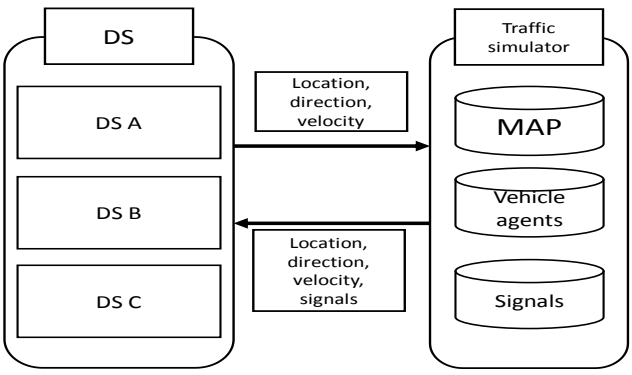


Figure 5: Proposed system

There have been related works for DS. In [3], the author reports a support system for traffic combined points of highway. In this study, the user interface of the system was evaluated on a simple DS. This evaluation method provides good hints.

Reference [4] presents a TS on a multi-agent model. In this study, the accident is simulated. In many multi-agent models, there are no accidents because those studies evaluate mainly the traffic flow. Therefore, this research may provide helpful results.

References [5] and [6] are studies that involve a combined TS.

The study in [5] is based on a macro TS. However, near to the target vehicle, the system is simulated by a multi-agent model. This research is used to obtain realistic simulation of a DS. The study in [6] is also based on a multi-agent model. This research is for the heart-related problems. These two studies have provided useful information in this regard.

However, these studies did not involve network-type simulators, and these studies did not consider the map problems. This may have been because from the beginning of the study, both maps were designed concurrently. However, the cost of constructing the map is very important, and we therefore focus on the map-conversion method.

6 MAP FORMAT

Figure 6 shows a map of the Suzuka circuit. This is a simple example because there are no junctions or intersections. From the corner of this map, Figure 7 are shown from the viewer. The map information contains mainly polygon data. The map in our driving simulator is in X-file format, which is supported by direct X. These types of files consist of a lot of polygon data and vertex data,

Table 2: Information of X file and shape file

	X file	Shape file
Presentation of roads	polygon	line
Road information	texture	Road width Number of lanes Road constraint
File format	Binary and Text	Binary

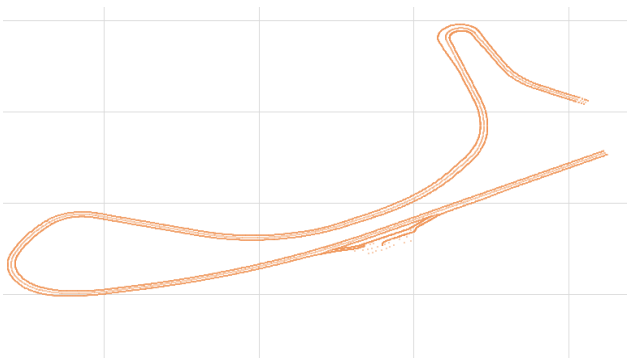


Figure 6: Map of Suzuka circuit

Figure 7: Example view of DS on the map of Figure

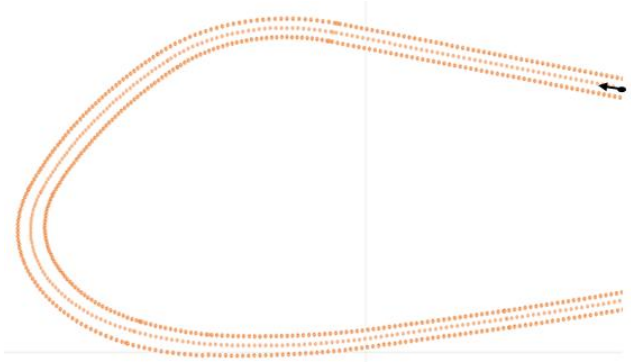


Figure 8: Polygons

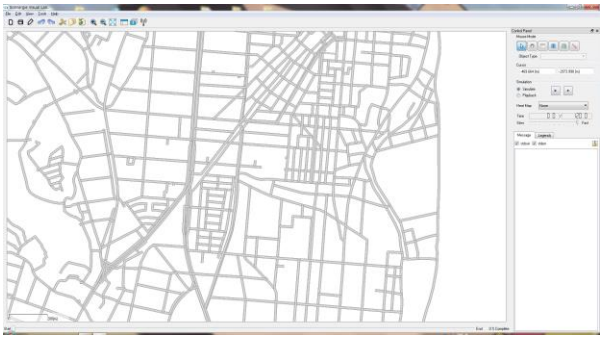


Figure 9: Map of traffic simulator

which are shapes that are presented in triangles. Figure 8 shows each vertex and the center point of each polygon.

On the other hand, Figure 9 shows the map of a TS. In many cases, the map of the TS is in vector format. Therefore, the map consists of multiple edges and nodes. Our proposed TS is based on the multi-agent model, which has a shape file, as shown in Table 2.

We believe that if the center points are connected, we can convert the continuous center points to edges. Table 3 shows the required information for each map.

Table 3: Required information for map

Driving simulator	Polygon, Texture mapping
Traffic simulator	Road width Number of lanes Road constraint

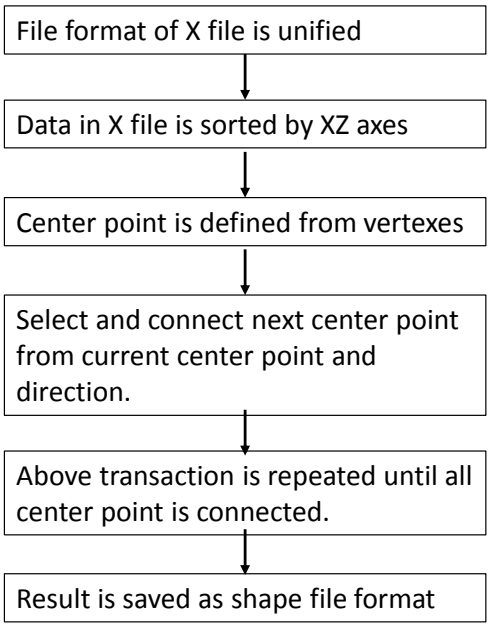


Figure 10: Proposed method

7 PROPOSED METHOD

Our proposed method is shown in Figure 10. First, the X file data should be consistent. There are binary and text files in the original X file format. Secondly, the data file should be sorted by the XZ axes. Continuous polygons along the road are created by this sorting. Third, in each polygon, the center point should be calculated. Next, one center point is chosen, and the nearest center point is found. After these two points are connected, the direction is defined. Then, the next candidate center point is found from the same direction, as shown in Figure 11. These operations are repeated until all of the center points are connected. Finally, the sets of nodes and edges (see Figure 12) obtained from the results of these operations are saved as shape files.

However, this conversion method does not account for all patterns. Therefore, after the conversion, the user has to check the map and make appropriate revisions. This

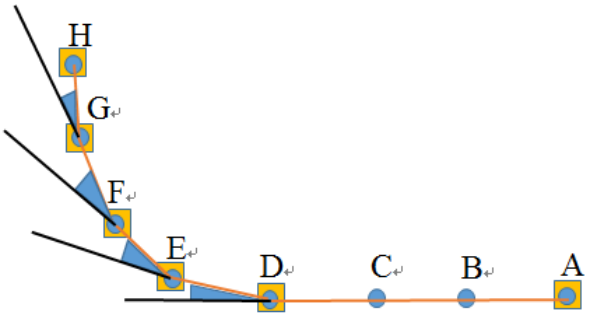


Figure 11: Proposed Method 2

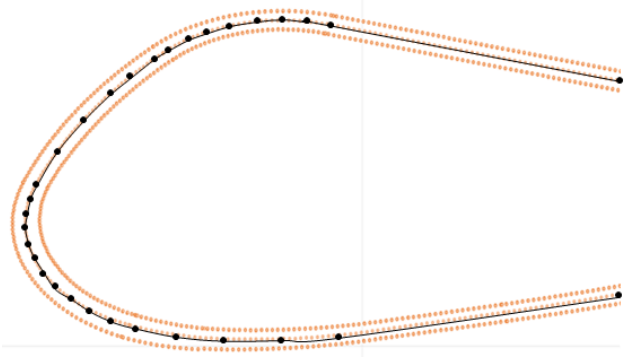


Figure 12: Conversion result

operation is not very difficult in the experiment environment. The following patterns are required for these operations.

(1) On a long and straight road, the nearest center points are very far, and errors may be made (see Figure 13).
(2) Junctions are complex cases, and errors may be made (see Figure 14).
(3) Intersections are also complex cases, and errors are possible (see Figure 15).

8 DISCUSSION

On a long, straight road, priority should be given to the search process. The presence of several vertices indicates a long and straight road, and there are not many junctions on the course. Therefore, user operation should be permitted. However, there are many intersections on the city map, as shown in Figure 15, so other processes may have to be developed.

9 CONCLUSION

We proposed a map-conversion method for a simple road. We showed that a map of the Suzuka circuit can be converted to a map for a traffic simulator. However, the presence of junctions or intersections would make it difficult for their automatic conversion. We therefore discussed these cases.

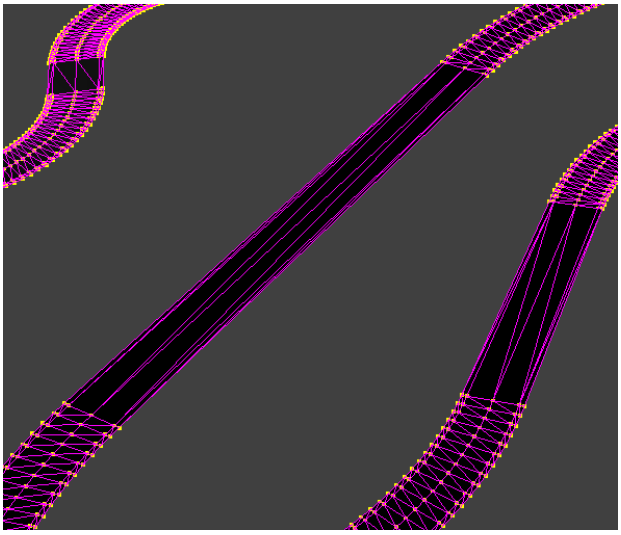


Figure 13: Example of long straight roads

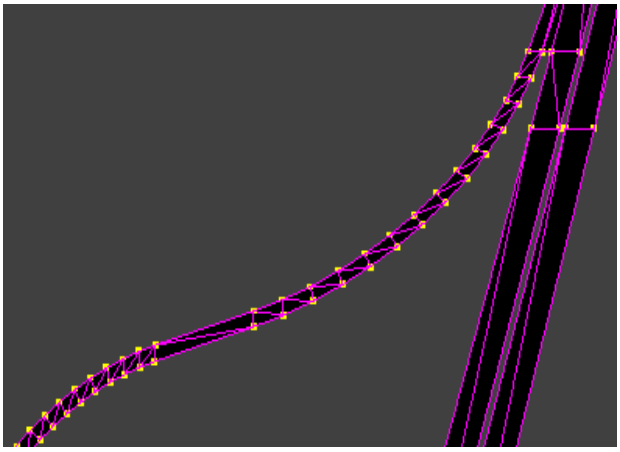


Figure 14: Example of junction

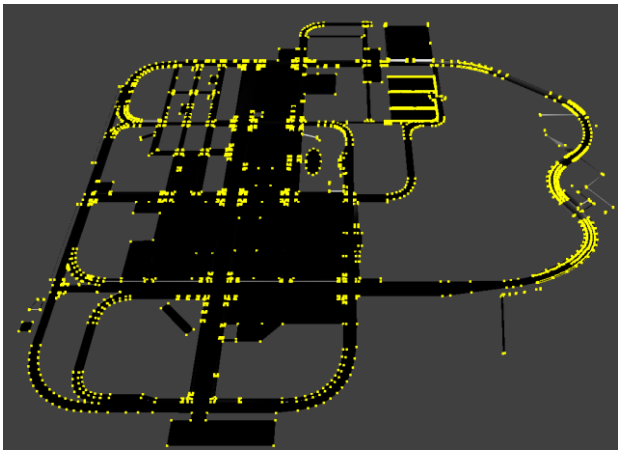


Figure 15: Example of city map

In our future work, we plan to develop a new process for intersections and junctions

REFERENCES

[1] Nakazawa S, Koichi M, Eto Y., Kano Y, Abe M,, " A study on effects of driver assists systems on reducing traffic accidents using three driving simulators connected with networks" Society of Automotive Engineers of Japan.2007..

[2] Kano Y., Ando T., Akiyama S., Abe M., "Networked Driving Simulators for Investigation into Interactive Behaviors among Driver-Vehicle-Systems in Road Traffic Accidents," FAST-Zero'11 Proceedings, TS3-8-1-3, p1～p6, (2011)

[3] Shimizu T., Ando T. "An Analysis of the Effect of Information System at Merging Section Using Driving Simulator," INFRASTRUCTURE PLANNING REVIEW Vol. 23, No.4, pp.833-840, 2006 (Japanese)

[4] Mizuno K., Yamada M., Fukui S., Nishihara S. "Multi-Agent Approach to Microscopic Urban Traffic Flow simulation," The journal of the Society for Art and Science, Vol. 5, No.2, pp.23-32, 2006 (Japnese)

[5] Honda K. "Evaluation of Driving Behavior using Virtual Reality Experiment," 12th World Congress on Intelligent Transport Systems proceedings 2005, 11(CD-ROM). (2005)

[6] Tajima J, Sakamoto N, Mochida T, Tanaka S, Yasuda S., "Verification of ASSTREET Driver-Agent Model by Collaborating with the Driving Simulator," SAE 2012 World Congress & Exhibition (2012)

Usability Improvement of RS-WYSIWYAS Navigation System

Yusuke Takatori *, Shohei Iwasaki **, Tatsuya Henmi **, Hideya Takeo*
Kanagawa Institute of Technology, Japan
*{takatori,takeo}@ele.kanagawa-it.ac.jp

Abstract -"RS-WYSIWYAS navigation" is a real-time and seamless intuitive navigation concept. A prototype application of RS-WYSIWYAS pedestrian navigation system for smart devices has been developed. In this application, a smart device captures the M-CubITS marker elements assigned by M-sequence on a corridor from a scene movie. Then the application obtains the position of the captured marker sequence in the building and heading of the smart device, and provides the user with the direction to a given destination in real time and seamlessly. However, in order for the prototype to become practical, usability improvements in the response and marker sequence recognition performance are required. In this paper, the application's response is improved by resizing the captured movie (sequential pictures) and using native code programming. At the same time, reliable positioning and heading performance is improved by pre-masking processing to the captured movie. Such processing omits the information of unnecessary areas where the M-CubITS marker elements do not exist, and extracts only the marker sequence information. Furthermore, an installation scheme to the RS-WYSIWYAS navigation system into multiple-layer buildings is presented.

Keywords: ITS, Indoor navigation system, WYSIWYAS navigation, M-CubITS

1 INTRODUCTION

Because of the spread of smart devices, the demand for pedestrian navigation systems has increased. Most smart devices use GPS for positioning. However, if these devices are used inside a building or underground areas that are inaccessible to the GPS signal, it can be difficult for such devices to position themselves. In addition, it is known that GPS precision and accuracy can decrease if smart devices are used in areas dominated by tall buildings [1]. For these reasons, pedestrian navigation systems not dependent on GPS have attracted research attention. In around 2000, many researches of the tag-based pedestrian navigation system that uses RFID or two dimension code are done as pedestrian navigation systems that does not rely on GPS[2]-[5]. Most of tag-based navigation system, tags are input their location information, then a mobile reader reads these information and recognizes its location, and the acquired information is pointed on a 2D map. However, the guidance is conducted only at the tag-installed place and uses 2D map. That is, the user looks for the ID-tag and understand his location on a 2D map, moreover he has to judge the way to go. Because the operation load on a user of the tag-based navigation systems are high, use of the system is difficult for the person who cannot read a map well. To realize more intuitive and easy navigation system, Hasegawa has proposed a navigation concept called "WYSIWYAS" (What you see is what you are suggested) [6]-[8] and a positioning system using M-sequence multi-modal markers called "M-CubITS" [6]-[8]. Moreover, Yamashita and Manabe developed navigation systems that recognize their position and heading by capturing images of M-sequence markers [9]-[11]. When the navigation application running on a cell phone with a camera device takes a picture which include M-CubITS elements arranged on the floor, it recognizes its position and heading based on the captured M-CubITS information, and displays the direction to the destination over the taken picture. In this manner, their WYSIWYAS navigation application systems achieve intuitive and virtual navigation without requiring the user to understand his or her own position on a 2D map.

Nevertheless, these systems urge users to stop and capture an image to determine the direction of the final destination. This could impede pedestrian traffic and lead to collisions with other pedestrians or vehicles. Therefore, the authors developed a WYSIWYAS indoor navigation system that provides real-time and seamless WYSIWYAS navigation as an android application [12].

In previous research, navigation applications captured a scenic movie and provided users with the direction to the final destination in real time and seamlessly. However, the fact that high processing load would cause navigation delays when using current smart devices with high-resolution cameras is concerning. Moreover, it has been difficult to distinguish non-marker objects of similar color to positioning markers placed on the floor. In such cases, navigation seamlessness deteriorates.

Furthermore, conventional applications have been evaluated only on a single floor in a building. It is important to introduce the RS-WYSIWYAS navigation system to multiple floors in a building.

In this paper, section 2 explains the RS-WYSIWYAS navigation system. In section 3, the application's response is improved by resizing the captured movie (sequential pictures) and using native code programming. In section 4, marker sequence recognition is improved by applying pre-masking processing to the captured movie. This omits the information of unnecessary areas where M-CubITS marker elements do not exist, and extracts marker sequence information. In section 5, introduction of the RS-WYSIWYAS navigation system to multiple floors in a building is described. In section 6, this paper is concluded.

2 REAL-TIME AND SEAMLESS WYSI-WYAS NAVIGATION SYSTEM [12]

2.1 WYSIWYAS navigation system concept

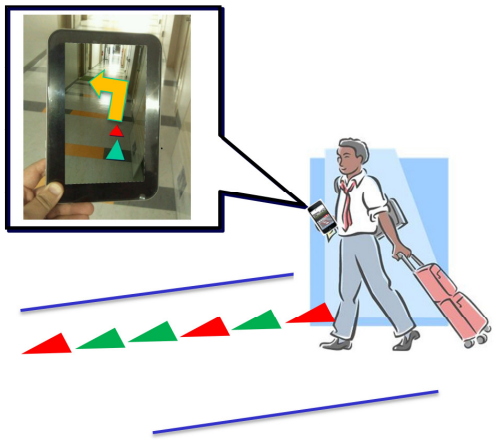
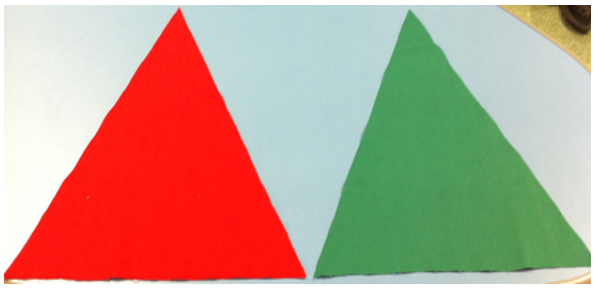


Figure 1 Example of WYSIWYAS navigation



Red: “1” Green: “0”

Figure 2 Marker elements for M-CubITS

WYSIWYAS is a type of Human Machine Interface (HMI) for virtual navigation systems [6]-[8]. When users employ smart devices with a WYSIWYAS navigation system, they capture images using the WYSIWYAS navigation application, and the application displays an arrow pointing to the final destination on the captured image. Figure 1 shows an example. If users navigate following the direction indicated by the arrow, they will arrive at their destination. That is, the smart device does not need to know the accurate and precise position information. To implement these features, the WYSIWYAS navigation system is required to recognize the relationship between the destination, location of the captured image, and orientation of the smart device.

2.2 M-CubITS

M-CubITS is a positioning scheme that places multimodal marker elements (as shown in Figure 2) according to M-sequences along passageways, detects a row of M-CubITS elements with a camera, compares the row with a database, and determines the position and direction of the captured marker elements. M-CubITS marker elements are designated as either “0” or “1” and distinguished by the color of the figure. Because an M-sequence code is generated from the Linear Feedback Shift Register (LFSR) of m stages with code length 2^{m-1}, the device can recognize its unique position by observing m-chips (Figure 3). For instance, if a passageway 1 km long is covered by markers that contain 1 bit of information placed 1 meter apart, the required number of shift register stages is ten. Therefore, marker positioning is required to capture at least 10 markers in an image. In this study, we adopt an M-sequence generated from a 7-bit LFSR, and the marker elements are

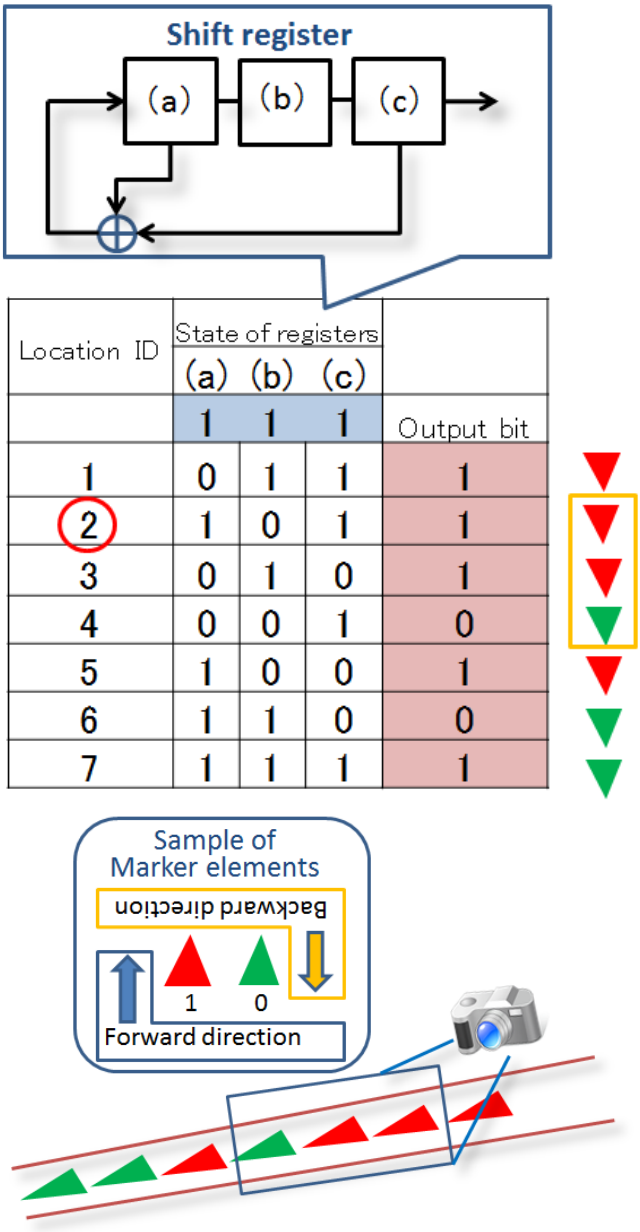


Figure 3 Example of marker sequence

arranged 0.5 meter apart. A passageway of approximately 125 m is covered, and a smart device recognizes the captured marker positions on the passageway by capturing an image of at least seven marker elements.

2.3 Overview of real time and seamless WYSIWYAS navigation

To indicate to users the direction to the final destination, the proposed system processes multiple images captured successively in a short period. The device recognizes the position of the captured marker sequence in the building and heading of the smart device, and displays a guiding arrow on the original landscape picture. This navigation application was developed on the Eclipse [13] with Android SDK [14]. Some image processes use the OpenCV for the Android library [15]. Although the process of marker recognition is based primarily on the scheme in [9], some of the process

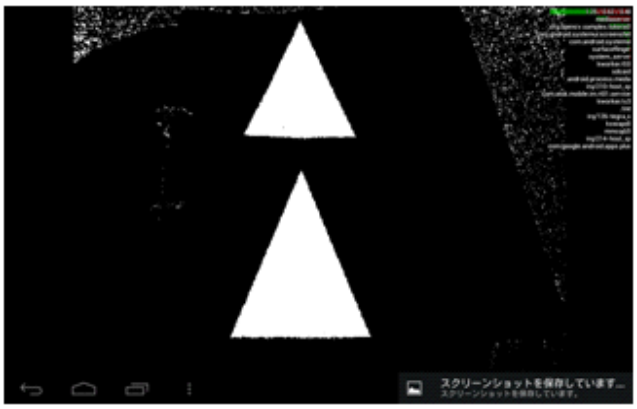


Figure 4 Particular color abstracted image

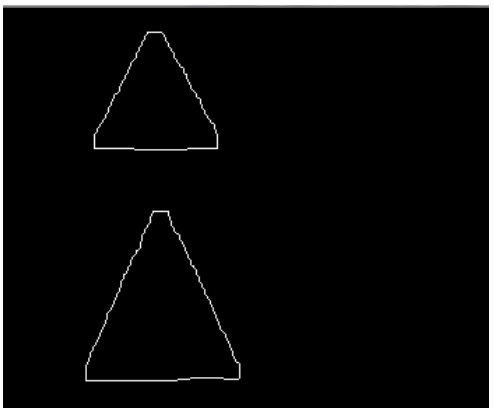


Figure 6 Contour extraction

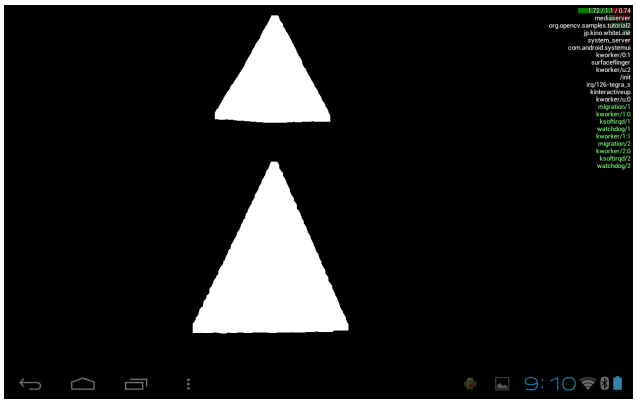


Figure 5 Noise reduction image

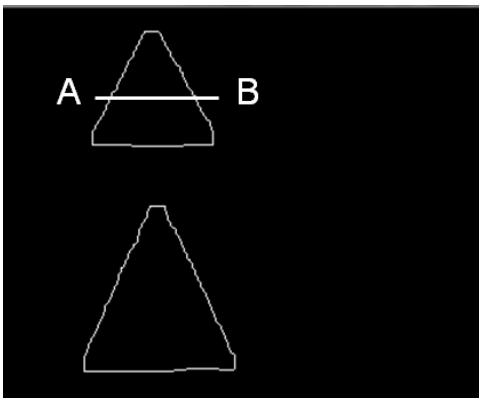


Figure 7 Recognition of marker element color

uses a different technique that considers processing time.

In this study, we adopted the M-CubITS marker elements shown in Figure 2. The elements are either red or green; this binary-information is used as chip information for the M-sequences. The shape of the elements is triangular to facilitate recognizing the direction of the M-sequence.

2.4 Flow of marker recognition process

The process for marker element recognition is as follows:

(1) Real-time image capturing by a rear camera

(2) Particular color abstraction

The marker-colored area is abstracted from the captured image and converted into an HSV (color model) image. In this process, the captured image becomes binary; red or green areas are designated as "1"; otherwise, they are designated as "0." The color-abstracted image is shown in Figure 4. To reduce the noise of the binary image, a smoothing and morphology operation is conducted. An example of a noise-reduced image is shown in Figure 5.

(3) Contour extraction

The contours of the marker elements are extracted and labeled with numbers. The label numbers are assigned in order from the bottom to the top of the image. During the marker element recognition process, described later, marker element information is acquired in a specific order, starting with the marker with the smallest label number. The contour image is shown in Figure 6.

(4) Marker element recognition

To extract information from a marker element, the centrobaric coordinate of the contour is calculated. Next, a horizontal line that passes through the centrobaric coordination is drawn, and two intersection coordinates with the contour line are calculated. A horizontal line with intersection coordinates (coordinates A and B) is drawn in Figure 7. Then, the most common pixel color between coordinates A and B is judges as the marker element color.

The direction of the extracted sequence of the marker elements is determined as follows:

- The application calculates the gradient of a line that passes through centrobaric coordinates on the first and second marker elements.
- It calculates two coordinates that intersect the horizontal line that passes through the centrobaric coordinate and contour line of the first marker element.
- By drawing two vertical lines from these intersection coordinates to the line that passes through the centrobaric coordinates, two line segments result.
- These line segments are compared to determine the direction of marker elements. These are shown in Figure 8. If the gradient of the line is a near right angle, the direction of the marker elements is determined by comparing the number of pixels within the contour line at $y = Y_g$ and $y = Y_g$-5 (pixel).

(5) Navigation arrow display

Using the marker color and the marker direction information, the device determines the sequence of marker elements. Then the system judges the direction to

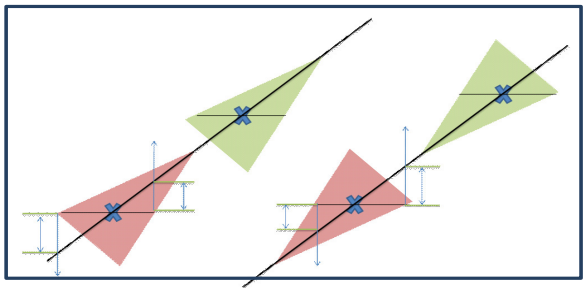


Figure 8 Determining marker direction

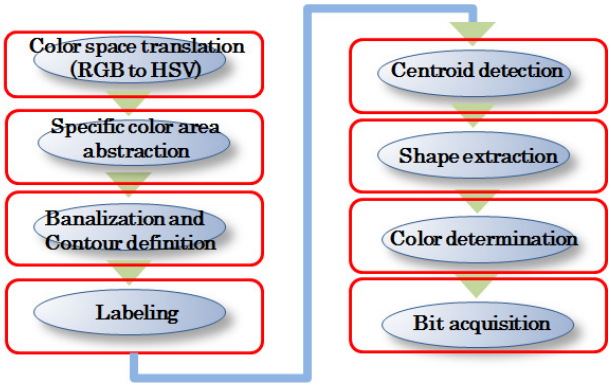


Figure 9 Flow of image processing

go, and overlay a guiding arrow symbol on the original scenic image. If the application cannot extract enough number of marker elements, the sequence of marker elements cannot be specified and arrow symbol is not updated in the processed frame.

2.5 M-CubITS marker database

Smart devices need to obtain their own location and direction that are referenced from a marker database. Therefore, a database and a method to access it need to be considered. With regard to this system, use in private areas is assumed. In private areas, a building manager might not want to offer marker data about the building to unauthorized individuals. Moreover, mobile phone use might be restricted within such private areas. In these cases, it is assumed that smart devices would access the M-CubITS database via the wireless local area network prepared by the manager.

3 RESPONSE IMPROVEMENT

3.1 Resizing captured images

There is a concern that the high processing load would cause navigation because the resolution of current smart devices has become larger, delays. To reduce the influence of the image processing load that depends on image size, the size of the captured image is reduced before the main process.

3.2 Implementation of native method

Because conventional applications work on the Android OS, the authors also have replaced several methods with the native method in order to reduce processing time. In the

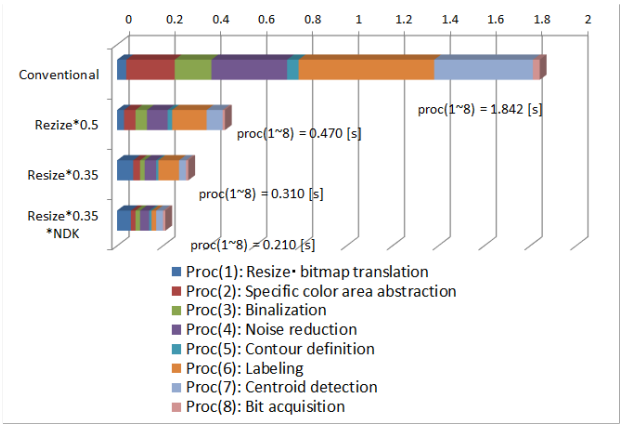


Figure 10 Result of performance evaluation

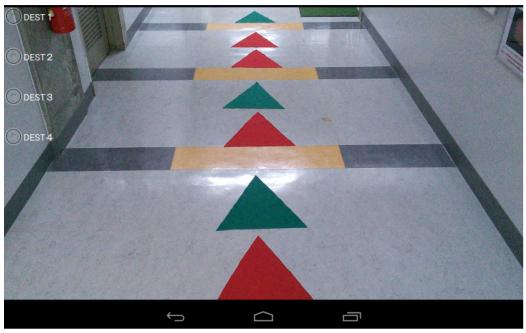
conventional navigation application, although some methods use the OpenCV library that provides the native method for image processing, other methods are coded as non-native methods.

3.3 Evaluation experiment

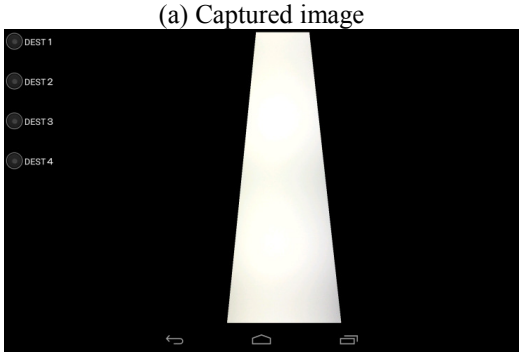
The authors divided the marker recognition process performed after each image capture into eight parts, as shown in Figure 9. First, the conventional RS-WYSIWYAS navigation application is installed into a smart tablet (Nexus 7 2013 model), and the average processing time for each process is recorded. The top of the data shown in Figure 10 is the result of the conventional application. This result shows that a large screen size costs approximately 1.8 seconds of the entire processing time. This processing delay could result unsatisfactory to users.

Next, the influence of image resizing is evaluated. The second and third data in Figure 10 are the results of the marker recognition process with pre-resizing of the captured image. In these results, a resizing rate of 0.35 shows approximately 0.31 seconds of processing time, which is approximately 1/6 of the conventional application, and this is the shortest processing time among the top three results. This processing time represents 3 Hz of the navigation updating cycle. When using a generic GPS receiver for navigation, the positioning updating frequency in most current smart devices is set at 1 Hz. Compared with this, the improved application can navigate with higher updating frequency. Meanwhile, although we attempted to test a lower than 0.35 resizing rate, the image processing program cannot present a navigation arrow on the screen because it cannot detect sufficient marker elements from the resized image.

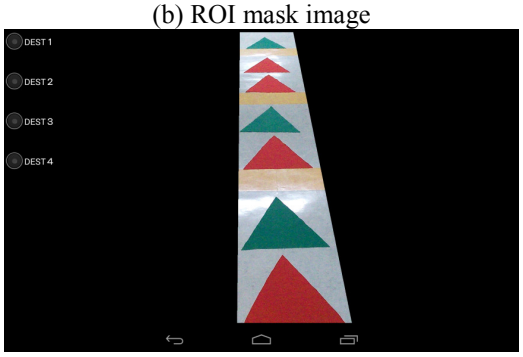
Furthermore, these results show that the processing time for labeling occupies the largest percentage of the total processing time. Accordingly, the authors replaced the labeling processing method implemented as byte code with native code by the Java Native Interface (JNI). The results of using the native code are shown at the bottom of Figure 10. This indicates that the labeling processing time is reduced. It also indicates that the total processing time has been reduced to 0.21 seconds; that is, a navigation update of approximately 5 Hz frequency is achieved.



(a) Captured image



(b) ROI mask image



(c) Pre-masked image

Figure 11 Generating pre-masked image

4 RELIABLE MARKER SEQUENCE RECOGNITION

4.1 ROI mask generation using past-extracted marker information

Typically, some of the objects on the corridor of a building have color that is similar to the marker. For the conventional application, it is difficult to distinguish non-marker objects with color similar to the positioning marker placed on the floor. If a non-marker object is abstracted and determined to be a marker, the application shows the wrong arrow symbol to the user. Therefore, the authors have implemented pre-masking processing for the captured movie that allows omitting the information of unnecessary areas where M-CubITS marker elements do not exist, and extract only the marker sequence information. To exclude unnecessary areas, the marker information detected in the previous processed frame is used. First, we derive a regression line using the centrobaric coordinates of the marker. Next, the Region of Interest (ROI) mask image is generated with a trapezoidal or triangular region along the regression line. In this process, if the regression line crosses the top of the screen, a trapezoidal region is placed along the regression line; meanwhile, if the regression line crosses the side of the screen, a triangular region is allocated along the regression line. Subsequently, an input image for marker detection is generated by multiplying the ROI mask and captured images. Figure 11 shows a generated pre-masked image.

Table 1 Device specifications (Google Nexus 7 2013)

CPU		Snapdragon S4 Pro (APQ8064)
	Frequency	1.5GHz
	Number of cores	quad core
memory		2GB
Camera	Front	120MP
	Rear	500MP
Screen	Size	7inch
	Resolusion	1920×1200
OS		Android 4.3

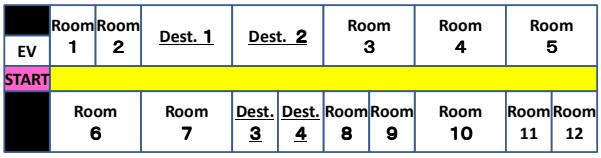


Figure 12 Diagram of experimental area

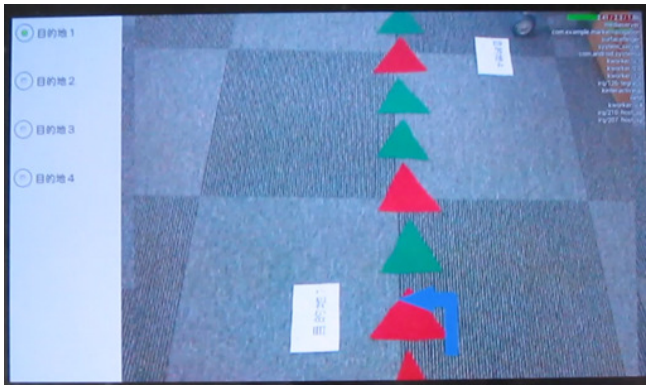


Figure 13 Navigation application user interface

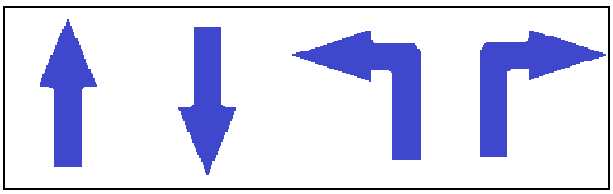


Figure 14 Arrow patterns

4.2 Evaluation experiment

4.2.1. Evaluation method

To evaluate the usability of the proposed system, M-CubITS marker elements for the WYSIWYAS navigation system were arranged for a building at the Kanagawa Institute of Technology. For this experiment, we made a marker arrangement where the triangle marker elements (40 cm × 40 cm and made of cloth) lined up in an interval of 50 cm on a corridor (40 m long), and made a database that

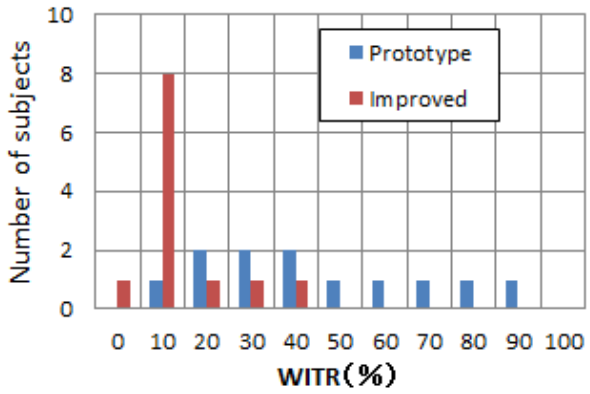


Figure 15 Wrong indication time ratio
(There are small non-marker objects)

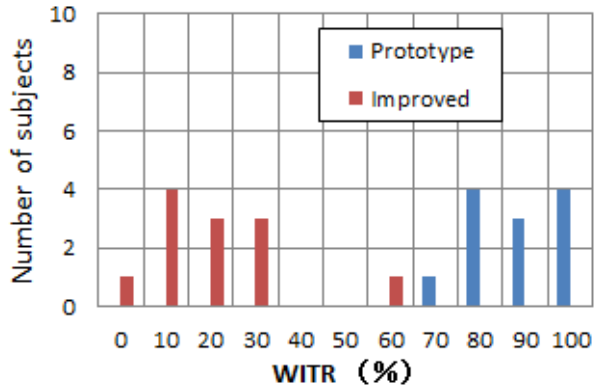


Figure 16 Wrong indication time ratio
(There are some non-marker objects)

corresponds to this marker sequence. We used an Android tablet, the Google Nexus 7 (2013 model). The specifications for this device are listed in Table 1. The navigation application is installed on the tablet. A subject first selects a destination from among four predetermined locations (classrooms). Then, the subject trains the backside camera to the floor of the corridor and captures in an image seven marker elements. The application then determines its position and orientation to the destination, and displays a guiding arrow on the movie to indicate the direction in which the subject should travel. Because all the subjects who participated in the experiment are students at the Kanagawa Institute of Technology, they could already know the location of each destination room. Therefore, we assigned each destination a name that was different from the actual room name. Figure 12 shows a diagram of the experiment area. Figure 13 shows the user interface for the application. There are four buttons on the left side of the screen; touching one selects a destination. During navigation, the device indicates the direction of the destination through a guiding arrow; the arrow patterns are shown in Figure 14. For this experiment, a right or left turn arrow indicates that the subject has arrived at the entrance of the destination room.

In order to evaluate the marker sequence recognition performance in marker object and non-marker object mixed environments, four types of experiment is conducted.

Table 2 Availability

	Fully available	Partially available	Little available
Prototype	2	4	6
Prototype (some non-marker object)	0	2	10
Improved	5	4	3
Improved (some non-marker object)	2	7	3

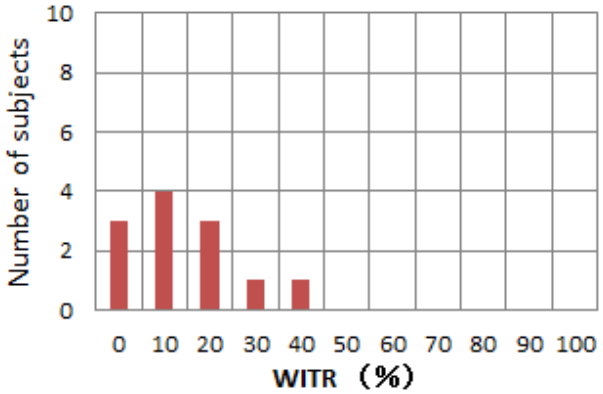


Figure 17 Acceptable WITR

i) Prototype app. / corridor without non-marker object
ii) Prototype app. / corridor with non-marker object
iii) Improved app. / corridor without non-marker object
iv) Improved app. / corridor with non-marker object

12 subjects conducted 4types of experiment by following procedure.

(1) The subject runs the navigation application (prototype or improved) on the smart device.
(2) The subject stands at the starting point.
(3) The subject is told the destination and sets the destination in the application.
(4) The subject searches for the destination according to the navigation instructions.

After subjects finish four experiments, they answer questionnaires. The questions are as follows:

Q.1 Wrong indication time ratio (WITR)
The wrong indication time ratio is defined as a time ratio that perceived as having been presented with the wrong indication. The subject estimates this value from 0% to 100% in step of 10%.

Q.2 Availability
A subject rates the usability of the application through navigation with three choices (fully available/partially available/not available)

Q.3 Acceptable WITR

This is defined as the value of WITR that the user may want to use the application. The subject answers this value from 0% to 100% in step of 10%.

4.2.2. Experiment results

Figure 15 shows the WITR in the case where there are few non-marker objects around M-CubITS marker elements. In this graph, the WITR using the prototype application distributes broadly. On the other hand, such distribution using the improved navigation application concentrates around 10%. This result indicates that most subjects feel that the improved application provides less wrong indications than the prototype application. Meanwhile, Figure 16 shows the WITR in the case where there are some non-marker elements. In this graph, the distribution of the WITR using the prototype application concentrates around 60% to 100%. We consider that this result is caused by marker sequence recognition error. On the other hand, such distribution using the improved navigation application concentrates around 0% to 30%.

Next, the results from questionnaire Q.2 are shown in Table 2. In this question, subjects choose availability of this system. For the prototype system, there are few users judging it to be available enough. Especially, in the case where there are some non-marker elements on the corridor, most users feel that the system is insufficient. On the other hand, for the improved system, the availability of subjects is increased. Whether there are non-marker objects or not, 75% of the subjects accept this improved application. However, three subjects feel that the improved application is not acceptable for navigation. This system is effective if it can extract correct marker information in previous frame. However if it cannot extract, it perform same processing as the prototype application. Because the user can recognize a marker on the screen, a function to input a centrobaric coordinates of extracted marker element into the application is expected.

Finally, the result from questionnaire Q.3 is shown in Figure 17. In this graph, the acceptable time ratio is distributed around 10%, and this result shows the same tendency as the time ratio perceived by subjects. On the other hand, three subjects answered that 0% of WITR is required, which corresponds to the results from questionnaire Q.2. Although the improved application reduces marker sequence recognition error, some subjects are not satisfied. This result will be an indicator for usability improvements.

5 INTRODUCTION TO MULTIPLE FLOORS

Conventional indoor navigation applications can only navigate buildings in a single layer. Moreover, the data from indoor maps are implemented in the application. However, if there are changes in the building, it is not possible for application developers to modify the map data.

In this study, the data from indoor maps are separated, and the application downloads the map data from an HTTP server installed in the private network of the building. Moreover, to navigate in multiple floor buildings, the indoor floor map is defined in the CSV format, and a function to guide to the elevator hall is added. Figure 18 shows a guiding example. Visitors to the building access the HTTP server to download the map data.

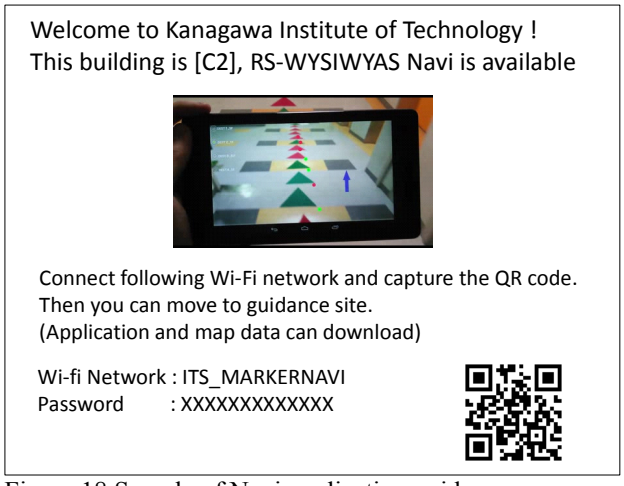


Figure 18 Sample of Navi application guidance

Figure 19 Navigating to EV hall

In order to evaluate the feasibility of downloading the map data, the time elapsed until the visitor downloads the map data, and the time to run the application, and we conducted another experiment. Seven subjects performed the following operation twice:

① Read a guidance paper
② Connect to the Wi-Fi network
③ Capture the QR code
④ Access the HTTP server
⑤ Download the map data
⑥ Run the application

Assuming the use in different buildings, different guidance sheets were used for the first and second operations. The average of the first operation time was approximately 100 seconds; the average of the second operation time was approximately 40 seconds.

Furthermore, navigation was tested in multiple floors. When the subject moved to a different floor from the current floor, the improved application navigated the subject to the elevator hall (EV hall). Then, the subject used the elevator to move to another floor. After the subject got off the elevator, the navigation application guided the subject to the destination room. Figure 19 shows the navigation display.

Table 3 Summary of improvements

	Conventional	Improved
Response time	Unsolved	Image resizing and use of native method
Reliable marker sequence recognition	Unsolved	Pre-mask processing using previous frame information
Database (DB) construction for indoor navigation	DB is included in application program	DB is separated from application program and stored in private server located in building

6 CONCLUSION

In this paper, the usability of a RS-WYSIWYAS pedestrian navigation system for smart devices was improved by improving the response time and marker sequence recognition performance. Improvements in the conventional RS-WYSIWYAS pedestrian navigation system for smart devices are summarized in Table 3. From the perspective of the response time, resizing captured images reduced the computation load. In addition, the implementation of native methods also reduced processing time. The results of processing time measurements indicated that a navigation updating cycle of approximately 5 Hz was achieved. Moreover, from the perspective of marker sequence recognition, pre-mask processing of captured images using the marker information obtained from a previous cycle improved the marker sequence recognition performance. The results of questionnaires for the navigation experiments indicated that the improved navigation application reduces the time ratio of wrong indication compared with the conventional application. In addition, most subjects considered that the availability of the improved application is higher than the conventional application.

REFERENCES

[1] J. Soubielle, I. Fijalkow, P. Duvaut, and A. Bibaut, “GPS positioning in a multipath environment,” IEEE Trans. Signal Process. vol. 50, no. 1, pp. 141–150, Jan 2002

[2]Jun Rekimoto and Katashi Nagao, " The world through computer," Proceeding of the ACM Symposium on User Interface Software and Technology (UIST95), pp.29-36. ACM Press, 1995.

[3]Itiro Siio, Toshiyuki Masui, Kentaro Fukuchi "Real-world Interaction using the FieldMouse" CHI Letters, Vol.1, Issue1, pp.113-119, ACM Press, 1999.

[4]Itiro Siio, "User Position Detection using RFID Tags" IPSJ-Technical Report 88, pp.45-50, 2000

[5]H. Matsubara, N. Fukasawa, K. Goto, K. Kawai, T. Sato, T. Aoki, N. Mizukami, K. Fujinami, A. Shinomiya, "Interactive Guidance System for Passengers," QR of RTRI 42(4), 201-206, 2001

[6] Takaaki Hasegawa, “A study on ITS and systems innovation,” Technical Report of IEICE, no. ITS 2002-120, pp. 13-17, Mar 2003 (in Japanese)

[7] Takaaki Hasegawa, “ITS platform EUPITS–approach to realization,” Technical Report of IEICE, no. ITS 2003-8, pp. 41-47, May 2003 (in Japanese)

[8] Takaaki Hasegawa, “ITS platform EUPITS–toward realization,” Technical Report of IEICE, no. ITS 2003-26, pp. 29-35, 2003 (in Japanese)

[9] Seiji Yamashita and Takaaki Hasegawa, “On the M-CubITS pedestrian navigation system by a camera-equipped mobile phone,” Proc. of ITSC2004, pp. 714-717, Oct 2004

[10] T. Manabe, S. Yamashita, and T. Hasegawa, “On the M-CubITS pedestrian navigation system,” Proc. 9th IEEE Int. Conf. on ITS, pp. 793-798, Toronto, Canada, Sept 2006

[11] T. Manabe, T. Hasegawa, Y. Matsuoka, S. Furukawa, and A. Fukuda, “On the M-CubITS pedestrian WYSIWYAS navigation using tile carpets,” Proc. 10th IEEE Int. Conf. on ITS, pp. 879-884, Seattle, WA, Oct 2007

[12] Yusuke Takatori, Hideya Takeo, “A real-time seamless WYSIWYAS navigation system for smart devices,” Proc. of ITST2013, pp. 163-168, Nov 2013

[13] http://www.eclipse.org/

[14] http://developer.android.com/index.html

[15] http://opencv.org/platforms/android.html

Implementation and Evaluation of a Road Information Sharing Scheme with a Still-Picture Internet Broadcasting System

Yoshia Saito* and Yuko Murayama*

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

Abstract – With an increase in navigation systems by portable terminals and car navigation devices, ITS which supports comfortable and effective driving has been evolved. However, in existing systems, it is difficult to offer real-time information to users since it needs to collect information from sensors and to analyze the collected information. Road information has to be provided timely so that drivers can pass through safe and comfortable roads. In this research, we use a still-picture Internet broadcasting system as a technique to share road information. It enables users to share the road information timely and to choose a road which is easy to pass. In addition, we implement an automatic photography function and conduct an experiment of the photography timing as a broadcaster. In this paper, we describe the design and implementation of our proposed system and evaluation experiments about the right photography timing of the automatic photography function. From the result of the experiments, we found road information was required more in the bad road situation such as rain and snow, and the road information could be grasped easily by introducing audience requests to the photography timing.

Keywords: Road Information Sharing, Still-Picture Internet Broadcasting, Photography Timing

1 INTRODUCTION

ITS (Intelligent Transport System) technology which addresses traffic information has been developed in recent years with the introduction of recent information technology to deal with increased volume of traffic. The road situation changes from moment to moment due to change of road surface conditions, weather conditions, traffic volume and various factors. The road information is an important factor and has high demand for drivers.

VICS (Vehicle Information and Communication System) [1] is one of systems to get road information in Japan. VICS provides road information which is collected by an information center via communication and broadcasting media such as FM multiplex broadcasting. Drivers can receive the road information by their car navigation systems and utilize it to select an appropriate route to the destination. However, VICS takes time to provide the road information to the drivers because the center needs to collect and analyze information. In addition, the ways of providing road information is limited to text, audio and map display. It is difficult for the drivers to understand the detailed road situation. The road information should be timely and provided in an easy-to-understand way.

Meanwhile we have studied a still-picture Internet broadcasting system [2-4] which uses still-pictures and audio streaming stead of video streaming to realize practical broadcasting via row-speed or limited high-speed cellular network by reducing data traffic. This system can broadcast anywhere using smartphones even if only connected to row-speed network. To provide timely and easy-to-understand road information, we propose a road information sharing scheme based on the still-picture Internet broadcasting system. The proposed system provide road information to users who want to know the road situation for route selection using still-pictures and audio streaming in real-time. To provide still-pictures and audio streaming, cooperative drivers set their smartphones on their cars. The broadcasting system works on the smartphones and automatically takes still-pictures and sends the still-pictures to the broadcasting server at the right time not to disrupt their driving operation. Users can select a car and view the broadcasting to get the road information in a certain area. The users can also communicate with the driver and get more detailed road information by the drivers through the communication.

In this paper, at first, we explain features and issues of existing traffic information systems. Secondly, we describe detail of the road information sharing scheme with a still-picture Internet broadcasting system. Then, preliminary experiment is conducted using implemented prototype system to study appropriate timing to take still-pictures. At last, we introduce several photography timing algorithms to the prototype system and evaluate the algorithms.

2 ISSUES OF EXISTING SYSTEMS AND OUR APPROACH

There are probe vehicle systems [5] to get road information. The probe vehicle system collects wide-area road information from probe vehicles which have various sensors and reduces cost for sensor installation. One of the services of the probe vehicle system, there is a vehicle tracking map which is provided by Honda [6]. This service shows whether the road is travelable or not on the map based on collected information from probe vehicles and aims to support driving in disaster area. It was used in 2007's the Niigataken Chuetsu-oki Earthquake and 2011's the Great East Japan Earthquake. This fact means road information is in great demand.

There are several issues in the probe vehicle system. At the first, the road information from probe vehicle system lacks of timeliness. For example, a service user checks road information to avoid traffic jams in advance but the road

could be backed up when the user arrived. This is because the probe vehicle system takes time to collect and analyze the road information from probe vehicles and the information lacks timeliness. Secondly, the road information is not flexible and intuitive. Typical road information which is provided by the probe vehicle system and VICS is predetermined by service providers and shown as text and icons. The users cannot know road situation in detail. Thirdly, the probe vehicle system requires dedicated sensor devices such as a specific car navigation system and it is difficult to prepare many probe vehicles. To get wide-are road information in real-time, the dedicated sensor devices should be eliminated. From these issues, it is important to provide flexible road information timely in an easy-to-understand way without dedicated sensor devices.

We focused on drive broadcasting which is one of broadcasting styles to show driving landscape using in-vehicle camera and communication devices. The audience enjoys the driving landscape and communicating with the broadcaster. The drive broadcasting is a popular content in live streaming services such as Ustream [7] and NicoNico Live [8]. The drive broadcasting can share road information timely. However, it has an issue about network communication. In the drive broadcasting, 3G/4G cellular network devices are used generally. Although 3G cellular network covers wide-area, it is too low-speed for video streaming. The video can be frequently stopped and low-quality. While 4G cellular network provides enough network bandwidth, it usually has limitation of amount of data traffic per day and month. Cellar carriers in Japan make communication speed slow when the subscriber uses hundreds of megabytes in a day or several gigabytes in a month. Therefore, the data traffic should be reduced for drive broadcasting.

3 PROPOSED SYSTEM

We have studied a still-picture Internet broadcasting system using smartphones which uses still-pictures and audio streaming stead of video streaming to reduce data traffic. Even if only 3G cellular network is available, the system realizes stable broadcasting. However, the previous study did not specify the use cases. In this research, we use the still-picture Internet broadcasting system for drive broadcasting to share road information.

The proposed system provides road information in real-time by live still-picture broadcasting so that users can choose safe and comfortable roads. The live still-picture broadcasting enables the users to understand road situation intuitively and also realize stable broadcasting anywhere by reducing data traffic. The users also can communicate with any broadcasters and ask a question about road situation.

Figure 1 shows the proposed system model. The broadcaster sets a smartphone on his/her car. The smartphone sends still-pictures and audio stream to the proposed system in order to share road information. It is on the assumption that the still-pictures are taken and sent automatically at the right timing. The broadcast programs are shown on a map.

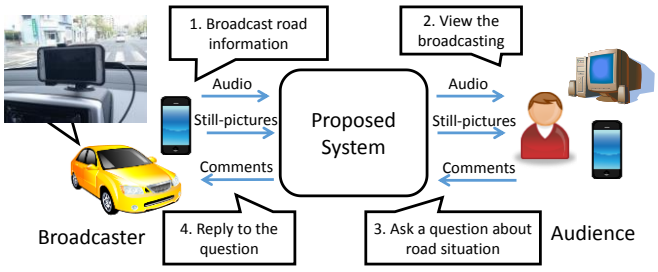


Figure 1: The proposed system model

Audience can select a broadcast program on the map and view the broadcasting by using their PCs or smartphones. The PCs/smartphones receive the still-pictures and audio stream from the proposed system. The audience can interactively ask a question about road situation to the broadcaster through the proposed system and the broadcaster can reply to the question to complement the road situation which is not understood by the still-pictures and audio stream.

The proposed system realizes timely road information sharing between broadcasters and audience and helps audience understand by still-pictures and interactive communication with broadcasters. Because it also does not require dedicated sensor devices and just uses smartphones, anyone can share road information and wide-area road information can be covered when there are a lot of broadcasters.

The use case of the proposed system is as follows. The broadcasters are people who drive for commuting or trip. The motivation of the broadcasters is to enjoy communicating with audience like fellow passengers. The audience are people who have a plan to go to the broadcasting place for commuting or trip and want to know the road situation. The audience can confirm traffic, road surface, and weather and so on by viewing the broadcasting to select routes.

4 PROTOTYPE SYSTEM

We implemented a prototype system based on the system model to conduct a preliminary experiment. A client software for broadcasters was developed on a smartphone and two client software for audience were developed on a smartphone and PC. Android smartphones were used for the client development.

Figure 2 shows the system architecture. At first, a broadcaster launches a broadcaster client on the smartphone and starts broadcasting. The audio broadcast function on the broadcaster client sends audio stream to a server using RTMP. The still-picture broadcast function on the broadcaster client sends still pictures which are encoded by JPEG2000 to the server. Since the broadcaster cannot touch the smartphone in driving, the automatic photography function takes still-pictures at right timing automatically. In this implementation, the timing is fixed time interval but it is variable.

On the server, Red5 which is a flash streaming server receives the audio stream from the broadcaster client. When

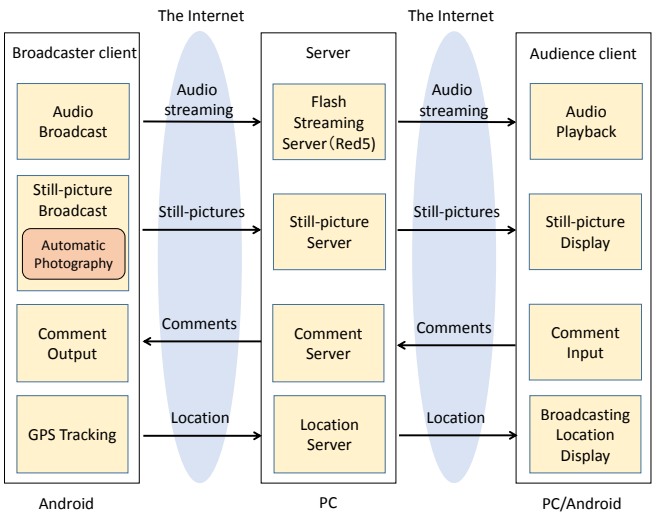


Figure 2: The system architecture

an audience client connects to the server and select a broadcasting, the Red5 sends the audio stream to the connected audience clients. The audience client receives and plays the audio stream. The still-pictures are also sent to the connected audience clients through a still-picture server which is implemented in Java. The audience client receives and display the latest still-picture.

The audience client can send text comments to the broadcaster through a comment server which is implemented in Java. The broadcaster client display the comments and read out the comments so that the broadcaster can communicate with audience without watching the smartphone in driving. The broadcaster client does not have the comment input function because the broadcaster cannot touch the smartphone in driving. The broadcaster hears the comments from audience and speaks about the reply to them.

The location of the broadcaster client is tracked by GPS on the smartphone and sent to the server. The server associates the location with the broadcasting and stores the location information in real-time. The audience can view the list of broadcastings on the map and select a broadcasting which they want to watch.

Figure 3 shows the user interface for the broadcaster client. The real-time camera image is displayed on the center. On the bottom part, there are control buttons. The connect button is used for connecting to the server. The login button starts to send audio stream and still-pictures to the server. The send button is used for manually sending still-picture to the server. The logout button stops the broadcasting. The comments from audience are displayed over the camera image and read out.

Figure 4 shows the user interfaces for the audience clients. The upside is smartphone version and the downside is PC version. The smartphone version is developed as an Android application and the PC version as a Web application. On start-up, the both audience clients show current broadcasting points on the map. When a broadcasting point is selected, the user interface is changed and the correspondent broadcasting is started. The audience can input comments. The inputted comments are displayed over the still-picture

Figure 3: The user interface for the broadcaster client

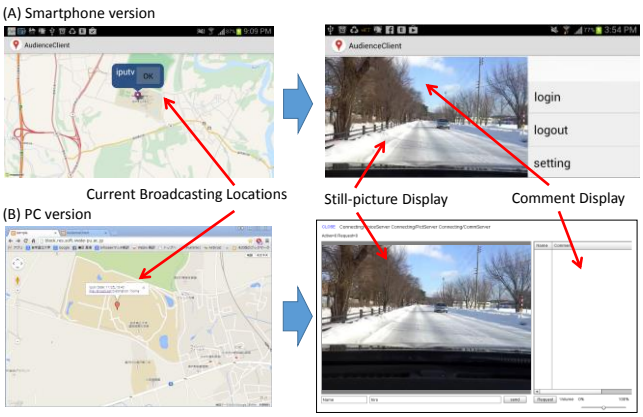


Figure 4: The user interfaces for the audience clients (A: smartphone version, B: PC version)

on the smartphone version and the comment display field on the PC version.

5 PRELIMINARY EXPERIMENT

To check operation of the prototype system and find its issues, we conducted a preliminary experiment. The experimental period was from May 5th to July 4th in 2014 and the broadcasting was performed in twice during daylight hours and night-time hours in a day. The subjects were 14 students of Iwate Prefectural University. They viewed the broadcasting using the audience client of PC version. The interval of the automatic photography was set to 30 seconds per a still-picture and the resolution of the still-pictures was 320x240.

From the experiment, we got several feedbacks from the subjects. Many subjects told about the automatic photography timing. The feedbacks were “I would like to shorten the interval of the automatic photography” and “I do not need so many still-pictures because the still-pictures are similar ones”. Furthermore, some subjects told about what still-pictures were required to help users understand the road situation. For example, the typical feedback was “During night-time, many rayless still-pictures were displayed and I did not understand the road situation when there were not streetlights”. The still-pictures have higher demand in well-lighted area than in dark area.

These results shows the automatic photography timing should be not fixed and more flexible. If the timing was a fixed interval, poor demand still-pictures could be sent to

Table 1: Utilization of road information in similar services

Service Name	Vehicle Speed	Sudden Braking	Road Surface	Traffic	Weather
SAFETY MAP		○			
EuroRAP	○		○		
SafeRoadMaps			○	○	○
Mi Drive	○				○

audience. To enable audience to understand road situation more easily, the photography timing must be determined when the audience can understand the road situation from the still-picture.

We make two hypotheses about the photography timing. The first hypothesis is that the demand of road information is changed depending on the road situation. For example, the demand of road information can be higher in a good weather condition than in a bad weather condition, and higher in a congested road than in a no traffic road. The second hypothesis is that the demand of road information is different from person to person. In the experiment, we got different feedbacks about the photography timing even if they watched same broadcasting. These are also mentioned in a related work. Münter [9] found drivers need more support when they don't have spatial knowledge and sense of direction of the person, and weather condition is bad. An effective photography timing algorithm needs to be studied based on the hypotheses to provide high demand still-pictures to the audience.

6 PHOTOGRAPHY TIMING ALGORITHM

We developed two photography timing algorithms to verify the hypotheses. The first algorithm changes the photography timing based on road situation utilizing sensors of a smartphone. This algorithm takes into account the first hypothesis. The second algorithm changes the photography timing based on audience request in addition to road situation. This algorithm takes into account the second hypothesis.

To develop the first algorithm, we researched similar road information services and what types of road information is utilized. Table 1 shows the result. The SAFETY MAP [10] which is provided by Honda uses sudden braking information for detecting unsafe points. The EuroRAP (European Road Assessment Programme) [11] which aims to reduce death and serious injury uses vehicle speed and road surface information. SafeRoadMaps [12] which is developed by University of Minnesota and Claremont Graduate University uses road surface, traffic and weather information for safety alerts. Mi Drive [13] which is provided by Michigan Department of Transportation (MDOT) uses vehicle speed and weather information for safety information. From these results, the first algorithm collects vehicle speed, sudden braking, road surface, traffic, and weather information by smartphone sensors and changes the photography timing based on these information.

Figure 5 shows the photography timing algorithm based on road situation. The algorithm starts operation when the driver puts on the brake. If the number of brakes for a given length of time is greater than 5, the algorithm shorten the

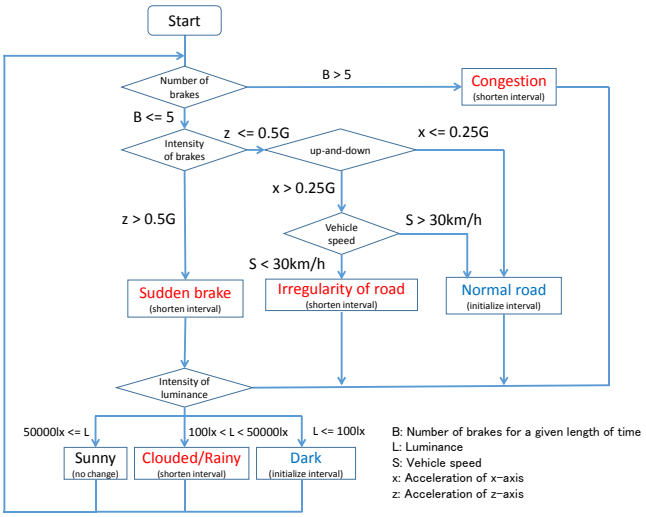


Figure 5: the photography timing algorithm based on road situation

interval of photography timing because a lot of brakes mean current road is congested [14]. If the number of brakes is below 5, it checks intensity of the brake to detect a sudden brake. The interval of photography timing is shortened when acceleration of z-axis which is anteroposterior acceleration of the vehicle is greater than 0.5G. If the acceleration of z-axis is below 0.5G, the algorithm checks acceleration of x-axis which is vertical acceleration of the vehicle and its speed to detect irregularity of road [15]. If the road is bumpy, the algorithm shortens the interval of photography timing. Otherwise, the interval will be initialized. After that, the algorithm checks weather and daylight sensing intensity of luminance. If the weather is clouded or rainy, the algorithm shorten the interval of photography timing. If it is in dark, the algorithm initializes the interval because the still-picture will be black one.

For the second algorithm which introduces audience request, in addition to the first algorithm, it shortens the interval when an audience request is received. The audience requests are sent from audience clients by pushing on a request button on the user interface. If our hypotheses are true, the first algorithm based on road situation will be more effective than the fixed interval one and the second algorithm will be more effective than the first algorithm.

7 EVALUATION

We introduced the two algorithms of photography timing to the prototype system and evaluate how well the system provide profitable road information to the audience. We compared the effects of the fixed interval scheme and the algorithm based on road situation and the algorithm based on road situation and audience request in an evaluation experiment.

7.1 Environment

A broadcaster drove on a predefined route near our university as shown in Figure 6 and broadcasted the driving

- **Zone from A to B**
  Using fixed interval
- **Zone from B to C**
  Using road situation algorithm
- **Zone from C to D**
  Using road situation and audience request algorithm

Figure 6: The driving route of the experiment

Table 2: The condition of the broadcastings

Date	Time	Weather	Num of Subjects
11/17	15:40~15:50	Rainy	5
11/25	13:00~13:40	Cloudy	2
11/26	15:30~16:10	Rainy	3
12/11	13:00~13:40	Cloudy	5
12/24	13:45~14:30	Snowly	5

scene with the prototype system switching the photography timing algorithms. The fixed or initial interval was set to 60 seconds. The smartphone which was used for the experiment was the au Galaxy S II. The maximum upload speed was 1.8 Mbps and download speed was 3.1 Mbps. The route included a broad road with heavy traffic, a narrow road in the neighborhood of housing estate, and a narrow road with many slopes and curves.

Subjects viewed the broadcasting using the audience client of PC version. A button for audience request, a question and answer section for the evaluation were added to the audience client. The subjects were 20 students who were from 19 to 22 years old, 17 male students and 3 female students. We conducted five broadcastings under the condition as shown in Table 2.

7.2 Results

Figure 7 shows the evaluation result comparing with each algorithm. We asked 5 questions to the subjects and they scored each question on 5-point scale. The blue bar shows the scores of the fixed interval one, the red bar is the algorithm based on road situation, and the green bar is the algorithm based on road situation and audience request. The first question shows usefulness of the proposed system for the route selection. The second question shows adequateness of the photography timing. The third, fourth and fifth questions show the understandability of weather, congestion and irregularity on the road respectively. About all questions, the green bar which is the score result of the algorithm based on road situation and audience request is highest. The score exceed 3 point which is the average score. Especially, the

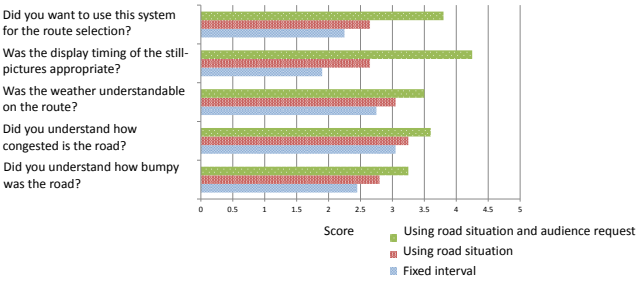


Figure 7: The result of the questionnaire comparing with each algorithm

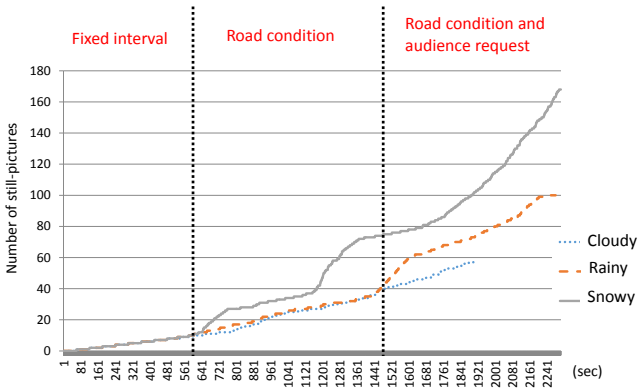


Figure 8: The number of still-pictures of each algorithm

score of the second question exceeds 4 point. The red bar which is the score result of the algorithm based on only road situation is higher than the blue bar which is that of the fixed interval one. Meanwhile the red bar scores below the average score on first, second and fifth questions. This result shows the audience request is effective to provide timely still-pictures for sharing road information. This means that one of our hypotheses, "the demand of road information is different from person to person" can be proven.

Figure 8 shows the number of still-pictures of each algorithm in different weather conditions. The horizontal axis indicates elapsed time from the beginning of the broadcasting. The vertical axis indicates the accumulated number of still-pictures. The algorithms were switched at the boundary of the vertical dotted line. From this graph, we found the number of still-pictures increased in worse weather conditions. Thus, the snowy condition increased the number of still-picture than rainy and cloudy conditions because the road surface was in the worse condition by fallen snow. The sensors of the broadcaster smartphone detects it and the algorithms shorten the photography interval. Considering the result of the questionnaire in Figure 7 and the result of the number of still-pictures in Figure8, one of our hypotheses, "the demand of road information is changed depending on the road situation" can be proven.

At last, we evaluated the amount of data traffic of each algorithm in order to realize stable broadcasting without

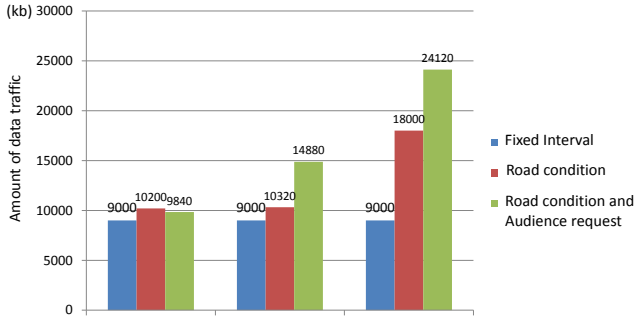


Figure 9: The amount of data traffic of each algorithm

large increase in data traffic. Figure 9 shows the amount of data traffic of each algorithm. The horizontal axis indicates weather and the vertical axis indicates the amount of data traffic. The green, red and blue bar means the same as Figure 7. From this graph, we found the data traffic increased in rainy and snowy weather conditions but the amount could be acceptable value. The amount of data traffic was highest in the snowy weather condition with audience request and it was about 24 MB. Comparing with the red bar, the data traffic increased about 30 % in the snowy weather condition. However, if we used a video streaming for sharing road information by Ustream in the same condition, the data traffic got about 114 MB. Since the algorithm based on road condition and audience request reduces about 80 % data traffic comparing with the video streaming scheme and realizes high user satisfaction, our proposed system can be effective and the photography timing algorithm should be based on road situation and audience request.

8 CONCLUSION

In this paper, we proposed a road information sharing scheme with a still-picture Internet broadcasting system. From the preliminary experiment, the right photography timing was an issue for the system. About the photography timing, we proposed two hypotheses that "the demand of road information is changed depending on the road situation" and "the demand of road information is different from person to person". Based on the hypotheses, we developed two photography timing algorithms which changes the photography timing based on road situation utilizing sensors of a smartphone, and based on road situation and audience request. From the evaluation, we found the algorithm based on road situation and audience request was most effective and our two hypotheses were proven.

In future work, we will improve the usability of the proposed system (e.g. the audience can see the vehicle information of the broadcasting in addition to still-pictures, and can watch the past broadcast programs by archiving environment.). We also study a business model to provide more motivation to broadcast the driving

REFERENCES

[1] VICS, http://www.vics.or.jp/english/vics/index.html
[2] Y. Nakano, Y. Saito and Y. Murayama: A Proposal for a Still Picture Internet Broadcasting System with Dynamic Picture Quality Adjustment based on Audience Requests for Smartphone Broadcasters, The 1st IEEE Global Conference on Consumer Electronics, pp. 360-364 (2012).
[3] Y. Nakano, N. Hirota, Y. Saito and Y. Murayama: An Implementation of a Still Picture Internet Broadcasting System with Audience-oriented QoS Control for Smartphones and PCs Audience, Computer Software and Applications Conference Workshops (COMPSACW), pp.523-527 (2013).
[4] Y. Nakano, Y. Saito and Y. Murayama: Still Picture Internet Broadcasting System with Audience-oriented Bandwidth Control for Smartphone Broadcasters, The Ninth International Conference on Systems, pp. 102-105 (2014).
[5] Takayuki Nakata and Jun-ichi Takeuchi: Mining traffic data from probe-car system for travel time prediction, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 817-822 (2004).
[6] Honda vehicle tracking map, http://www.honda.co.jp/internavi/LINC/service/disastermap/
[7] Ustream, http://www.ustream.tv/
[8] NicoNico Live, http://live.nicovideo.jp/
[9] Daniel Munter, Anna Kotteritzsch, Tobias Islinger: Improving Navigation Support by Taking Care of Drivers' Situational Needs, Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '12), pp.131-138 (2012).
[10] Honda SAFETY MAP, https://safetymap.jp/
[11] EuroRAP, http://www.eurorap.org/
[12] SafeRoadMaps, http://saferoadmaps.org/
[13] MDOT – Mi Drive, http://mdotnetpublic.state.mi.us/drive/
[14] Prashanth Mohan, Venkata N. Padmanabhan, Ramachandran Ramjee: Nericell: rich monitoring of road and traffic conditions using mobile smartphones, Proceeding SenSys'08 Proceedings of the 6th ACM conference on Embedded network sensor systems, pp. 323-336 (2008).
[15] Jakob Eriksson, Lewis Girod, Bret Hull, Ryan Newton, Samuel Madden, Hari Balakrishnan: The pothole patrol: using a mobile sensor network for road surface monitoring, Proceeding MobiSys '08 Proceedings of the 6th international conference on Mobile systems, applications, and services, pp. 29-39 (2008).

Session 4:
Data Analytics
(Chair : Tomoo Inoue)

PBIL-RS: An Algorithm to Learn Bayesian Networks Based on Probability Vectors

Yuma Yamanaka[†*], Takatoshi Fujiki[†], Sho Fukuda[†], and Takuya Yoshihiro[‡**]

[†]Graduate School of Systems Engineering, Wakayama University, Japan
[‡]Faculty of Systems Engineering, Wakayama University, Japan
[*] s161065@sys.wakayama-u.ac.jp
[**] tac@sys.wakayama-u.ac.jp

Abstract - Bayesian Network is used as a probabilistic model to analyze causal relationship between events from data. To learn a near-optimal Bayesian Network model from a set of target data, efficient optimization algorithm is required to search an exponentially large solution space, as this problem was proved to be NP-hard. To find better Bayesian Network models in limited time, several efficient search algorithms have been proposed. In recent years, several algorithms to learn Bayesian Network structures based on genetic algorithms (GAs) have been proposed. Among them, algorithms based on PBIL (Population-Based Incremental Learning) are regarded as a better sort of algorithms to learn superior Bayesian Networks in a practical computation time. In this paper, we propose PBIL-RS (PBIL-Repeated Search), which is an improvement of PBIL. In PBIL-RS, if the search area becomes sufficiently small in the process of converging the probability vector, we in turn spread the search area and again begin the converging process, repeatedly. We performed an evaluation of PBIL-RS, clarified its characteristics, and showed the superiority in its performance.

Keywords: Bayesian Networks, PBIL, Evolutionary Algorithm, EDA, Information Criterion

1 INTRODUCTION

Bayesian Network is used as a probabilistic model to analyze causal relationship between events from data. Recently, rapid growth of the Internet and processing speed of computers have made us possible to analyze the causal relationship from a large amount of data, and Bayesian Network is one of the important data analysis methods that are useful in various research fields with large data such as bioinformatics, medical analyses, document classifications, information searches, decision support, etc.

However, there is one difficulty that learning Bayesian Network models is proved to be NP-hard [1]. In other words, solution space exponentially increases as the number of events in the Bayesian Network increases. Therefore, several near-optimal algorithms to find better Bayesian Network models within a limited time have been proposed so far. Cooper et al. proposed an algorithm to learn Bayesian Networks called K2 that reduced execution time by limiting the search space [2]. To limit the search space, K2 applies a constraint in the order of events. The order constraint, for example, means that future events cannot be caused of events in the past. However, in many practical cases, we cannot assume such an order constraint. Therefore, to learn Bayesian Networks in general cases, several approaches have been proposed. Many of them use genetic algorithms (GAs), which find better Bayesian Network models when we take more time for computation [3][4][5]. Meanwhile, recently, requirements for large-data analyses arise due to the growth of the Internet. To meet these requirements, more efficient algorithms to find better Bayesian Network models within smaller time are strongly expected.

On the background above, a number of authors have proposed a new category of algorithms. Those algorithms called EDA (Estimation of Distribution Algorithm) have been reported to find better Bayesian Network models [6][7][8]. EDA is a kind of genetic algorithms that evolves statistic distributions from which we produce individuals over generations. Namely, EDA is a stochastic optimization algorithm. From the result of Kim et al., PBIL-based algorithm performed the best among several EDA-based algorithms [7].

Blanco et al. presented the first PBIL-based algorithm for Bayesian Networks [9]. They showed that their PBIL-based algorithm outperforms the traditional K2 algorithms. However, his algorithm has a drawback that his algorithm easily falls into local minimum solutions because it does not include any mutation operation to avoid converging into local minimum solutions. To overcome this drawback, several mutation operations were proposed for PBIL-based algorithms to learn Bayesian Networks. Handa et al. introduced bitwise mutation (BM), which apply mutation operations in which each edge is added or deleted with a constant probability [6]. Kim et al. proposed transpose mutation (TM) that is designed specific to Bayesian Networks [7]. This operation changes the direction of edges in the individuals produced in each generation. Fukuda et al. proposed a mutation operator called probability mutation (PM) for PBIL-based algorithms to learn Bayesian Networks [8]. Probability mutation manipulates the probability vector to avoid converging at local minimum solutions. These mutation operators improved the performance of PBIL-based algorithms to learn Bayesian Networks by avoiding local minimum solutions. However, mutation operators also have a drawback that the searching area jumps to other areas due to mutations before searching the local areas deep enough to explore good solutions.

In this paper, we propose a new PBIL-based algorithm called PBIL-RS (PBIL-Repeated Search), which is an improvement of PBIL. In PBIL-RS, if the search area becomes sufficiently small in the process of converging the probability vector, we in turn spread the search area and again begin the converging process, repeatedly. By searching local areas deeply until the probability vector converges into a sufficiently small

area, PBIL-RS improves the performance of PBIL-based algorithms to learn Bayesian Networks. We performed an evaluation of PBIL-RS, clarified its characteristics, and showed the superiority in its performance.

The rest of this paper is organized as follows: In Section 2, we give the basic definitions on Bayesian Networks and also describe related work in this area of study. In Section 3, we propose a new efficient search algorithm called PBIL-RS to achieve better learning performance of Bayesian Networks. In Section 4, we describe the evaluation of PBIL-RS, and finally we conclude this paper in Section 5.

2 PRELIMINARY DEFINITIONS

2.1 Bayesian Network

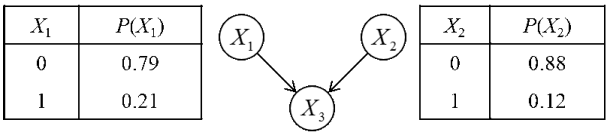
A Bayesian Network model visualizes the causal relationship among events through graph representation. In a Bayesian Network model, events are represented by nodes while causal relationships are represented by edges. See Fig. 1 for a concise example. Nodes X_1, X_2, and X_3 represent distinct events, where they take 1 if the corresponding events occur, and take 0 if the events do not occur. Edges $X_1 \rightarrow X_3$ and $X_2 \rightarrow X_3$ represent causal relationships, which mean that the probability of $X_3 = 1$ depends on events X_1 and X_2 . If edge $X_1 \rightarrow X_3$ exists, we call that X_1 is a parent of X_3 and X_3 is a child of X_1. Because nodes X_1 and X_2 do not have their parents, they have own prior probabilities $P(X_1)$ and $P(X_2)$. On the other hand, because node X_3 has two parents X_1 and X_2, it has a conditional probability $P(X_3|X_1, X_2)$. In this example, the probability that X_3 occurs is 0.890 under the assumption that both X_1 and X_2 occur. Note that, from this model, Bayesian inference is possible: if X_3 is known, then the posterior probability of X_1 and X_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 $O = \{o_j\}$, $(1 \leq j \leq S)$ be a set of observations, where S is the number of observations. Let $o_j = (x_{j1}, x_{j2}, \ldots, x_{jN})$ be j-th observation, which is a set of observed values x_{ji} on event X_i for all $i(1 \leq i \leq N)$, where N is the number of events. We try to learn a good Bayesian Network model θ from the given set of observations. Note that, good Bayesian Network model θ is the one that creates data sets similar to the original observation O. As an evaluation criterion to measure the level of fitting between θ and O, we use AIC (Akaike's Information Criterion) [10], which is one of the best known criterion used in Bayesian Networks. 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, compute a Bayesian Network model θ that has the lowest AIC criterion value.

2.2 PBIL

Recently, a category of the evolutionary algorithms called EDA (Estimation Distribution Algorithm) appears and reported



X_1	$P(X_1)$
0	0.79
1	0.21

X_2	$P(X_2)$
0	0.88
1	0.12

X_1	X_2	$P(X_3\|X_1, X_2)$	
		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: A Bayesian Network Model

to be efficient to learn Bayesian Network models. As one of EDAs, PBIL was proposed by Baluja et al. in 1994, which is based on genetic algorithm designed to evolve a probability vector [13]. Later, Blanco et al. applied PBIL to the Bayesian Network learning, and showed that PBIL efficiently works in this problem [9]. In PBIL, an individual creature s is defined as a vector $s = \{v_1, v_2, \ldots, v_L\}$, where $v_i(1 \leq i \leq L)$ 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 $v_i = 1$. 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 S that consists of C individuals according to probability vector P, i.e., element v_i of each individual is determined by the corresponding probability p_i.

(3) Compute the evaluation score for each individual $s \in S$ (In this paper we use AIC as the evaluation score).

(4) Select a set of individuals S' whose members have evaluation scores within top C' in S, and update the probability vector according to S' . Specifically, the formula applied to every p_i to update the probability vector is shown as follows.

$$p_i^{new} = ratio(i) \times \alpha + p_i \times (1 - \alpha), \qquad \text{(i)}$$

where p_i^{new} is the updated value of the new probability vector (p_i is soon replaced with p_i^{new}), $ratio(i)$ is the function that represents the ratio of individuals in S' that include edge i (i.e., $v_i = 1$), and α is the parameter called learning ratio.

(5) Repeat steps (2)-(4) until P converges.

By merging top-C' 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.

2.3 PBIL-based Bayesian Networks

In this section, we describes a PBIL-based algorithm that learns Bayesian Network models. Because our problem (i.e. Problem 1) to learn Bayesian Network 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 Bayesian Network model. Namely, with the number of events N, an individual model is represented as $s = \{v_{11}, v_{12}, \ldots, v_{1N}, v_{21}, v_{22}, \ldots, v_{N1}, v_{N2}, \ldots, v_{NN}\}$ where v_{ij} corresponds to the edge from an event X_i to X_j, i.e., if $v_{ij} = 1$, the edge from X_i to X_j exists in s, and if $v_{ij} = 0$ it does not exist. Similarly, we have the probability vector P to generate individual models as $P = \{p_{11}, p_{12}, \ldots, p_{1N}, p_{21}, p_{22}, \ldots, p_{N1}, p_{N2}, \ldots, p_{NN}\}$ where p_{ij} is the probability that the edge from X_i to X_j exists. A probability vector can be regarded as a table as illustrated in Fig. 2. Note that, because Bayesian Networks do not allow self-edges, p_{ij} is always 0 if $i = j$. The process of the proposed algorithm is basically obtained from the steps of PBIL, as described in the following.

(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 S as a set of C individual models according to P. (This step (2) is illustrated in Fig. 3)

(3) Compute the evaluation scores for all individual models $s \in S$.

(4) Select a set of individuals S' whose members have top-C' evaluation values in S, and update the probability vector according to the formula (i). (These steps (3) and (4) are illustrated in Fig. 4.)

(5) Repeat steps (2)-(4) until P converges.

Same as PBIL, the proposed algorithm evolves the probability vector so that we can generate better individual models. However, there is a point specific to Bayesian Networks, that is, a Bayesian Network model is not allowed to have cycles in it. To consider this point in our 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 v_{ij} according to P; every time v_{ij} is determined, if v_{ij} is determined as 1, we check whether this edge from X_i to X_j creates a cycle with all the edges determined to exist so far. If it creates a cycle, let v_{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 Bayesian Network models within the framework of PBIL. Note that the algorithm introduced in this section 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 improve the performance of the algorithm, several mutation operations have been proposed. A mutation operator called bitwise mutation (BM) was introduced by Handa [6]. BM applies mutations to each edge in each individual with a certain mutation probability. Kim et al. proposed a mutation operator called transpose mutation (TM), which is specifically designed for Bayesian Networks [7]. TM changes the direction of edges in the individuals produced in each generation. Fukuda et al. proposed a mutation operator called probability mutation (PM) for PBIL-based Bayesian Network learning [8]. PM manipulates the probability vector to avoid converging at local minimum solutions. These mutations avoid converging at local minimum solutions, and it improves the efficiency to learn Bayesian Networks with PBIL-based algorithms.

P		Parent Node					
		X_1	X_2	…	X_i	…	X_N
Child Node	X_1	0.0	0.5	…	p_{i1}	…	0.5
	X_2	0.5	0.0	…	p_{i2}	…	0.5
	⋮	⋮	⋮	⋱	⋮	…	⋮
	X_j	p_{1j}	p_{2j}	…	p_{ij}	…	p_{Nj}
	⋮	⋮	⋮	⋮	⋮	⋱	⋮
	X_N	0.5	0.5	…	p_{iN}	…	0.0

Figure 2: A Probability Vector.

3 PROPOSED ALGORITHM: PBIL-RS

We propose PBIL-RS (PBIL- Repeated Search), which is an algorithm to learn Bayesian Networks based on PBIL. To search for good Bayesian Networks efficiently, we introduce a new technique instead of mutation operators. Because mutation operators work with a certain mutation probability, they tend to change the search space before we deeply search the current search area to explore good solutions. As a result, efficiency of the algorithm decreases by skipping the search areas where many superior solutions are likely to be buried. In contrast, in PBIL-RS, we transit the search space only after we search the current search area deeply, i.e., only after PBIL-RS judged that the search space gets converged. With this technique, we can search deeply the specific space in which superior solutions would exist while avoiding local minimums.

Figure 5 shows the outline of PBIL-RS. In general, in the search space of Bayesian Network models, there are many local minimum points. Because models with similar structures tend to have similar evaluation scores, superior solutions would likely be collected at several local areas in the search space. Our algorithm PBIL-RS explores these areas with the following steps: (1) Initially, PBIL-RS sets the search space as the whole solution space. (2) As the algorithm proceeds and the generation grows, the search space usually gets smaller

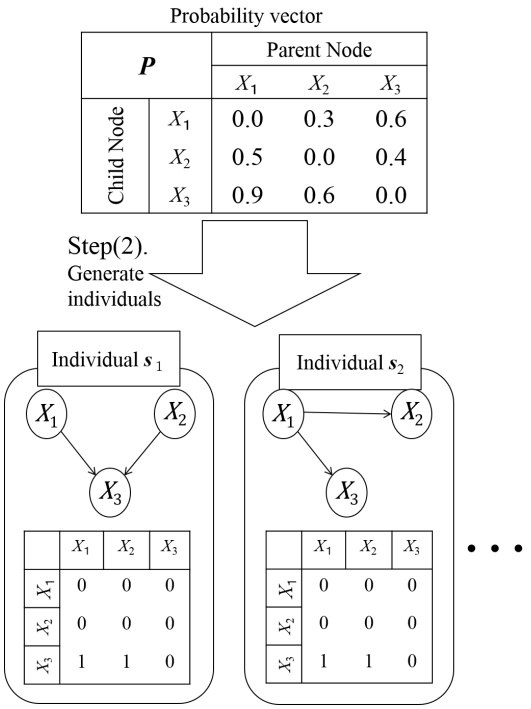


Figure 3: Generating Individuals from Probability Vector.

by focusing on an area in which superior solutions would be likely to exist. When the search space converges to a sufficiently small area, and PBIL-RS judges that the current area is sufficiently searched out, and (3) PBIL-RS in turn spreads the search space to explore different local minimum areas. Here, if the size of the spread search space is not sufficiently large, it may again fall into the same local minimum area. In order to avoid this, PBIL-RS spreads it to be larger search spaces step-by-step. Specifically, the size of the spread search space is firstly small to search near local minimum areas, and if we cannot find superior solutions in the next convergence, we then try to spread to larger search spaces to reach more distant search areas.

PBIL-RS controls the search space with probability vector P. Each element $p_{i,j}$ of vector P represents the probability to have the corresponding edge (i, j) in the generated Bayesian Network models. Thus, if each element $p_{i,j}$ approaches to 0 or 1, then naturally we have a probabilistic bias in the structure of the generated Bayesian Network models: The closer to 0.5 each element of probability vector P is, the larger the variation of generated models, and the closer to 0 or 1 each element is, the smaller the variation is. Namely, the probability vector P controls the variation and the bias of the generated structures of Bayesian Network models. Based on this, for probability vector P, we define *convergence level* S as follows:

$$S = \frac{\Sigma_{i,j(i\neq j)}\{0.5 - |P_{ij} - 0.5|\}}{N(N-1)}. \qquad \text{(ii)}$$

Convergence level S takes the average of the difference between 0 (or 1) and each element of probability vector P. Namely, the less this value is, the smaller search space is. In

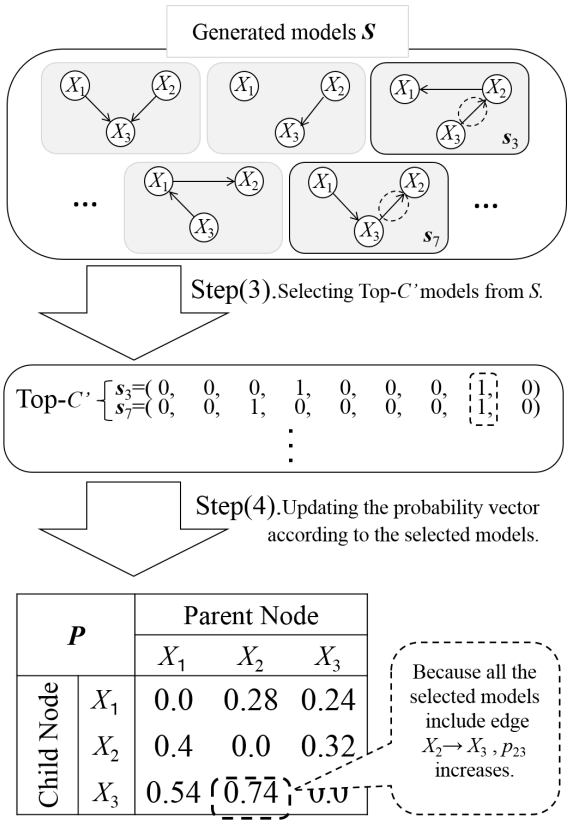


Figure 4: Step(3)(4): Updating Probability Vector.

PBIL-RS, generally *convergence level* S gets smaller as generation proceeds. Thus, PBIL-RS spreads the search space when the search space shrinks to be sufficiently small. To judge that the search space is sufficiently small, we introduce the number of search limitation k. Specifically, when *convergence level* S does not update the smallest value in the past k generations, i.e., the *convergence level* S in the k-th last generation takes the smallest value in the past k generations, PBIL-RS judges that the search space is sufficiently small and has converged.

When PBIL-RS detects the search space convergence, it in turn spreads the search space. We define H as the level to spread the search space. PBIL-RS modifies the probability vector P to increase the *convergence level* S to H. Specifically, we choose an element of P randomly, and reset it as $P_{ij} = 0.5$. This operation repeats until $S \leq H$ holds.

In addition, as mentioned previously, we change the value H dynamically to spread the search space and so avoid converging to the same local minimum areas repeatedly. More specifically, (a) we firstly initialize H with the initial value H_{min}, (b) secondly every time the search space is converged we increase H by a constant value *spread width* H_{inc}, and (c) lastly when we find the solution that has the best score so far, we again initialize H with the initial value. This operation enables PBIL-RS to leave a local area quickly when good solutions would hardly be found, and guide to the bigger search space.

The formal description of PBIL-RS is as follows. Processes (i)-(iv) are inserted into the steps (4) and (5) described in sub-

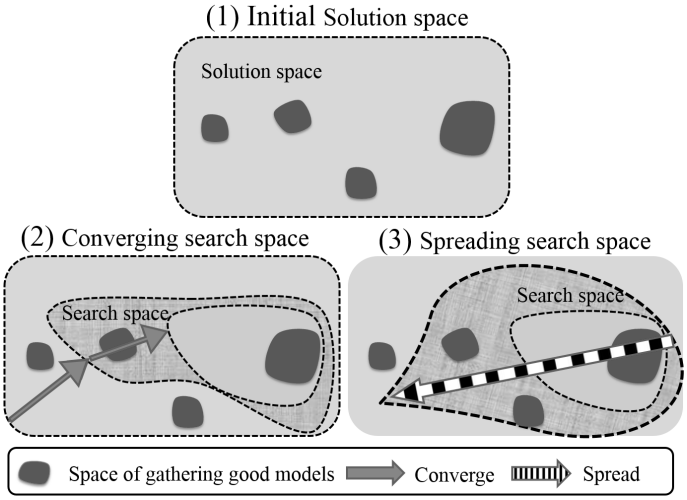


Figure 5: PBIL-RS Method

section 2.3.

(i) If the Bayesian Network model that has the best score so far is found, H is initialized by the initial value H_{min}.

(ii) Choose an element p_{ij} in P randomly, and reset it as $p_{ij} = 0.5$.

(iii) If $S < H$, then return to step (ii).

(iv) Augment H by *spread width* H_{inc}.

4 EVALUATION

4.1 Performance Comparison with Existing Methods

We designed our evaluation procedure as follows: We select Bayesian Network models used in our evaluation. In this paper, we use two well-known Bayesian Network models called Alarm Network [11] and Pathfinder [12], where Alarm Network represents the causal relation among events to monitor patients in intensive care units, and Pathfinder represents that related to the diagnosis of lymph node diseases. Note that Alarm Network includes 37 nodes and Pathfinder does 135 nodes. We generate an observation data set from each of the two models. In a Bayesian Network model, each node has a set of conditional probability so that we can obtain a set of values corresponding to all nodes according to the probability. Figure 6 shows an example of the data set generated from the example of Bayesian Networks shown in Fig 1, where j-th row represents an example of j-th observation set o_j generated according to the conditional probabilities of the model. We generated a data set that consists of 1,000 observations from each of two models. We use AIC criterion as the evaluation score, which is one of the representative criterion to measure the distance between the input data set and a Bayesian Network model.

We perform an evaluation of PBIL-RS in comparison with existing methods. We compare the performance of PBIL-RS

Data set

Observation No.	X_1	X_2	X_3
1	0	1	0
2	1	0	1
3	1	1	1
⋮	⋮	⋮	⋮

Repeatedly determine values of all nodes to generate a data set

Figure 6: Generating an Observation Data Set

Table 1: Parameters of PBIL-RS

Parameters	Values
# of observations	1000
Individuals in a generation (C)	1000
# of selected individuals (C')	10
Learning Ratio (α)	0.1
Search limitation (k)	10
Initial spreading level (H_{min})	0.4
Spread width (H_{inc})	0.1
Evaluation Score	AIC

with K2 that order restriction is evolved by genetic algorithms (K2-GA) [3], PBIL without mutations, and PBIL with three different mutation operators BM, TM, and PM. Parameter values used in the evaluation are shown in Table 1. Note that the mutation probability for BM, TM, and PM that performs the best is different for each mutation operators. Thus, we carefully chose those through preliminary experiments. For BM we use 0.005 that is the best performance mutation probability in range [0.001:0.2]. Similarly, for TM and PM, we use 0.1 and 0.002 that are the best in range [0.001:0.2] and [0.001:0.009], respectively.

Table 2 shows the comparison result summarizing the value of AIC calculated by each method. In Table 2, we show the performance of each method running 500 generations for two Bayesian Network models. In this result, we use the mean of 10 repetitions. Also, in Fig. 7, we show the transition of AIC values in the case of Alarm Network. From these results, we found that the PBIL-series methods perform far better than the traditional K2 although its order restriction is evolved by genetic algorithms, which proves the excellent ability of PBIL-based algorithms. Note that we could not compute the score of K2-GA for Pathfinder because it requires very large amount of time; it took 650 hours to proceed only 45 generations, whereas PBIL-RS took only 3 hours.

We also found that PBIL-RS has the best performance among those PBIL-based algorithms. This is because PBIL cannot continue searching after convergence (e.g., it finishes running at 160 generations in Alarm and 302 generations in Pathfinder), while BM, TM, and PM frequently change the searching area before exploring there deeply.

Table 2: AIC Values at 500 Generations

Methods	Bayesian Network Models	
	Alarm (37 events)	Pathfinder (135 events)
PBIL-RS	8536.4	30138.6
PBIL	8627.2	30243.7
PBIL + BM	8563.1	35240.9
PBIL + TM	8654.3	34784.2
PBIL + PM	8582.9	33003.0
K2-GA	13347.7	-

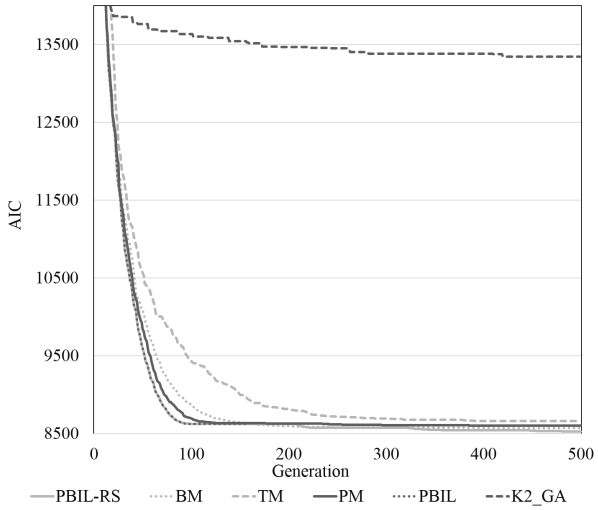


Figure 7: AIC Transition in Case of Alarm Network

4.2 Behavior in Varied Spread Level

In this section, we examine the behavior of PBIL-RS in which the value of spreading width is varied. We ran PBIL-RS in 3000 generations. Figure 8 shows the result where the line shows the transition of *convergence level S* as generation proceeds, and the dotted points shows the timing at which the best model is updated. We see the basic behavior of PBIL-RS such that each time P converges it spreads the search space, and we see that PBIL-RS continues finding better models around 3000th generations by changing the exploring areas adaptively.

5 Conclusion

In this paper, we proposed a new algorithm called PBIL-RS, which is an algorithm to learn Bayesian Network models. PBIL-RS is an extension of PBIL that avoids convergence to local minimum solutions by means of spreading the search space repeatedly whenever it converges to a small area. We evaluated the performance of PBIL-RS in comparison with existing algorithms, and we showed that PBIL-RS outperforms other existing algorithms regardless of the number of nodes in the Bayesian Network models used in the evaluation. Moreover, by examining the behavior of PBIL-RS, we verified that it properly controls the search space depending

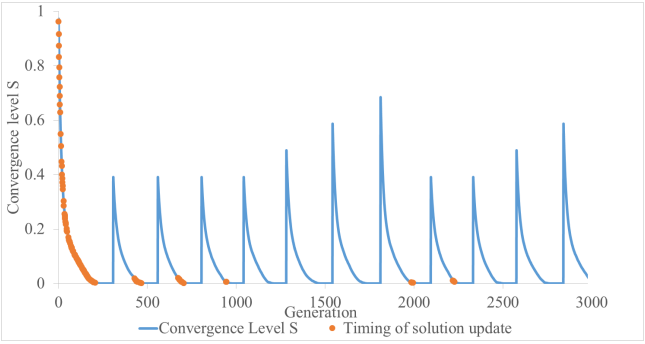


Figure 8: Behavior of PBIL-RS in Changing Convergence Levels

on the situation, and which leads to the superior performance.

As future work, more extensive evaluation using various Bayesian Network models is important. Especially, we would like to apply the models that include several thousands of nodes.

Acknowledgment

This work was partly supported by "the Program for Promotion 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] W.H. Hsu, H. Guo, B.B. Perry, and J.A. Stilson, "A Permutation Genetic Algorithm for Variable Ordering in Learning Bayesian Networks from Data," In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'02), pp.383-390 (2002).

[4] O. Barriére, E. Lutton, P.H. Wuillemin, "Bayesian Network Structure Learning Using Cooperative Coevolution," In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'09), pp.755-762 (2009).

[5] A.P. Tonda, E. Lutton, R. Reuillon, G. Squillero, and P.H. Wuillemin, "Bayesian Network Structure Learning from Limited Datasets through Graph Evolution," In Proceedings of the 15th European conference on Genetic Programming (EuroGP '12), pp.254-265 (2012).

[6] H. Handa, "Estimation of Distribution Algorithms with Mutation," Lecture Notes in Computer Science, Vol.3448, pp.112-121 (2005).

[7] 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).

[8] S. Fukuda, Y. Yamanaka, and T. Yoshihiro, "A Probability-based Evolutionary Algorithm with Muta-

tions to Learn Bayesian Networks," International Journal of Artificial Intelligence and Interactive Multimedia, Vol.3, No.1, pp.7-13 (2014).

[9] R. Blanco, I. Inza, P. Larrañaga, "Learning Bayesian Networks in the Space of Structures by Estimation of Distribution Algorithms," International Journal of Intelligent Systems, Vol.18, pp.205-220 (2003).

[10] H. Akaike, "Information theory and an extension of the maximum likelihood principle," Proceedings of the 2nd International Symposium on Information Theory, pp.267-281 (1973).

[11] 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. of Second European Conference on Artificial Intelligence in Medicine, Vol. 38, pp.247-256 (1989).

[12] D.E. Heckerman, E.J. Horvitz, B.N. Nathwani, "Towards Normative Expert Systems: Part I - the Pathfinder Project," Methods of Information in Medicine, pp. 90-105 (1992).

[13] S. Baluja, "Population-Based Incremental Learning: A method for Integrating Genetic Search Based Function Optimization and Competitive Learning," Technical Report CMU-CS-94-163, Carnegie Mellon University (1994).

Can three-month time-series data of views or downloads predict the highly-cited academic papers in open access journals?

Hiroshi Ishikawa*, Masaki Endo*, Iori Sugiyama**, Masaharu Hirota***, and Shohei Yokoyama****
*Graduate School of System Design, Tokyo Metropolitan University, Hino, Japan
**Faculty of System Design, Tokyo Metropolitan University, Hino, Japan
***Department of Information Engineering, Oita National College of Technology, Oita, Japan
****Graduate School of Science and Technology, Shizuoka University, Hamamatsu, Japan

Abstract- Currently, academic papers and their authors can be mainly evaluated by the statistics in Bibliometrics such as number of citations, h-index, and impact factor. However, it usually takes at least half a year or more for Bibliometrics-based approaches to evaluate academic papers. Open access journals, which do not restrict browsers, are spreading especially in US in recent years. Further, the number of viewing academic papers published in open access journals and the number of posting articles about the papers to social media continue to increase year after year. Such data can be treated as time-series data with immediacy. Therefore it is thought that if the academic papers as time-series data can be analyzed by proper machine learning methods such as clustering, it will be possible to extract the characteristics of highly-cited scientific papers. Instead of conventional evaluation of scholarly papers based on Bibliometrics, this paper discusses a method for estimating scientific papers with the potential of being highly cited in future based on the associated time-series data.

Keywords. *Open access journal; time-series data; Dynamic Time Warping; clustering; BIRCH*

1 INTRODUCTION

Recently, open access journals (OAJ), which do not restrict viewing, are spreading in the world, especially in US and the ratio of such journal papers over all academic papers is increasing accordingly [1]. In general, since OAJ publish accepted papers worldwide in one week or so, the academic papers can be accessed and viewed without restrictions by any user. Compared to traditional journals, OAJ can ensure the immediacy of the scientific papers by proving shorter periods from submission to publication. On the other hand, traditional surveys about academic papers are based on Bibliometric indices such as the total number of papers published by one author and the number of citations per paper. And such traditional surveys based on Bibliometrics tend to take long time just like submitted papers take long time to be published.

Very recently, as methods to evaluate the scientific papers in the immediate term, alternative Bibliometrics called Altmetrics, which use the posts to social media, such as Twitter (micro blogging), Facebook (blogging), and Mendeley (social bookmarking), and analyze the contents and numbers, are gathering attention. Typical services for evaluation of scientific papers by using Altmetrics include altmetric.com and ImpactStory. In Japan, Ceek.jp Altmetrics, a university-originated venture has begun to provide Altmetrics-based services [2]. In this paper, we mean by Altmetrics both methods for quantitative measurement of impacts of research products such as journal papers and data sets using the social media responses and activities as to measuring the future influences of emergent researches based on the results [3].

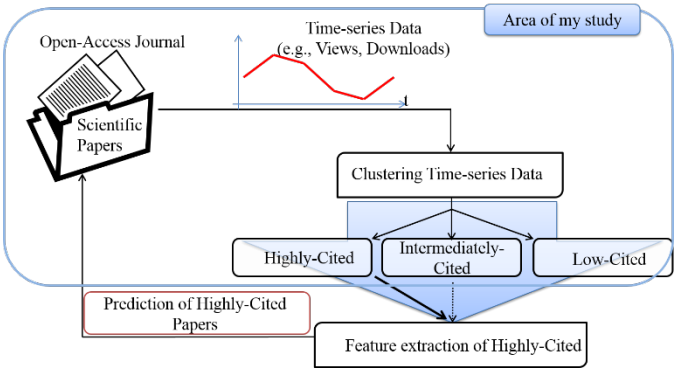


Fig.1. Big picture of our research.

So we have thought that it is necessary to clarify the relationships between Altmetric indices such as traffic data (i.e., number of views) and social media posts and Bibliometric indices such as citation data. In the preliminary experiment, we have found weak or very weak correlations between the number of views or downloads as of the first month after publication and that of citations of the scientific papers. This detail will be

TABLE I. Correlation between numbers of citations and numbers of views as of 1st month.

	Cited	Views
Cited	1	---
Views	0.0695	1

TABLE II. Correlation between numbers of citations and numbers of downloads as of 1st month.

	Cited	Downloads
Cited	1	---
Downloads	0.2852	1

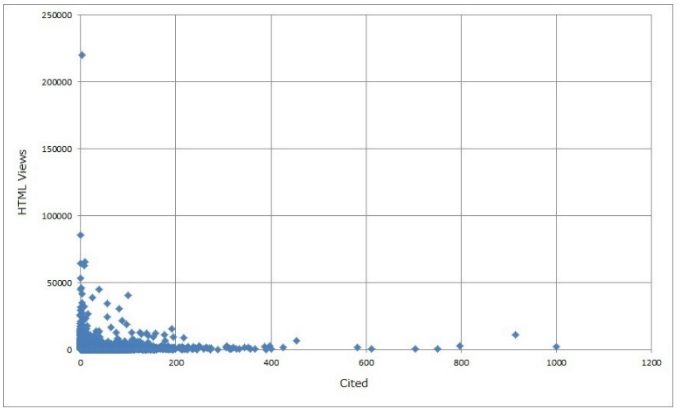


Fig.2. Number of citations vs number of views.

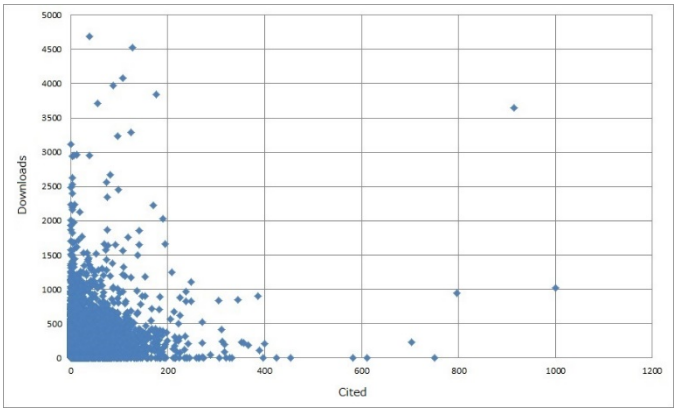


Fig.3. Number of citations vs number of downloads.

explained in Section 3 of this paper. Since these correlations are negligible, we have clustered academic papers published in OAJ with immediacy by paying attention to the counts of views and downloads so as to predict Bibliometric indices such as the counts of citations. We have used time-series data consisting of the numbers of views and downloads of papers per month for three or six months in clustering. Our final objective, which we have not fully obtained yet, is to estimate papers with the possibility of being highly cited in the future by performing machine learning, that is, learning classifiers for such papers based on clustering results. If the objective is fully attained, it will be possible to know the technological trends at the very early stage. Fig. 1 illustrates the big picture of our research as the flow of the associated processes. In this paper, we focus on the clustering of the academic papers based on the time- series data. Section 2 describes the relevant works. Section 3 explains the normalization of the time series data. Section 4 and section 5 describe our clustering method and experiments using the method, respectively. Section 6 summarizes the contribution of our work and describes future challenges.

2 RELATED WORK

Most of works that cluster time-series data focus on the frequency. As clustering time-series data based on the frequencies as features, subsequence time-series clustering [4] is often used. So as to exactly handle the frequency of time-series data as the features, time-series data longer than a certain length are required. However, as our work focuses on the immediacy of the scientific papers published in OAJ, we use only data recorded every month for at most 12 months from publication, that is, 12 pieces of time-series data. Therefore, we use the shape of a wave as features of the time-series data. Then we can calculate dissimilarity (i.e., distance) of the time-series data using Dynamic Time Warping DTW [5].

Recently, studies on Altmetrics are becoming very active and are expected to complement Bibliometric evaluation of scientific papers. Posts to Twitter (i.e., Tweets) referring to academic papers published in OAJ are observed for several days just after publication, according to Gunther [6], thereby enabling the prediction of the number of citations. However, the number of collected Tweets mentioning scholarly papers is not so large, partially because it is rather difficult to exhaustively find correspondences between academic papers and tweets. Nakahashi et al [7] have done automatic mapping between academic papers and Tweets and have attained better performance than the previous works. However, their research have used only tweets referring to papers presented in the traditional academic meetings. There still remain a lot of works to do so as to automatically find correct correspondences between Tweets and scholarly papers on the web, such as OAJ.

3 NORMALIZATION OF TIME-SERIES DATA

We used the data of academic papers published in Public Library Of Science (PLOS) [8], one of US-based OAJ. We used the API provided by the PLOS and obtained the data of academic papers for two times, on 7/20/2013 and 12/22/2013. While the numbers of views are equal to those of accesses to web pages provided to individual academic papers by PLOS (html views), the numbers of downloads are equal to those of the saved PDF (pdf views). The numbers of views and downloads are monthly calculated from the published month. They constitute time-series data. The numbers of citations as to some papers are also calculated at fixed intervals from the published month. The numbers of citations as to others, however, are those of citations accumulated from the published date to the collected date. Then we used the accumulated numbers of citations in a uniform fashion.

We acquired academic papers on 7/20/2013 whose number n = 52,386 and calculated the correlation coefficient r between the numbers of citations and those of views as of the first month of papers since publication and we obtained a very weak correlation (r = 0.0695) (See Table I). As to the same set of papers, we also calculated a correlation coefficient r between

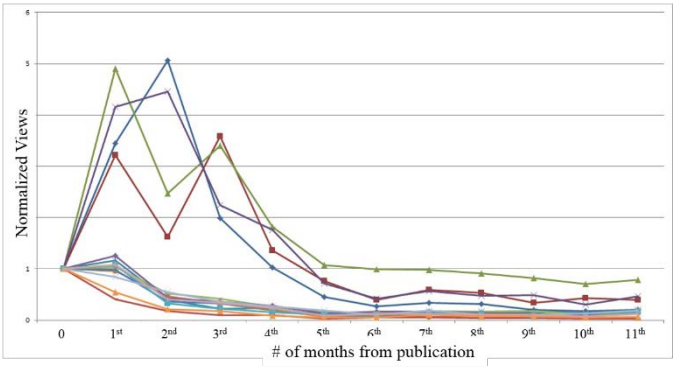


Fig.4. Examples of normalized time-series data.

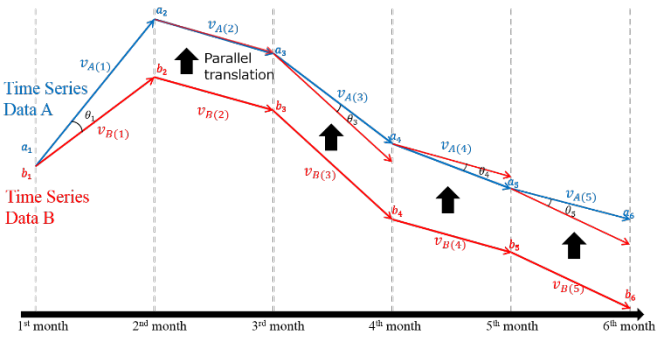


Fig.5. Vectorization of time-series data.

the numbers of citations and those of downloads as of the first month and obtained a weak correlation (r = 0.2852) (See Table II). The corresponding scatter charts are shown in Fig. 2 and Fig. 3. Then we judged these correlations negligible and decided to use time-series data of views and downloads instead. We normalized these time-series data by dividing all the values by those as of the first month. Fig. 4 shows examples of normalized time-series data as to the numbers of downloads of 14 samples published in 2/2004 for the first 12 months.

4 CLUSTERING

We examined two different clustering techniques for time-series data. First, we made vectors out of the time-series data and conducted k-means clustering based on the cosine measure as to the vectors. Next, we calculated dissimilarities between two pieces of the time-series data by Dynamic Time Warping (DTW) and conducted clustering based on the dynamic time warping squashed trees, an extension of CF tree.

4.1 K-means clustering

1) Vectorizing time-series data

First, let a_i be the number of views or downloads (exactly, the ratio over the value for the first month) recorded for i-th month as to the scientific papers A. As a whole, the time series data for the paper A is represented as A = a_1, a_2,..., a_I. Next we vectorize the time-series data as follows. So as to focus on the rate of change of each component of the time series data along the time line, we consider a sub-vector for the change of two consecutive elements. The x-component of the sub-vector is constantly 1 because the value is always one month. Each y-component of the sub-vector is represented as (a_{k+1}-a_k). Therefore, vectorized time-series data for the paper A is represented as a set of sub-vectors by the following formulas:

$$V_A = \{\overrightarrow{v_{A(1)}}, \overrightarrow{v_{A(2)}}, \dots, \overrightarrow{v_{A(k)}}, \dots, \overrightarrow{v_{A(kI-1)}}\} \quad (1)$$

$$\overrightarrow{v_{A(k)}} = \{1, a_{k+1} - a_k\} \quad (2)$$

Fig. 5 shows two examples of vectorized time series data.

2) K-means method

Here we consider the similarity between the academic papers A and B. The similarity between corresponding sub-vectors is calculated by using the cosine measure. Let the angle between them be θ_k. Then the cosine similarity between two vectors $\overrightarrow{v_{A(k)}}$ and $\overrightarrow{v_{B(k)}}$ is represented by the formula (3) (See Fig. 5). Note that as 0 <= cos θ_k <= 1, two vectors is similar if the cosine similarity is close to 1 or dissimilar if close to 0. Further, by using the total sum $\cos_{sum}$ and the total product $\cos_{prod}$ of the cosine similarity between corresponding sub-vectors, which are defined in the formulas (4) and (5), respectively, the similarity of the academic papers A and B (i.e., the vector sets $\boldsymbol{V_A}$ and $\boldsymbol{V_B}$) s is expressed as s = ($\cos_{sum}$, $\cos_{prod}$). Every piece of time series data has the same length, that is, I = J.

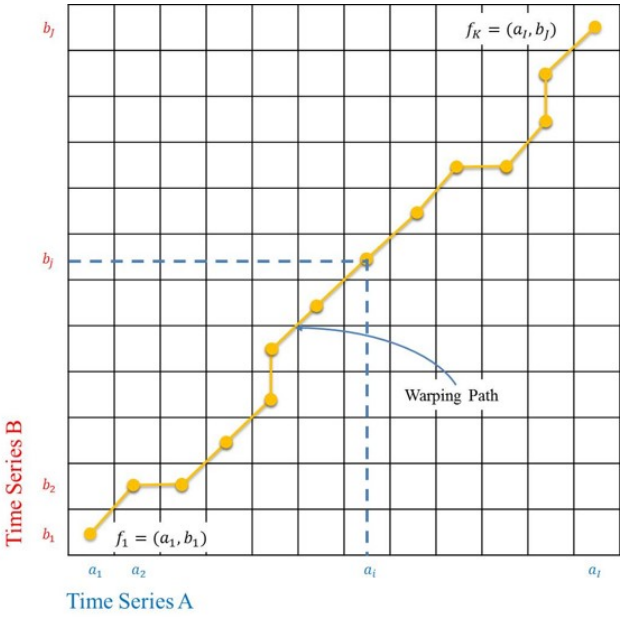


Fig.6. Example of warping path in DTW.

$$\cos\theta_k = \frac{\overrightarrow{v_{A(k)}} \cdot \overrightarrow{v_{B(k)}}}{|\overrightarrow{v_{A(k)}}| \cdot |\overrightarrow{v_{B(k)}}|} \quad (3)$$

$$\cos_{sum} = \sum_{k=1}^{I-1} \cos\theta_k \quad (4)$$

$$(0 \le \cos_{sum} \le I - 1)$$

$$\cos_{prod} = \prod_{k=1}^{I-1} \cos\theta_k \quad (5)$$

$$(0 \le \cos_{prod} \le 1)$$

Here we cluster academic papers by using k-means clustering based on this similarity measure s. Beforehand, we sorted data for the whole academic papers according to the descending order of DOI (Digital Object Identifier) and divided into k groups as initial clusters. The centroid of the cluster is calculated as the vector set $\boldsymbol{Vm}$ expressed by the formula (6). Let N be the total number of the papers. The initial centroids are calculated by using the initial clusters.

$$V_m = \frac{\sum_{n=1}^{N} V_n}{N} = \{\frac{\sum_{n=1}^{N} \overrightarrow{v_{n(1)}}}{N}, \dots, \frac{\sum_{n=1}^{N} \overrightarrow{v_{n(I-1)}}}{N}\} \quad (6)$$

In one repetition, the similarities between each paper and all the centroids are calculated. If there is another cluster to which the paper is more similar than the current cluster, the paper is moved to the former cluster. In case of any movement at the

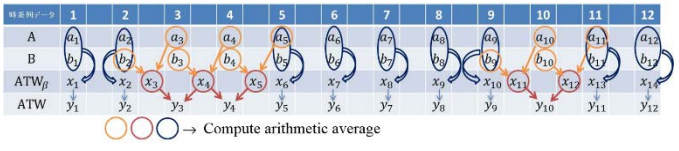


Fig.7. Computation of ATW.

end of the repetition, the centroids of the clusters are updated. This process is repeated until the number of academic papers to move is less than a certain threshold, that is, stable.

4.2 Dynamic time warping squashed tree clustering

1) Dynamic Time Warping

The Dynamic Time Warping (DTW) algorithm can calculate the dissimilarity between a pair of time-series data. The algorithm can map multiple points in one piece of time-series data to a single point in another piece of time series data, thereby allowing non-linear transformation as to the time axis.

Let us consider a pair of 12-month time-series data for two academic papers A and B. The time-series data are represented as $A = (a_1,a_2,\ldots,a_{11}, a_{12})$ and $B = (b_1,b_2,\ldots,b_{11}, b_{12})$. Then we construct a 12×12 matrix (i.e., table) by corresponding A and B to the row and column, respectively. In this table, grid points $f_k = (a_{ik}, b_{jk})$ represent correspondences between A and B. The series $F = (f_1,f_2,\ldots, f_k,\ldots, f_K)$ is called warping path. An example is shown in Fig. 6. The distance between a_{ik} and b_{jk} is denoted by $\delta(f_k)$ and is calculated by the formula (7). Using this distance, the evaluation function $\Delta(F)$ for the warping path F is calculated by the formula (8). Here w_k is a positive weight to f_k and is calculated by the formula (9).

$$\delta(f_k) = \left|a_{ik} - b_{jk}\right| \quad (7)$$

$$\Delta(F) = \frac{1}{I+J}\sum_{k=1}^{K} w_k \cdot \delta(f_k) \quad (8)$$

$$w_k = (i_k - i_{k-1}) + (j_k - j_{k-1}) \quad (9)$$

$$i_0 = j_0 = 0$$

The smaller $\Delta(F)$ is, the smaller the dissimilarity between A and B is. In other words, mapping is better in that case. The warping path F with the minimum $\Delta(F)$ is the shortest warping path of A and B. In that case $\Delta(F)$ is used as the dissimilarity between the two time series data. This is calculated by using the function mlpy.dtw_std provided by the machine learning library mlpy [9] in the programming language Python.

2) Dynamic time warping squashed trees

Dynamic time warping squashed tree (DTWS tree) [10] is a height-balanced binary tree, a variant of CF tree, when BIRCH, a typical hierarchical clustering method is adapted to time-series data. This method clusters time-series data based on DTW-dissimilarities by exhaustive searching, doing data compression.

First, let us consider the node vector CF of the original CF tree. Generally using N_0 as the number of elements belonging to the node, the linear sum of vectors $\boldsymbol{LS}_0 = \Sigma_{k=1,\,N_0} \boldsymbol{X_k}$, and the squared sum $SS_0 = \Sigma_{k=1,\,N_0} (\boldsymbol{X_k})^2$, the node CF of the CF tree is represented as $CF = (N_0, \boldsymbol{LS}_0, SS_0)$. On the other hand, the node vector $DTWS$ of the DTWS tree corresponds to the CF vector but the squared sum SS_0 is omitted from the CF and is represented as $DTWS = (N, \boldsymbol{ATW})$. Here N is the number of time-series data belonging to the node $DTWS$ and $\boldsymbol{ATW}$ is the average vector of time-series data. $\boldsymbol{ATW}$ is calculated as follows: First the average of two time-series data A and B is calculated as $\boldsymbol{ATW}_\beta = (x_1, x_2, x_k,\ldots, x_K)$. Here x_k correspond to f_k, calculated by the formula (10). As the average vector $\boldsymbol{ATW}_\beta$ is, in general, longer than the original time-series data A and B ($K > = I = J$), the average vector is compressed as $\boldsymbol{ATW} = (y_1, y_2,\ldots, y_{k'},\ldots, y_I)$ in accordance with the original time-series data. If the path for x_k extends diagonally on the warping path, its length will be 1; If the path for x_k extends either horizontally or vertically, its length will be 0.5. Based on this calculation, $\boldsymbol{ATW}_\beta$is compressed to $\boldsymbol{ATW}$. The value of $y_{k'}$ is calculated, that is, equal to that of x_k by the formula (11) if the length to x_k from the starting point of the time-series data is integer, otherwise it is calculated as linear interpolation between the two consecutive elements by using the formula (12). As to the time warping path in Fig. 6, the processes of the above calculation are shown in Fig. 7.

$$x_k = \frac{a_{ik}+b_{jk}}{2} \quad (10)$$

$$y_{k'} = \begin{cases} x_k & (11) \\ \dfrac{x_k + x_{k+1}}{2} & (12) \end{cases}$$

TABLE III. Data used in the experiments.

Length of Time-series Data	Published	# of Papers	Data1		Data2	
			Intermediately- and Higly-Cited	Highly-Cited	Intermediately- and Higly-Cited	Highly-Cited
3 months	Before 2013/4	48261	10978	398	13890	579
6 months	Before 2013/1	43363	10978	398	13885	579
12months	Before 2012/7	33934	10978	398	13715	579

TABLE IV. Results of k-means clustering (k=10).

	Data1		Data2		Difference	
# of Papers in a cluster	Cited>=10	ratio(%)	Cited>=10	ratio(%)	Cited>=10	ratio(%)
13,659	2,586	18.93	3,298	24.15	712	5.21
4,542	1,459	32.12	1,790	39.41	331	7.29
4,174	516	12.36	760	18.21	244	5.85
3,392	1,256	37.03	1,476	43.51	220	6.49
1,765	809	45.84	911	51.61	102	5.78
1,707	231	13.53	346	20.27	115	6.74
1,460	538	36.85	640	43.84	102	6.99
1,045	190	18.18	269	25.74	79	7.56
955	58	6.07	115	12.04	57	5.97
801	341	42.57	397	49.56	56	6.99
780	54	6.92	94	12.05	40	5.13
762	234	30.71	296	38.85	62	8.14
751	65	8.66	113	15.05	48	6.39
721	124	17.20	182	25.24	58	8.04
545	195	35.78	219	40.18	24	4.40
524	143	27.29	211	40.27	68	12.98
510	106	20.78	151	29.61	45	8.82
491	39	7.94	65	13.24	26	5.30
485	17	3.51	36	7.42	19	3.92
469	198	42.22	220	46.91	22	4.69
436	17	3.90	39	8.94	22	5.05
432	122	28.24	155	35.88	33	7.64
429	177	41.26	196	45.69	19	4.43

TABLE V. Kruskal-Wallis test of the results by *k*-means clustering (*k*=10).

	χ^2	# of degree of freedom	p-value
Data1	3283.75	96	$p_1<2.2\times10^{-16}$
Data2	2962.84	96	$p_2<2.2\times10^{-16}$

The procedures of the DTWS tree are similar to those of the CF tree. The dissimilarity between the time-series vector to be added and the average time-series vector of the node is calculated based on DTW. If there exists a node with the minimum dissimilarity, less than a prescribed threshold, the new vector is added to the node.

5 EXPERIMENTS

5.1 Outline of the experiments

For the experiments, we used data that were collected from the scientific open access journal PLOS twice on 7/20/2013 and 12/22/2013, respectively. The data for the scientific papers include the numbers of views and those of downloads of each paper as of each month as time-series data starting just after the publication and the number of citations of each paper as of the time of collection. We collected data as to 52,555 scientific papers on 7/20/2013, among which 52,386 papers have information as to citations. Further, from the collection, we selected scientific papers which have 3-month, 6-month, and 12-month numbers of views and of downloads as of 7/20/2013. Data collected on 7/20/2013 and 12/22/2013 are called Data1 and Data2, respectively (See Table III). We used two clustering methods described in Sections 4.1 and 4.2 for comparison. In one run of experiments, we clustered the same collection of academic paper data separately based on the numbers of views and of downloads. As a result, we obtain two sets of clusters. By making intersections of two sets of clusters, we obtained one set of clusters as a final result. We evaluated the final set of clusters in terms of the number of citations. The average number of citations is 11.81 and the median is 5 in Data1. The low-cited papers are defined as those with less than 10 citations. The highly-cited papers are assumed to belong to the top 10 % of the whole collection with respect to citations. In the top 10% subset of Data1, the minimum number of citations is 87. For the simplification of the judgment when evaluating clustered results, the threshold for the highly-cited papers is set to 90 in our experiments. Therefore, the intermediately-cited papers are defined as those with 10<=citations<90. Table III shows some statistics about our scientific paper collections prepared for the experiments. The intermediately- and highly-cited papers and the highly-cited papers denote the items cited times ">=10" and ">=90", respectively.

5.2 Results

1) K-means clustering

We conducted k-means clustering on 3-month time-series data with respect to both views and citations as k = 10 and 25. Because each result consists of 10 or 25 clusters, the academic papers are divided into 10 × 10 or 25 × 25 clusters. However, there also exist clusters which contain no elements. Then clusters with more than 400 elements are picked up and described in Table IV for k = 10. In the table, "# of papers", ">=10", and "ratio" denote the number of elements in the cluster, the number of elements cited for 10 or more times, and the ratio over the cluster (%), respectively. By k-means clustering, no clusters with only intermediately- and highly-cited academic papers could be found. However, the Kruskal-Wallis test, a nonparametric method, was conducted on 100 clusters constructed by k-means clustering (i.e., k=10) using Kruskal.test function of the R language [11]. Table V shows the results of the Kruskal-Wallis test. As p_1, p_2 <= significance level α = 0.01 from Table V, it has been confirmed that there exist significant differences between the median numbers of citations of clusters. Note that, because there exist three clusters with less than two elements in the results and the tests were performed on the rest and thus the number of degrees of freedom was 96.

TABLE VI. Recall, precision, and *F*-value of highly-cited papers.

Length of Time-series Data	Threshold		# of Papers in a Cluster	Data	# of Highly-Cited Papers in a Cluster	R (%)	P (%)	F (%)
	Views	Downloads						
12	47	42	506	Data1	396	99.50	78.26	97.50
12	47	50	516	Data1	396	99.50	76.74	97.32
12	47	40	496	Data1	393	98.74	79.23	96.94
12	47	30	494	Data1	392	98.49	79.35	96.73
3	25	35	500	Data1	389	97.74	77.80	95.88
12	47	22	459	Data1	379	95.23	82.57	94.14
12	47	12	466	Data1	375	94.22	80.47	93.02
6	25	40	473	Data1	375	94.22	79.28	92.90
6	25	32	473	Data1	375	94.22	79.28	92.90
6	25	47	479	Data1	375	94.22	78.29	92.80

2) DTWS tree clustering

DTWS tree clustering was performed on 3-month, 6-month, and 12-month time-series data with respect to the number of views and of citations. DTWS tree is a clustering method using exhaustive searching based on thresholds. Then, the size of clusters and the number of clusters change, depending on the given thresholds. For both the numbers of views and of downloads, the threshold X takes one of 26 different values:{0.3, 0.5, 0.7, 1, 1.5, 2, 3, 5, 7, 10, 13, 15, 17, 20, 22, 25, 27, 30, 32, 35, 37, 40, 42, 45, 47, 50}. And time-series data have 3 different lengths. Then we get 2,028 clusters as a final result. From the result, we excluded clusters whose size were less than 10. Thus, we evaluated the result, focusing on clusters in the result whose size is larger than or equal to 10.

However, using 3-month time-series data, with 32 as the *view*-threshold and 47 as the *download*-threshold, approximately 80% of the intermediately- and highly-cited papers in Data1 could be successfully extracted. Similarly, approximately 72% of the intermediately- and highly-cited papers in Data2 could be extracted with 42 as the *view*-threshold and 15 as the *download*-threshold. As for 12-month time-series data, approximately 93% of the intermediately- and highly-cited papers in Data1 could be extracted with 37 as the *view*-threshold and 50 as the *download*-threshold. Approximately 79% of the intermediately- and highly-cited papers in Data2 could be extracted with 50 as the *view*-threshold and 40 as the *download*-threshold.

Further, evaluation of the clustering results was done focusing on one of the clusters using the *F*-value calculated from the precision ratio *P* and recall ratio *R* of the results. *R*, *P*, and *F* are calculated by the formulas (13), (14), and (15), respectively. The precision ratio is expected to increase as the

TABLE VII. Number of clusters in top clusters ranked by *F*-value for highly-cited papers.

Length of Time-series Data	# in Top 100 clusters	# in Top 200 clusters	# in Top 300 clusters
3 months	48	87	110
6 months	24	71	122
12 months	28	42	68

numbers of citations increase in a course of time. The weight β for the recall ratio is set to $\beta = 3.5$ in the formula (15).

$$R(\%) = \frac{relevant\ papers\ in\ the\ cluster}{relevant\ papers} \times 100 \qquad (13)$$

$$P(\%) = \frac{relevant\ papers\ in\ the\ cluster}{papers\ in\ the\ cluster} \times 100 \qquad (14)$$

$$F(\%) = \frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R} \qquad (15)$$

The top 10 clusters for the highly-cited papers sorted by *F*-values are shown in the descending order in Table VI. From the table, it is known that there exist clusters with recall ratios over 90% and precision ratios over 75%. The numbers as to the top 100 clusters, top 200 clusters, and top 300 clusters sorted by the *F*-values with respect to highly-cited papers for the 3-month-, 6-month-, and 12-month time-series data are shown in Table VII. From the table, it is known that clusters based on 3-month- and 6-month time-series data are relatively highly ranked.

3) Discussion

By k-means clustering, only highly-cited academic papers could not be extracted. In other words, at least under the k-means clustering, vectors made from time-series data of numbers of views and of downloads cannot effectively represent the characteristics of only highly-cited academic papers. However, by the results of the Kruskal-Wallis test, it is confirmed that there exist significant differences among the medians of numbers of citations as to clusters.

By merging multiple clusters with 90% or more precision ratios as to intermediately- and highly-cited papers based on DTWS tree clustering, 80% or more of the papers could be recalled. Further, it was found that there existed clusters containing intermediately- and highly-cited papers with the recall ratios of 75% or more and the precision ratios of 90% or more. It was also found that there existed clusters containing highly-cited papers with the recall ratios of 90% or more and the precision ratios of 75% or more. This indicates that vectors made from time-series data consisting of numbers of views and downloads can predict intermediately- and highly-cited papers under the DTWS tree clustering. The numbers of citations in the Data2 represent those of the scientific papers published 8 or more months ago. Therefore, by clustering 3-month time-series data by DTWS tree clustering, at best approximately 77% of the intermediately- and highly-cited papers published 8 or more months ago could be detected. Also, at best approximately 98% of the highly-cited papers academic papers could be detected. Further, clusters containing rather many low-cited scientific papers could be found. In summary, by clustering the 3-month time-series data by using DTWS tree clustering, each of low-, intermediately- and highly-, and highly-cited academic papers could be detected with high likelihood..

In this experiment, the number of days from publication was not sufficiently considered. Thus, the numbers of citations of papers with different publication dates were equally treated. It is expected that a clustering scheme considering exactly the number of days from publication can cluster the academic papers with higher accuracy.

6 CONCLUSION

We have clustered academic papers published in open access journals by using time series data of the numbers of views and downloads. We used k-means clustering and clustering DTWS tree clustering and compared the performances. As a result, clusters mostly containing highly-cited papers could be discovered. In other words, it has been confirmed that highly-cited papers in the first year in open access journals can be predicted by at least the first three-month time series data of views and downloads based on the extracted features of these clusters.

DTWS tree clustering experiments in 3 different lengths of time-series data were compared. However, in k-means clustering, k was uniquely fixed and only 3-month time-series data were used. For this reason, there remain possibilities that the results were not properly compared among the two clustering methods. It is thought that other advanced methods such as x-means can remedy some of the above problems and improve the comparison results.

We used the numbers of views and downloads as of every month. This is because PLOS releases time-series data every month. However, if finer-grained time-series data, for example, of every day, are available, more accurate clustering of scholarly papers may be possible. Further, a wide range of academic papers have been posted on PLOS. Therefore, by clustering papers in restricted areas, more specialized characteristics may be detected.

Further, as using 12-month time-series data for clustering lacks the immediacy of estimated papers, clustering based on data provided by Altmetrics such as posts in social media and links in social bookmarks is expected to overcome the problem. From our experimental results, the clustering results are expected to provide the features required for machine learning, that is, classification of highly-cited papers.

REFERENCES

[1] M. Laakso, et al., "The Development of Open Access Journal Publishing from 1993 to 2009," PLoS ONE 6(6): e20961. doi:10.1371/journal.pone.0020961, 2011.

[2] Ceek.jp Altmetrics http://altmetrics.ceek.jp/ Accessed 2013

[3] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger, "Altmetrics in the wild: Using social media to explore scholarly impact," arXiv preprint arXiv:1203.4745, 2012.

[4] Tsuyoshi Ide, "Why does Subsequence Time-Series Clustering Produce Sine Waves?" Knowledge Discovery in Databases: PKDD 2006, Lecture Notes in Computer Science, Vol. 4213, pp. 211-222, 2006.

[5] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," Readings in speech recognition, Alex Waibel and Kai-Fu Lee (Eds.). Morgan Kaufmann Publishers Inc., 1990, pp.159-165.

[6] Gunther Eysenbach, "Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact," Journal of Medical Internet Research, Vol.13, No.4-e123, 2011.

[7] Hikaru Nakahashi, et al., "Automatic Alignment of Tweet," IEICE Technical reports, Vol.113, No.105, .pp.65-70, 2013 (in Japanese).

[8] PLOS http://www.plos.org/ Accesses 2014

[9] Mlpy http://mlpy.sourceforge.net/ Accesses 2014

[10] Kazuki Nakamoto, et al., "Fast Clustering for Time-series Data with Average-time-sequence-vector generation Based on Dynamic Warping," Trans. JSAI, Technical papers, vol.18, No.3, pp.144-152, 2003 (in Japanese).

[11] R http://www.r-project.org/ Accessed 2014.

Real-time Log Collection Scheme using Fault Tree Analysis

Naoya Chujo[†], Akihiro Yamashita[‡], Nobuyuki Ito[‡], Yukihiko Kobayashi[‡], and Tadanori Mizuno[†]

[†]Faculty of Information Science, Aichi Institute of Technology, Japan
[‡]Mitsubishi Electronic Engineering Co., Ltd., Japan
{ny-chujo, tmizuno}@aitech.ac.jp
{Yamashita.Akihiro, Ito.Nobuyuki, Kobayashi.Yukihiko}@ma.mee.co.jp

Abstract - The increasing complexity of embedded systems in information and communication technology causes a problem with locating faults during system failures. One reason for this problem is that complicated systems consist of so many components that basic log data do not contain useful information about abnormal system behavior by faulty components. Since available time resources in real-time systems are limited, we cannot use much time for logging all data to specify the faulty components.

In this paper, we present a real-time log collection scheme using Fault Tree Analysis for locating the faulty components[1]. Fault Tree Analysis is applied for assumed system failures, and then specific data in fault trees are defined to locate the faulty components. Log tasks are scheduled to collect the specified data in cooperation with system tasks. Once the assumed system failure is observed during system operation, the related log tasks wake up and collect the specified data to diagnose system faults. The experimental results have shown that specified data related to faulty components are collected by log task and the overhead for logging is predictable.

Keywords: Fault Tree Analysis, Log Data, Fault Diagnosis, Real-time, Embedded System

1 INTRODUCTION

In recent years, while the embedded software in such as automobiles or medical devices has grown increasingly complicated, numerous real-time control systems have been developed as well. These complexities have caused to a range of problems, including reduced productivity and increasing difficulty in pinpointing fault origins. Thus, improving the reliability of such systems has become an important objective. Logging data has been popular method in order to improve the reliability.

The primary role of the log data associated with faults is to record items such as fault occurrence time and the nature of the fault. In addition, fault-diagnosis functions allow the collection of log data describing the basic status of the system at the fault occurrence time. However, using only this basic level of log data, it remains difficult to determine the factors that caused the fault to arise.

The remainder of this paper is organized as follows. In Section 2, we review related work and discuss case studies involving automobile fault diagnosis and reliability improvements using a fault-analysis model. In Section 3, we describe our proposed method to collect log data based on FTA for real-time system. In Section 4, the experiments using the miniature car and the motor control system are presented. The results shows that faults of real-time system can be detected by the proposed method. Moreover, the overhead for collecting log data is evaluated, since predictable overhead is important for real-time system. In Section 5 we discuss these results. Our conclusions are presented in Section 6.

[1]This research is partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 2014-2016(26330074)

2 RELATED RESEARCH

Real-time control systems for applications such as automobiles and medical devices require extremely high reliability, and various methods exist for improving system reliability. In this section, we consider various methods of improving the reliability of real-time control systems and discuss a case study involving automotive fault-diagnosis functionality. We then describe a fault tree analysis (FTA)[1] for fault diagnosis.

2.1 Fault Diagnosis in Automobiles

The field of automotive fault-diagnosis functionality provides a case study of log data collection in a real-time control system. For example, on board diagnosis (OBD) [2][3], which is a tool for diagnosing system status in automobiles, conducts automatic diagnosis via computers embedded in automobiles, and most automobiles in service today are equipped with OBD.

The basic scope of OBD encompasses both monitoring and data recording and communication. Monitoring refers to the flashing of malfunction indicator lamps (MILs) when relevant items exceed fault detection criteria, while data recording and communication refers to the recording (in the event of a fault) of a code to specify the fault. A diagnostic tool can subsequently be used to read this code.

It is generally believed that monitoring frameworks in OBD systems should grow increasingly sophisticated in the future. Monitoring frameworks of this sort include an automotive fault-diagnosis function that allows computers to detect faults.

When an automobile detects a fault, it records diagnostic trouble code (DTC) that encodes information on the sensors involved in the fault, the events that have been diagnosed and the basic status of the automobile. The basic status data is called Freeze Frame Data (FFD). However, since DTC and FFD record anything other than basic status information, it is

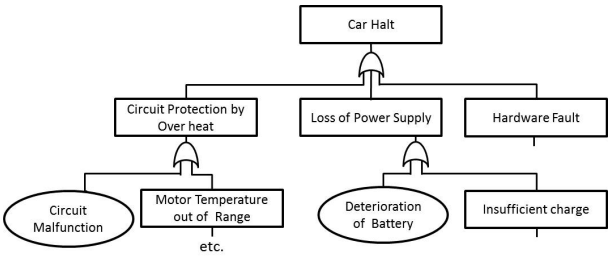


Figure 1: FT diagram for the case of car halt.

difficult to specify the causes of a system fault based on those data alone.

Moreover, the real-time nature of the system makes it difficult to diagnose faults, since the period of control cycle by sophisticated controllers is the order of milliseconds. It is much shorter than the period of FFD, which is typically 500 msec[4], depending on the system.

2.2 Fault Tree Analysis to Improve Reliability

As an example of a fault tree analysis (FTA), consider the case of car halt. Figure 1 shows the fault tree (FT) diagram corresponding to the halt of Electric Vehicle (EV).

We note first that the event at the top of the diagram is an undesirable event within the system. In this case, we have positioned "Car Halt" as the top-level event. Possible causes for this top-level event include "Circuit Protection by Overheat", "Loss of Power Supply" and "Hardware Fault". Among these events, possible causes for "Circuit Protection by Overheat" include "Circuit Malfunction" and "Motor Temperature out of Range". We proceed in this way to trace the possible causes of each event.

Rectangles in the diagram show intermediate events that should have possible causes for the lower level events. Circles in the diagram show basic events that could cause system faults.

By specifying events that could cause system faults (top-level events), an FTA enumerates the causes of top-level events and the events that can become structural elements. This enumeration may then be used to analyze the causes and fault events that contributed to the system failure.

3 PROPOSED METHOD FOR COLLECTING LOG DATA

In this section, we describe our proposed method, which we have named *Log data collection using Fault Tree Expansion* (LoFTE). We then present our method by using a simple example of a system fault and the FT, after which we will demonstrate an example of fault-event identification.

3.1 Philosophy of LoFTE Method

The LoFTE method implements FTA at the system-design stage and determines both the collection schedule and data to be collected at the time of fault detection. Then, during the system-operation stage, log data are collected with proper consideration paid to the real-time nature of the system at fault detection times. Thus this method aspires to achieve real-time fault diagnosis by collecting log data during system operation. More specifically, our method executes the following procedures at the system-design stage and during system operation.

System design stage:

1. FTA based on the system design specifications.
2. Within the FT, specify the data to be collected by software.
3. Store collection schedules for the specified data within the control software.

System operation stage:

1. Identify fault(s) that arise during control tasks.
2. Report the log-data-collection task responsible for information associated with the fault(s)
3. The data associated with the fault(s) are then collected as log data based on the relevant scheduling.

We will now describe the system in more detail by referring to the example depicted in Figure 2. The results of FTA at the system-design stage are stored as FT_1, FT_2, and so on. In this case, we assume that fault in FT_2 has occurred and initiate the log-data collection task. Upon receiving this instruction, the log-data collection task collects the log data displayed for FT_2.

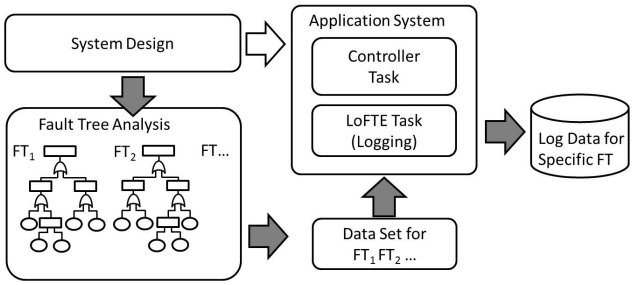


Figure 2: Schematic diagram of the LoFTE method.

3.2 Collection of Log Data

In the LoFTE method, collection of log data is executed based on information of the fault tree. Once a faulty event detected, all events in the fault tree should be recorded. However, log data of large number of events is not preferable, because the work for collection of log data takes long time and it makes difficult to analyze faulty events. To this end, the control software is analyzed to determine its module structure[5][6]. The collection of log data will be executed based on the structure.

In this subsection, we use the example of EV discussed in subsection 2.2, as an example of a real-time control system and consider the collection of log data.

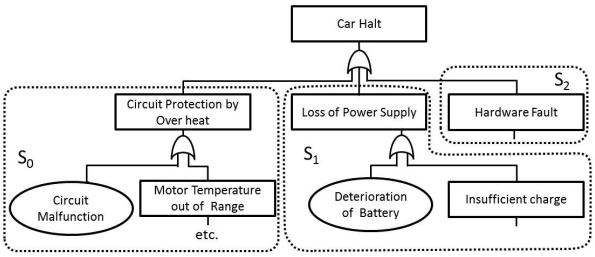


Figure 3: Subtrees for the case of car halt.

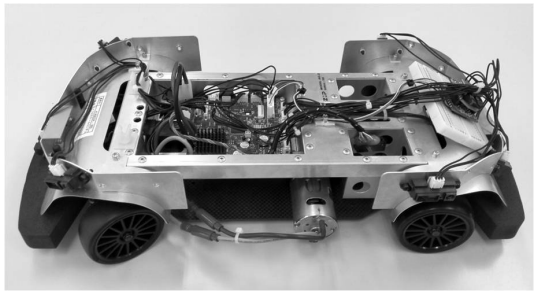


Figure 4: RoboCar 1/10 for AP.

The fault tree diagram in Figure 1 has three subtrees: S_0 of control circuit module, S_1 of power supply module and S_2 of other hardware module as shown in Fig 3. We assume that control software for each module is designed to be independent from others. Therefore, the work of collection of log data for fault tree will be divided to three parts.

4 EXPERIMENTS

In this Section, we describe our experiments conducted to test whether it is possible to identify the cause of a fault from collected sensor data and the FT. The overhead of the proposed method is estimated through experiments.

4.1 Experiment with Miniature Car

In the first experiment, we used a miniature car as an experimental device.

4.1.1 Experimental Device

The experimental device was a 1/10-scale miniature car, Robocar 1/10 for AP (Automotive Platform) [7], which was designed to be a research platform for autonomous driving, hereafter referred to as a RoboCar. Figure 4 shows a photograph of the experimental device, while Figure 5 shows a structural diagram of the RoboCar system.

The device is equipped with V850/FG4 CPU [8], multiple input devices, including a three-axis acceleration sensor, eight infrared range sensors, a three-axis gyro sensor, two FET temperature sensors, an motor encoder, and four wheel encoders. Sensor data from all of these input devices may be obtained from the RoboCar API. The device is also equipped with two output devices: a servo motor and a DC motor. These devices may also be controlled via the RoboCar API. Communications specifications correspond to CAN [9] and UART.

4.1.2 Software Used in the Experiment

In this experiment, we used TOPPERS/ATK2 [10], a real-time OS designed for next-generation automotive embedded systems. This OS was designed by the Center for Embedded Computing Systems at Nagoya University (NCES) and was designed to comply with AUTOSAR [11], a standard specification for automotive embedded software.

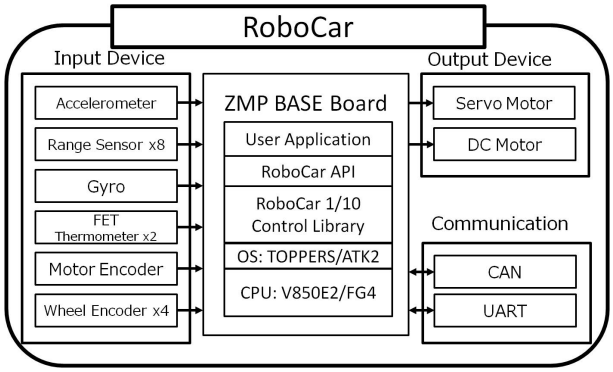


Figure 5: Strcuture of the RoboCar control system.

4.1.3 Experimental system

The experimental system used in this work consists of the RoboCar (the system in which the fault occurs) and a computer that monitors sensor data transmitted from the RoboCar via Bluetooth. Figure 6 shows a schematic diagram of the experimental system[12].

Three tasks were implemented as TOPPERS/ATK2 applications (LoFTE task: a task to collect log data, a control task, and a communication task), and the system was realized by periodically executing these tasks. The collection task collects data from the sensors installed on the RoboCar. These tasks are cyclically executed and the cycle period is 100 msec.

The control task controls the various system actuators based on the sensor data collected by the sensor-data collection task. For example, this task controls the motor torque to ensure that the driving motor maintains a constant velocity. The communication task transmits the data collected to the computer. In this experiment, the Bluetooth wireless communication is used for the communication. Data transmission via Bluetooth is at a rate of 115,200 bits per second (bps) to allow the computer to monitor data transmission from the RoboCar.

4.1.4 Assumed Fault and Experimental Procedure

For our experiments, we designed a specific fault event subject to which the experiment was conducted. Throughout this research, single-fault was assumed. As the RoboCar is traversing a circular track in the clockwise direction, its motion is obstructed by hand, bringing the RoboCar to a halt. We take the halt of the RoboCar as our fault event in this experi-

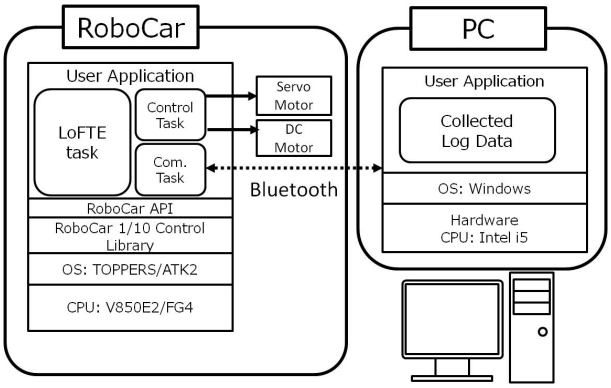


Figure 6: Schematic depiction of the experimental system.

ment.

At the system-design stage, we anticipate the causes of the RoboCar halt and prepare the fault tree diagram as shown in Figure 7. The shaded events indicate the causes of the fault assumed in our experiment.

We assigned the code for all events of the fault tree according to the level. The event of level 0, car halt, was assigned code 0x01. Three events of Level 1, circuit protection by overheat, loss of power supply, and hardware fault were assigned code, 0x10, 0x11, and 0x12, respectively. The events from Level 2-4 were assigned codes in the same manner. The codes of shaded events of Figure 7 and the related log data are listed in the Table 1.

In our experiment we configured the overheat protection function to go into operation at a temperature of 80° C. In our experiments we measured the velocity of both wheels powered by the driving motor, the current, and the FET temperature to record the status of RoboCar.

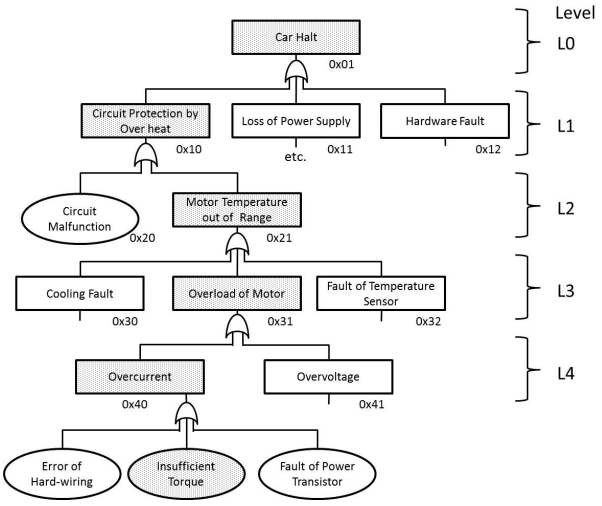


Figure 7: FT diagram for the case of RoboCar halt.

4.1.5 Experimental Results

Table 2 shows the experimental results. The RoboCar's motion was obstructed at 3'59"7 after measurements began. The driving motor continued to operate after this time, but stopped 10 seconds later.

Table 1: Assigned codes for the case of RoboCar halt.

Event	Log Data	Code
Car Halt	Wheel Speed	0x 01
Circuit Protection by Overheat	Flag of Circuit Protection	0x 10
Out of Range of FET Temp.	FET Temperature	0x 21
Motor Overload	Flag of Motor Overload	0x 31
Overcurrent	Motor Current	0x 40

In the interval prior to 3'59"7, the drive velocity of the RoboCar (which is moving clockwise) were approximately 0.8 m/sec for the left wheel and 0.6 m/sec for the right wheel. At the time the car reached the obstruction, the both wheels velocity fell to 0.3 m/sec. At 4'00"4, the left wheel velocity jumped to 1.4 m/sec, while the right wheel velocity fell to 0.0 m/sec which shows the left wheel was locked. At 4'09"7, the driving motor halts and the velocity of both wheels fell to 0.0 m/sec.

Looking at the current of RoboCar, the current lied in the range from 1.0 A to 4.0 A in the interval prior to 3'59"7, but it jumped to a maximum value of 7.7A after encountering the obstruction. At 4'00"4, the current rose up to 10.2A, which showed the overcurrent.

In terms of the temporal variation of the temperature, we saw that the temperature remained roughly constant until 3'59"7, but it jumped to 80.2 ℃ at 4'09"4 after encountering the obstruction. The driving motor halted at 4'09"7, and the temperature remained high.

Regarding the event codes, 0x40 and 0x31 were detected at 4'00"04, which corresponded to overcurrent and overload of motor. In addition, 0x21, 0x10 were detected at 4'09"04, which corresponded to motor temperature out of range, circuit protection by overheat. 0x01 was detected at 4'09"07, which correspond to car halt.

In addition, we saw that the execution time for logging increased after the detection of the codes. The execution time was 4 μsec at 3'59"7. The execution time jumped to 124 μsec at 4'00"4, and it increased with the number of the detected codes.

Table 2: Detected events induced by RoboCar halt.

Event	Time	wheel Speed (Left) [m/sec]	wheel Speed (Right) [m/sec]	Current [A]	FET Temp. [℃]	Detected Event Code	Execution Time for Logging [μsec]
Normal	0'00 "0 – 3'59"6	0.8	0.6	1.0~4.0	49.5	-	4
Drive Obstruction	3'59"7	0.3	0.3	7.7	50.5	-	4
Overcurrent and Overload	4'00"4	1.4	0.0	10.2	52.2	0x40, 0x31	124
Out of Range of FET Temp.	4'09"4	1.4	0.0	7.6	80.2	0x40, 0x31 0x21, 0x10	139
Car Halt	4'09"7	0.0	0.0	9.9	79.9	0x40, 0x31 0x21, 0x10 0x01	149

4.2 Experiment with Wiper Control System

In the second experiment, we used a wiper control system as an experimental device.

4.2.1 Experimental System

Figure 8 shows a structural diagram of the wiper control system. The device is equipped with CPU board(KED-SH101[13]), a servo motor and a monitoring computer. They are connected with RC232C and RS485 network. The servo motor may be controlled from KED-SH101 via networks.

In this experiment, we used TOPPERS/JSP [14], a real-time kernel designed for embedded systems.

Three tasks were implemented as TOPPERS/JSP applications (LoFTE task, a control task, and a communication task), and the system was realized by periodically executing these tasks. LoFTE task was executed every 700ms. LoFTE task collects data from the servo motor, and the data are sent to the monitoring computer.

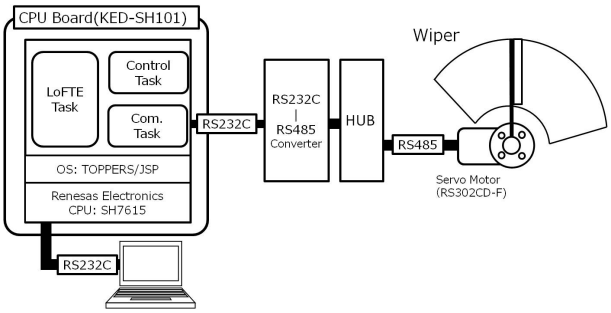


Figure 8: Wiper control system.

4.2.2 Assumed Fault and Experimental Procedure

For our experiments, we designed a communication fault event subject to which the experiment was conducted. As the wiper is swinging back and forth, the communication is interrupted intentionally by setting the disable flag of communication, bringing no response from the servo motor.

At the system-design stage, we anticipate the causes of the wiper halt and prepare the fault tree diagram as shown in Figure 9. The shaded events indicate the causes of the fault assumed in our experiment.

We assigned the codes for all events of the fault tree according to the level. The event of level 0, wiper halt, was assigned code 0x01. Three events of Level 1, circuit protection by overheat, communication fault, and hardware fault were assigned code, 0x10, 0x11, and 0x12, respectively. In Table 3, the codes of shaded events in Figure 9 are depicted.

4.2.3 Experimental Results

Table 4 shows the experimental results. The wiper was started to swing at 0'07"18 after measurements began. The communication was obstructed at 0'16"47 after measurements began. The wiper halted at 0'17"07.

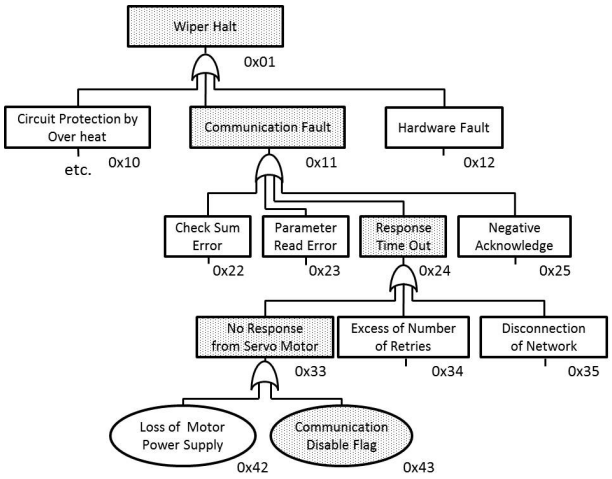


Figure 9: Assumed fault in wiper control system.

Table 3: Assigned codes for the case of wiper halt.

Event	Code
Wiper Halt	0x 01
Communication Fault	0x 11
Response Time Out	0x 24
No Response from Servo Motor	0x 33
Communication Disable Flag	0x 43

In the interval prior to 0'15"79, the register value of the communication register was 0x30 (00110000), which showed transmit enable bit and receive enable bit were set, and the communication was enabled. At 0'16"47, the register value was changed to 0x00, which showed that the communication was disabled. At 0'17"07, the number of retry for the communication reached 6, which showed the maximum number of retry was exceeded.

Regarding the event codes, 0x43 and 0x33 were detected at 0'16"47, which corresponded to disable flag of communication and no response from servo motor. In addition, 0x24, 0x11, and 0x01 were detected at 0'17"07, which corresponded to response timed out, communication fault, and wiper halt, respectively.

Since the time resolution of TOPPERS/JSP kernel is 1 msec, we could not see that the execution time increased with the number of the detected codes.

5 DISCUSSIONS

Our experiments confirmed that the LoFTE method was capable of logging the sequences of faulty events up to the top-level event. The sequences of faulty events are useful information for identifying the cause of the fault.

In the first experiment using RoboCar, we see first that over current and overload of the motor can be detected, then the events on the path of the fault tree can be detected, and finally RoboCar halt (the top-level event) can be detected.

Table 4: Detected events induced by wiper halt.

	Time	Register Value	Num. of Retry	Detected Event Code	Execution Time for Logging [msec]
No Operation	0'00 "00 – 0'7"18	0x30	0	-	1
Normal Operation	0'07"19 – 0'15"79	0x30	0	-	1
Communication Fault	0'16"47	0x00	0	0x43, 0x33	1
Wiper Halt	0'17"07	0x00	6	0x43, 0x33 0x24 0x11, 0x01	2

In the second experiment using a wiper control system, we see that the events on the path of the fault tree can be detected, and finally the wiper halt (the top-level event) can be detected.

Regarding the load of LoFTE tasks, the maximum execution time was estimated 149 μsec in the first experiment. It was fewer than 1% of that of the control cycle period, 100 msec. However, more important thing for real-time system is that overhead is predictable [15]. Since the execution time for logging increased with the number of the detected codes, the load of LOFTE task is predictable by heights of fault trees.

Since we assume a single fault at a time in this study, the order of logged events can be straightforward on the path of the fault tree. However, applications to practical products must consider multiple faults. This can cause the fault trees to grow to an unwieldy size.

Regarding the unexpected faults such as not predefined faults, they cannot be detected in this scheme basically. The unexpected faults do not belong explicitly to any fault trees. However, if unexpected faults had impact on system failure, they would have implicit edge to some fault trees. Such faults are considered to be detectable as other faults at higher level.

On the other hand, the quality of FTA depends on the skill of designers, and constructing FTA of complex system requires long time.

For this reason, our future work will be to access data set for logging through remote networks. We would be able to access real-time controllers by wireless communication, and modify data set for logging about unexpected faults depending on the designer's request. Fault diagnosis of real-time controllers through remote networks would be available. After detecting the faulty device or software, the system would be fixed at service stations or by wireless network. This direction will reduce the problem of designing complete fault trees of complex systems prior to the market release.

6 CONCLUSIONS

In this study, we proposed the LoFTE method for log data collection of faults based on fault tree expansion. We then conducted experiments involving a miniature car and a wiper control system.

The results of our experiments indicate that the LoFTE method is capable of logging the sequences of faulty events. It is useful for identifying the cause of the fault. The load of logging task is predictable by fault trees.

Regarding our future work, it will be to control the logging of fault trees through remote network.

ACKNOWLEDGMENTS

We extend our deepest gratitude to Associate Professor Shinya Honda and the members of the Center for Embedded Computing Systems at Nagoya University (NCES), who provided extensive support and resources regarding TOPPERS/ATK2 during the completion of this work.

We are also particularly grateful for the assistance and support given by Shogo Fukuoka, Hiroki Kitagawa and Kazuhiro Tsujita.

REFERENCES

[1] N. Leveson, Safeware: System Safety and Computers, ACM, pp. 305-313(1995).

[2] J. Shaeuffele, and T. Zurawka, Automotive Software Engineering - Principles, Processes, Methods and Tools, SAE International, pp. 118-125(2005).

[3] Robert Bosch GmbH, Bosch Automotive Handbook, John Wiley & Sons Inc. (2011).

[4] Toyota Motor Corp., Toyota Prius New Model Reference and Repair Manual, Parts Number NM12B1J, pp. IN-36-38 (2010).

[5] S. Fukuoka, et al. ,Scheduling for Logging of Realtime Control Systems, the 77th National Convention of IPSJ, No. 1, pp. 59-60(2014).

[6] M. Takahashi, A Study of Fault Tree Analysis for Embedded Software, IPSJ Technical Reports (JPN), Vol.2013-SE-182 No.24, pp. 1-8(2013).

[7] ZMP Inc., RoboCar 1/10 for AP (Automotive Platform), accessed June 13(2015). https://www.zmp.co.jp/products/robocar-110_package_option#ap

[8] Renesas Electronics, Data Sheet V850E2/FG4 32-bit Single-Chip Microcontroller(2013), accessed June 13(2015). http://documentation.renesas.com/doc/ DocumentServer/R01DS0139ED0100_FG4.pdf

[9] Robert Bosch GmbH, CAN Specification Version 2.0(1991), accessed June 13(2015). http://www.kvaser.com/software/7330130980914/V1/ can2spec.pdf

[10] TOPPERS Project, TOPPERS/ATK2, accessed June 13(2015). https://www.toppers.jp/en/atk2.html

[11] AUTOSAR, GbR., AUTOSAR—Technical Overview V2. 0.1 (2006).

[12] H. Kitagawa, et al., Logging for Fault Diagnosis of Realtime Control System, Proceedings of Workshop on Informatics 2014, pp. 158-164(2014).

[13] Kyoei Electronics, KED-SH101, accessed June 13(2015). http://www.kyoei-ele.com/products/index.php/prod/ info/28/8_13

[14] TOPPERS Project, TOPPERS/JSP kernel, accessed June 13(2015). http://www.toppers.jp/en/jsp-kernel.html

[15] P. Dodd and C. Ravishankar, Monitoring and Debugging Distributed Real-time Programs, Software: Practice and Experience 22.10 pp. 863-877 (1992).

Session 5:
Sensor Applications
(Chair : Yoshia Saito)

Evaluation of an Unconscious Participatory Sensing System with iOS Devices

Takamasa Mizukami†, Katsuhiro Naito‡, Chiaki Doi*, Ken Ohta*,
Hiroshi Inamura*, Takaaki Hishida‡, and Tadanori Mizuno‡

†Graduate School of Business Administration and Computer Science, Aichi Institute of Technology, Japan
‡Faculty of Information Science, Aichi Institute of Technology, Japan
* Research Laboratories, NTT DOCOMO, Inc., Japan
naito@pluslab.org hishida@aitech.ac.jp mizuno@mizulab.net

Abstract - This paper describes a prototype implementation of an unconscious participatory sensing system and evaluates its performance about a measurement process. The proposed system consists of various beacon devices and smartphone devices. When a beacon device requires a communication function, it tries to find a neighboring smartphone device that implements a special application, and requests the relay of measurement information to a management server. Hence, wide-area data collection is feasible without the conscious participation of the smartphone owners in the measurement process. Additionally, we achieve low power consumption in the smartphone application by employing iBeacon, which is a short-range recognition mechanism for BLE. The iBeacon function can trigger the unique dedicated application for background processing even if the application has been suspended because of limitations on background processing in a mobile OS. Therefore, the above processes can be performed without the participants' interaction, and the participants are not required to recognize the measurement operation. In field tests, Raspberry Pi is used on the beacon device side, and the smartphone application side is realized using iOS. We examine the field test using actual sensors and evaluate the performance of the measurement process in the proposed system. The evaluation result shows that the proposed system works on practical devices and collects information unconsciously.

Keywords: Unconscious participatory sensing, iBeacon, BLE, Smartphone device, Beacon device.

1 INTRODUCTION

Sensor networks have received considerable attention to collect various information [1]. Moreover, participation sensing systems, where smartphones measure with high-performance sensors, have been studied [2]–[4]. Conventional participatory sensing systems usually assume that participants join a sensing network voluntarily, and require numerous voluntary participants to collect information. Those participants who accept the demand, perform the requested sensing job, and report measurement information to the system [5]. The participatory sensing process thus realizes flexible and intelligent measurements.

The types of measurement information collected in participatory sensing systems are classified as abstract information, which is evaluated by the participants [6]–[8], and quantitative information, which is measured by sensors [9]–[11]. Measurements of abstract information are difficult to acquire by sensors. Therefore, participants should join the sensing process voluntarily to realize effective participatory sensing systems. Hence, many studies have actively investigated techniques that incentivize participants to join the sensing process [12], [13]. Some researchers have attempted to handle information from participatory sensing and social media in a composite manner [14].

Participatory sensing systems for quantitative information also require participants' interaction behavior, such as checking measurement requests, moving to a measurement location, launching the application, and reporting measurement information. Therefore, participants in conventional systems take part in measurement operations consciously, even if they use sensors to acquire the required quantitative information. For the above reasons, conventional participatory sensing systems currently attract early adopters who have an interest in the new service.

Some studies employ smartphones as a communication device and special measurement devices to achieve accurate measurement by the same type of sensors and under the same implementation conditions [15]. The accuracy of measurement information given by acceleration sensors, magnetism sensors, etc. is generally stable, even if the sensor type is different, because the implementation conditions are not liable to cause a difference in accuracy. However, the accuracy of measurement information given by temperature sensors, illumination sensors, etc. is expected to vary according to the precision of the sensor and implementation environment, because different smartphones have different types of sensors, and the implementation condition of the sensors is also different. Therefore, different measurement values are obtained with different smartphones, even when the measurement is conducted in the same environment. As a result, the built-in sensors in smartphones are not sufficient to acquire accurate measurement information in real situations.

It has been suggested that processes can be carried out without the participants' interaction [16]. This method aims to take continuous environmental measurements at a specific place with a beacon device, with the participants' smartphones cooperating and realizing the desired intelligence. Moreover, the iBeacon [17] function can trigger unique dedicated applications for background processing, even if the application has been suspended because of limitations on background processing in a mobile OS. Therefore, the above processes can be performed without the participants' conscious interaction, and the participant need not recognize the measurement op-

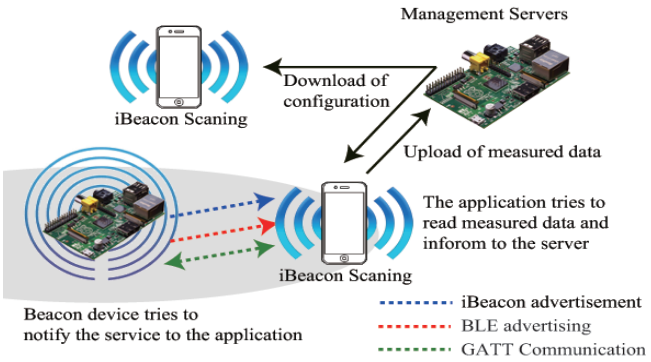


Figure 1: Unconsciousness participation sensing system

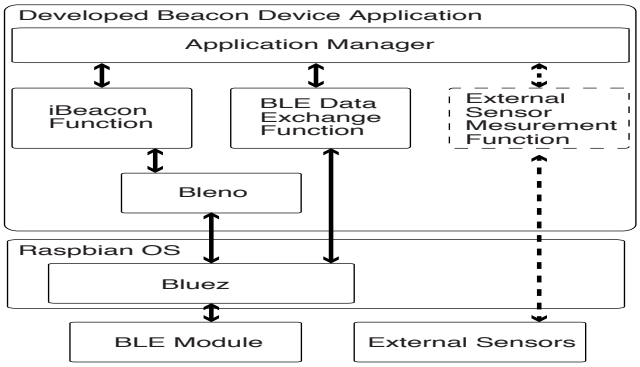


Figure 2: Implementation model of a beacon device

eration. Moreover, participants are not conscious of the cost of the sensing act, and it is predicted that people will be more likely to participate in such a sensing system. This paper assesses the efficiency of the suggested system, and performs a field test using the actual sensors to examine its performance. The proposed system uses an implementation of a smartphone OS, namely iOS, for the field test. The results clearly show that unconscious participatory sensing is suitable for collecting information.

2 THE PROPOSED UNCONSCIOUS PARTICIPATORY SENSING SYSTEM

2.1 Overview of Unconscious Participatory Sensing System

Sensor networks can be easily developed using low-power sensors and cost-effective short-range communication equipment. However, long-range communication devices such as cellular phones are needed to collect information from sensor networks. Participatory sensing systems collect wide-ranging information at low cost. Moreover, various applications have been developed in the last decade. The proposed system can collect widespread information that allows a large number of smartphones to cooperate with a large number of beacon devices with a proximity distance communication facility and a sensing function. The beacon device does not use a specific smartphone. Instead, a smartphone cooperates with a beacon device that is present in the local neighborhood. Moreover, a smartphone reports the data that the beacon device has observed. It is difficult for this proposed system to acquire information in real time at specific time. Therefore, it's suitable for the regular observation is required service. And it assumes the existence of a large number of beacon devices and a large number of smartphones installed in the scanning application. The difference from conventional research is that the beacon device requests a neighborhood smartphone device to relay information from the beacon device measurements taken with built-in sensors, rather than to request a specific smartphone device. Moreover, participatory sensing is often have been discussed motivation of participants. However, the proposed system is not aware of the motivation. Because it handles all processing in a background. It is possible that this proposed

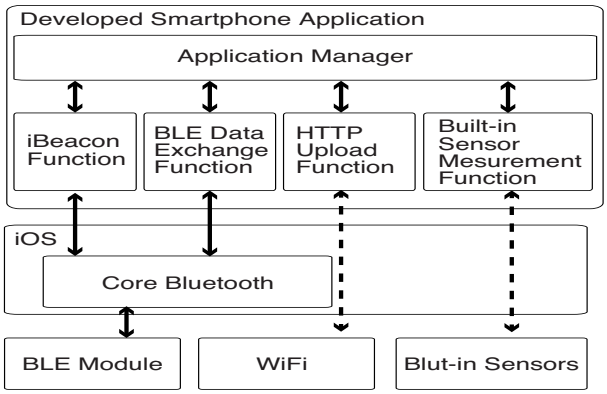


Figure 3: Implementation model of iOS

system is operating at low power consumption by using the technology of the BLE and iBeacon[18], [19]. As mentioned above, we change the viewpoint to ask for sensing and data collection. This change in requesting device gives the following benefits.

- Measurements can be performed with the same sensor.
- A specific place can be observed continuously.
- Preparation of a communication line for an individual beacon device is unnecessary.
- Privacy is maintained as participants' positions are not revealed.
- Communication of sensing requests from the server is unnecessary.
- Participants' interaction behavior for sensing is unnecessary.

Fig. 1 shows an overview of the unconscious participatory sensing system.

2.2 Structure of Unconscious Participatory Sensing System

The proposed system consists of a management server for the management of beacon devices and measurement infor-

mation, a scanning application in the smartphone OS to search for beacon devices and acquire measurement information with built-in sensors, and detectable beacon devices for beacon announcements and measurements with sensors. These components have the following functions.

- Beacon device
  A beacon device has a sensor and the iBeacon function. The main functions of the beacon devices are to trigger a scanning application installed on neighboring smartphones using the iBeacon function, and to allow the application to detect the beacon devices themselves. The application determines the legitimacy of the beacon device by evaluating the hash value after it has detected the beacon device, and starts dedicated operations according to the beacon device's configuration. The proposed system uses built-in sensors in the smartphone and sensors in the beacon devices because measurement information may be affected by differences in implementation conditions or sensor specifications. Therefore, the application transfers measurement information from a beacon device using BLE when the sensors on the beacon device are used. The beacon device performs measurements depending on predefined rules. For example, it starts the iBeacon function to trigger a neighborhood smartphone device after performing duplicate measurement operations.

- Scanning application
  The functions of the scanning application can be roughly classified into searching for a beacon device, acquiring measurement information from the beacon device, taking measurements, and sending measurement information to the management server. The acquired data is transferred to the server in real time. Generally, it is not preferable for a smartphone to constantly search for a beacon device, because this would require excessive electricity consumption. The power consumption of searching for beacon devices generally cannot be ignored if the smartphone wishes to maintain long-time operations. The proposed system employs the iBeacon function to search for beacon devices with low power consumption.

- Management server
  The functions of the management server are information management for each beacon device and the storage of measurement information. iBeacon uses a UUID, major value, and minor value to identify each beacon device. Therefore, the management server should handle these parameters to identify beacon devices and determine suitable measurement rules. Additionally, the server should store measurement information from the scanning application.

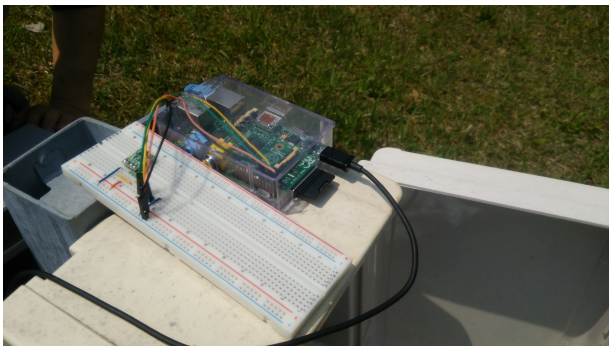


Figure 4: Beacon device

3 IMPLEMENTATION OF UNCONSCIOUS PARTICIPATORY SENSING SYSTEM

3.1 Beacon Device

We have implemented a beacon device for the proposed method using Raspberry Pi, a single-board microcomputer based on ARM. Fig. 2 shows the design of the beacon device. We have been developing a beacon device consisting of an application manager and the iBeacon function, with a BLE data exchange function and outside sensor measurements. The prototype implements the iBeacon function and BLE communication. We employ Raspbian OS, the Bluez [20] library for BLE implementation, and the Bleno [21] library for iBeacon implementation. The developed application on Raspbian OS can handle the iBeacon function and the BLE function using these libraries. Moreover, the outside sensor connected a model number ADT7410 to the outside of a breadboard, and could observe temperatures at a fixed point. The measured temperature is then advertised using the BLE library, and the scanning application in the smartphone can acquire the sensor information. We have confirmed that the developed application can be detected by smartphone devices with the iBeacon function, and the smartphone application can communicate with the beacon device via BLE. We have a schedule for implementing the same mechanisms on System-on-a-Chip (SoC) using BLE, because SoC is the best device for realizing a feasible beacon device with low power consumption.

3.2 Scanning Application

The requirements of the scanning application are to support BLE and iBeacon. We developed a scanning application using an iPod Touch (fifth generation, version 8.1.1). Fig. 3 shows the design of the implemented scanning application. We have been developing a scanning application consisting of an application manager and iBeacon, a BLE data exchange function, hypertext transfer protocol (HTTP) upload function, and built-in sensor measurement function. We employ Core Bluetooth for the Bluetooth implementation in the iOS framework. We have confirmed that triggering the dedicated special application for background processing is effective, even if the application has been suspended because of limitations on background processing, in enabling sensing and sensor data to be recognized.

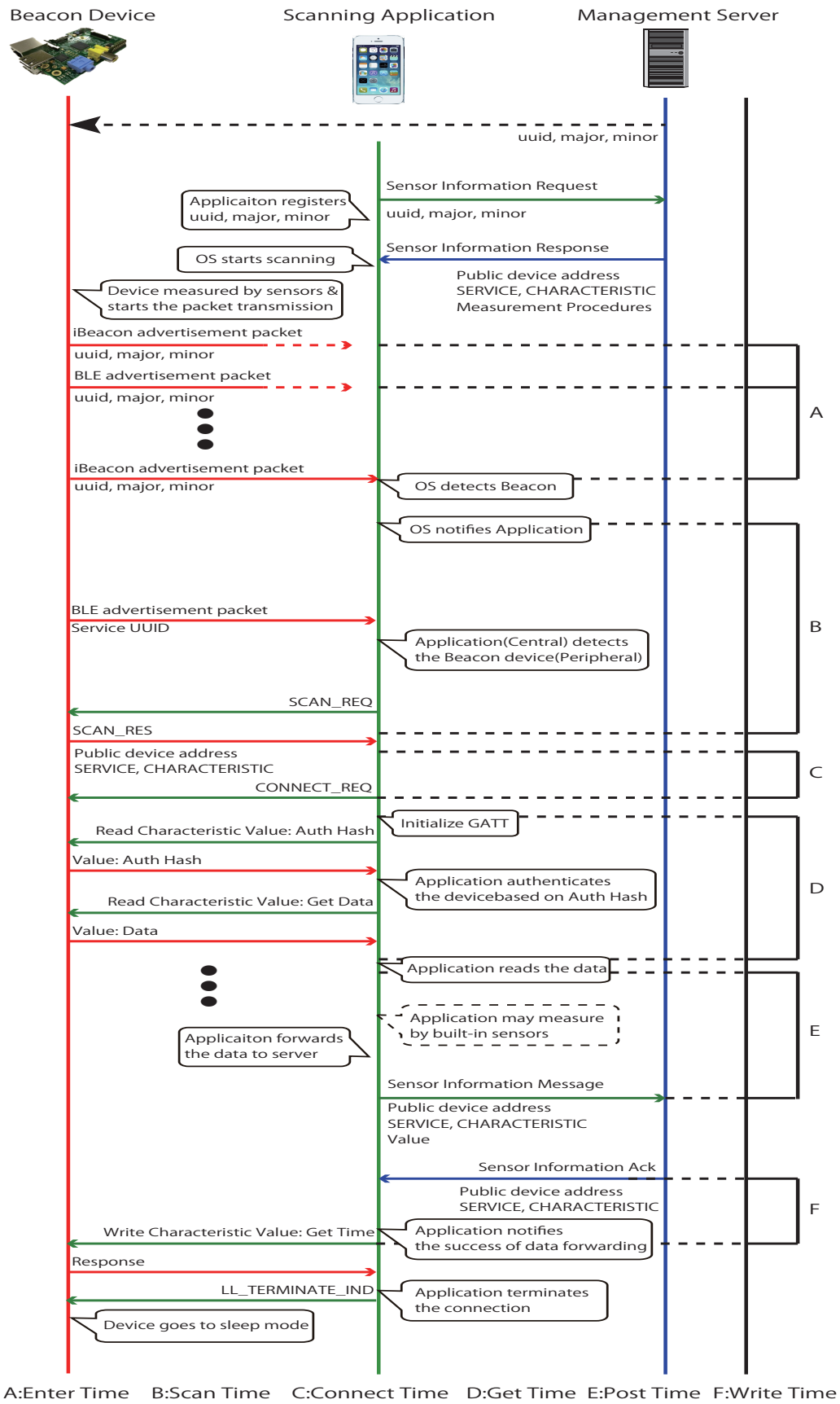


Figure 5: Proposed signaling process

3.3 Management Server

The management server incorporates a beacon identifier, context information management and acquisition, and data storage. We employ the HTTP communication protocol between the management server and the scanning application, because smartphone OSs prepare some APIs for HTTP communication. Therefore, the management server consists of the Apache [22] HTTP server function and MySQL [23] as a database server function. We implemented the management server using Raspberry Pi. Moreover, the management server and scanning application communicate using JavaScript Object Notation (JSON).

4 EXPERIMENTS AND EVALUATION

4.1 Test Objective and Evaluation Points

The scanning application measured the time taken to finish a series of processing tasks on the basis of the signaling process of unconscious participatory sensing. There is a limitation on the iOS side. The time required for processing is approximately 10 s, whereas the time for which a smartphone is within the iBeacon domain and can freely handle applications in the background is limited. Therefore, the processing that can be performed under this limitation must be investigated. This experiment was conducted in the athletic grounds of a university. The Raspberry Pi was used as the beacon device, the scanning application was an iPod touch, and the server was a MacBook Air. Moreover, We prepared for communication environment using mobile Wifi. Fig. 4 shows that the beacon device was set on the ground. The temperature was observed at a fixed point and advertised via BLE. We tested the movements within 15 m that could detect an iBeacon identifier at approximately 20 m from the beacon device. I pocketed iPod touch and repeated the comings and goings detectable area of iBeacon. We conducted this field test 30 times. Moreover, Fig. 5 shows that we evaluated each of the following items in the signaling process.

A Once within range, the time taken for the data collection application to recognize iBeacon and acquire identifier information

B Time required for the scanning application to search for a beacon device by checking a BLE advertisement packet in background processing

C Time required for the scanning application to connect to the detected beacon device once it has been identified

D Time taken before the scanning application obtains the desired measurement information from the beacon device by BLE communication

E Time taken before the scanning application transfers the measurement information to the management server

F Time taken before the scanning application notifies the beacon device that all processing has finished

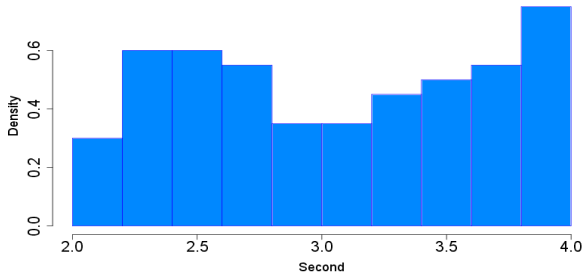


Figure 6: A: Time taken to obtain iBeacon identifier from a beacon in the area

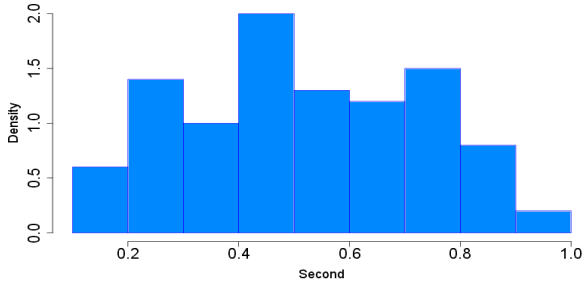


Figure 7: B: Time taken to detect a peripheral

4.2 Experimental Results

Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11 show the field test results for the evaluation of items A–F. Fig. 12 shows the total time required by the scanning application for background processing. As the scanning application moved into the detectable area of iBeacon, the scanning application takes approximately 2.5 $\sim$ 3.0 s to detect identifier information. Moreover, the scanning application searches for a beacon device for 0.4 $\sim$ 0.6 s. The scanning application connects with the beacon device after 0.5 $\sim$ 0.8 s, and takes 0.5 $\sim$ 0.8 s to acquire the sensor data. Transferring the measurement information to the management server takes approximately 0.8 $\sim$ 1.2 s. Finally, the scanning application takes 0.2 $\sim$ 0.4 s to notify the beacon device that all processing has finished. The overall background processing time required by the scanning application approximately 4.25 $\sim$ 4.5 s.

5 APPLICATION EXAMPLES

The prototype that we developed in this report is roughly application to the following applied application.

Applied the example using the beacon device's sensor.

- Discomfort index Heat illness-warning
  Heatstroke is used as necessary information of the temperature and humidity of the present location. The proposed system can warn of potential heat stroke for nearby location by smartphone automatically by implementing a thermometer and a hygrometer to a beacon device.

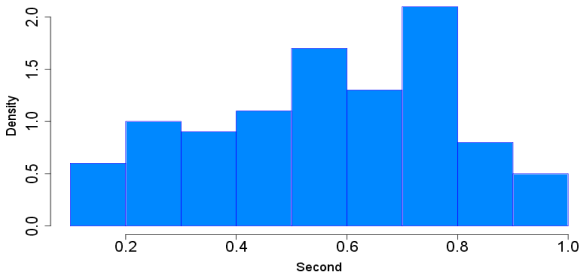


Figure 8: C: Time required to connect to the peripheral

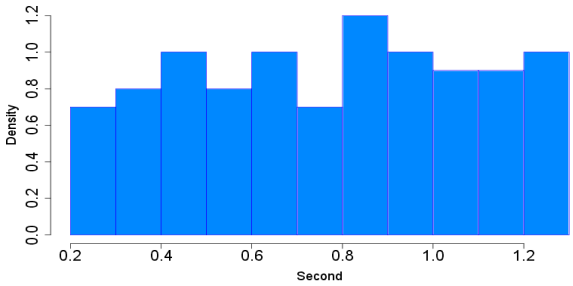


Figure 10: E: Time taken to post sensor data

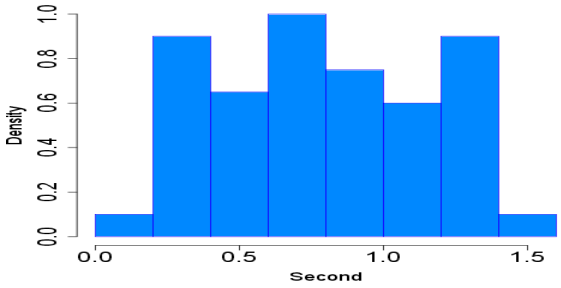


Figure 9: D: Time taken to obtain sensor data

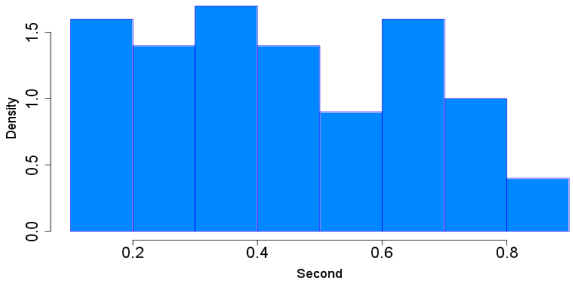


Figure 11: F: Time taken to write to the peripheral

- Amenity
  It is important that facility management understand the environmental condition. The proposed system can learn the proper conditions of the beacon device's location by implementing a thermometer, a hygrometer, a microphone and a illumination sensor to the beacon device.

Applied the example using the built-in sensors in the smartphone

- Congestion degree
  It is not easy to grasp the movement of the person when considering privacy within public facilities. In contrast, the proposed system can monitor precise of the smartphone around the beacon device by using built-in acceleration sensors in the phone. It is possible that the movement of the person around the beacon device quantitatively estimates the amount of congestion by processing the degree of movement in the area.

- Environmental control
  It is important to understand each environment to learn the most suitable use of energy for that environment. The proposed system can gather information of the smartphone owner's location by using the smartphones around the beacon device. The beacon device will use the illumination sensor and thermometer sensor. We think more comfortable environmental control is feasible by using the information presented above.

6 CONCLUSION

This paper has developed a prototype implementation of the unconscious participatory sensing proposed in previous studies, and has assessed the efficiency of such a system using Raspberry Pi and a temperature sensor. A sensing participant operates unconsciously by utilizing the BLE mechanism. We show that it is possible to report sensor data and unconscious operation automatically. Moreover, we think that this helps reduce the participation cost compared with conventional participation sensing. In addition, to perform the field test using an actual machine sensor and assess its efficiency, we examined the performance of the suggested system. The field test used an application based on iOS, which supports iBeacon by default. Using the sleep application function of iBeacon, we confirmed that the scanning application could start the necessary background processing. The application confirmed that temperature information could be acquired from a beacon device during background processing. Moreover, we confirmed that this information could be reported to a server. We found that all processing could be completed within the limited time available. The test results indicate that an average of approximately 8 s was required for processing. This is within the limit of approximately 10 s in which information can be freely processed in the background under iOS. It is thought that this limitation of iOS is to prevent the excessive consumption of the smartphone's electricity resources. Therefore, the scanning application must finish background processing within 10 s. The proposed beacon devices and sensing applications of previous studies were found to be suitable for unconscious sensing in the field tests carried out in this paper.

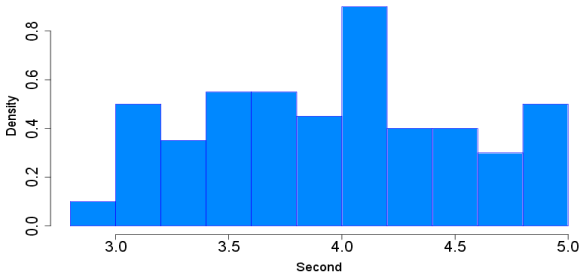


Figure 12: Total processing time

REFERENCES

[1] F. Viani, P. Rocca, G. Oliveri, and A. Massa: Pervasive remote sensing through WSNs, In Antennas and Propagation (EUCAP), 2012 6th European Conference on, pp. 49-50. IEEE, 2012.

[2] N. D. Lane, S. B. Eisenman, M. Musolesi, E. Miluzzo, and A. T. Campbell: Urban sensing systems: opportunistic or participatory?, In Proceedings of the 9th Workshop on Mobile Computing Systems and Applications, HotMobile'08, pp. 11-16, 2008.

[3] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell: A survey of mobile phone sensing, IEEE Communications Magazine, Vol. 48, No. 9, September 2010.

[4] P. Mohan, V. N. Padmanabhan, and R. Ramjee: Nericell: rich monitoring of road and traffic conditions using mobile smartphones, SenSys'08: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, November 2008.

[5] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava: Participatory sensing, Mobile Device Centric Sensor Networks and Applications, In Workshop on World-Sensor-Web (WSW), pp. 117-134, 2006.

[6] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le: Using Humans as Sensors: An Estimation-theoretic Perspective, In IPSN'14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, pp. 35-46, 2014.

[7] E. Niforatos, A. Vourvopoulos, M. Langheinrich, P. Campos, and A. Doria: Atmos: a hybrid crowdsourcing approach to weather estimation, UbiComp'14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, September 2014.

[8] A. H. Lam, Y. Yuan, and D. Wang: An occupant-participatory approach for thermal comfort enhancement and energy conservation in buildings, The 5th International Conference on Future Energy Systems (e-Energy'14), June 2014.

[9] M. Budde, R. E. Masri, T. Riedel, and M. Beigl: Enabling Low-Cost Particulate Matter Measurement for Participatory Sensing Scenarios, In Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia (MUM'13), December 2013.

[10] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden: CarTel: a distributed mobile sensor computing system, In SenSys'06, pp. 125-138, 2006.

[11] F. Zeiger and M. Huber: Demonstration abstract: participatory sensing enabled environmental monitoring in smart cities, The 13th International Symposium on Information Processing in Sensor Networks (IPSN'14), April 2014.

[12] A. Tomasic, J. Zimmerman, A. Steinfeld, and Y. Huang: Motivating Contribution in a Participatory Sensing System via Quid-Pro-Quo, In CSCW'14 Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 979-988, 2014.

[13] D. Zhang, H. Xiong, L. Wang, and G. Chen: CrowdRecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint, UbiComp'14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, September 2014.

[14] M. Demirbas, M. A. Bayir, C. G. Akcora, Y. S. Yilmaz, and H. Ferhatosmanoglu: Crowd-sourced sensing and collaboration using twitter, WOWMOM'10 Proceedings of the 2010 IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1-9, 2010.

[15] L. Li, Y. Zheng, and L. Zhang: Demonstration abstract: PiMi air box: a cost-effective sensor for participatory indoor quality monitoring, IPSN'14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, pp. 327-328, 2014.

[16] T. Mizukami, K. Naito, C. Doi, T. Nakagawa, K. Ohta, H. Inamura, T. Hishida, and T. Mizuno: Fundamental Design for a Beacon Device Based Unconscious Participatory Sensing System, International MultiConference of Engineers and Computer Scientists 2015, Vol. 2 2015.

[17] iBeacon for Developers, https://developer.apple.com/ibeacon/, Retrieved October 2014.

[18] BLUETOOTH SPECIFICATION Version 4.0, https://www.bluetooth.org/ja-jp/specification/adopted-specifications, Retrieved October 2014.

[19] Getting Started with iBeacon, https://developer.apple.com/ibeacon/Getting-Started-with-iBeacon.pdf, Retrieved October 2014.

[20] Official Linux Bluetooth protocol stack http://www.bluez.org, Retrieved October 2014.

[21] A node.js module for implementing BLE (Bluetooth low energy) peripherals, https://github.com/sandeepmistry/bleno, Retrieved October 2014.

[22] Apache, http://www.apache.org, Retrieved October 2014.

[23] MySQL, http://www.mysql.com, Retrieved October 2014.

Gait-based Authentication using Trouser Front-Pocket Sensors

Shinsuke Konno*, Yoshitaka Nakamura**, Yoh Shiraishi**, and Osamu Takahashi**

* National Institute of Technology, Hakodate Collage, Japan
** School of Systems Information Science, Future University Hakodate, Japan
skonno@hakodate-ct.ac.jp
{y-nakamr, siraisi, osamu}@fun.ac.jp

Abstract - Recently, to reduce the inconvenience caused by authentication operations in portable terminals, various authentication methods based on behavior characteristics have been studied. Gait-based authentication is one of them. This authentication method identifies individuals based on walking motions measured by wearable sensors such as acceleration sensors. This study aims to improve the authentication accuracy using trouser front pocket sensors. In this study, we consider two analyses to achieve this goal. First, we investigate the relation between walking motion and gait signals from trouser pocket sensors to extract signals of same-gait motion intervals in different subjects. Next, we verify an authentication method that uses both an acceleration sensor and a gyro sensor to improve the authentication accuracy.

Keywords: gait-based authentication, acceleration sensor, gryro sensor, dynamic time warping, fusion

1 INTRODUCTION

The use of portable terminals such as smartphones has increased in the various situations and can be expected to increase in the future. Accordingly, smartphones and other portable terminals have equipped various personal authentication functions to prevent imposters from misusing. Recently, authentication functions such as pattern locks, which are more difficult for an imposter to break, has been incorporated into the devices. However, there are reports and news items showing that approximately 50% of users do not lock their devices by inconvenient their operations.

Previous studies proposed easier authentication methods by various device operations, such as swinging their terminals. However these methods require conscious action, so they cannot perform authentication in the background.

On the other hand, it is conceivable that individual authentication might be established through daily repeated activities. With such a method, a user can unlock a terminal without conscious operations. Gait-based authentication is one of this type authentications. We think that walking is performed in various situations. If gait authentication was established by sensors on a portable terminal, the inconvenience users feel in individual authentication would be reduced.

We work with multi-modal authentication to improve authentication performance by combining multiple methods in individual authentication [1]. Fernand et al. [2] combined faces and fingerprints to improve accuracy. Zhou et al. [3] combined features of side face and gait using principal component analysis to identify people, and many other researchers have also attempted to improve accuracy using biometric authentication.

However, wearing multiple sensors on various body parts sacrifices convenience, the advantage of gait authentication. For this reason, we adopt a method that combines multiple sensor methods measuring the same body parts using multiple sensors, and a multi-sample method that measures a modality several times to improve performance. It is possible to equip a terminal with multiple sensors, enabling us to authenticate using multiple sensors without imposing a burden on users.

In this study, we use two sensors (a three-axis acceleration sensor and a three-axis gyro sensor) to measure human walking motion. We show that the proposed method, which combines distance information recorded by these two sensors, improves authentication accuracy in comparison with previous studies.

2 RELATED WORK

2.1 Position of sensors

Table 1 summarizes the related work. These studies explored features and authentication methods primarily to improve performance. However, they did not investigate which sensor positions would be acceptable for daily use.

Those studies measured mainly using devices attached on the belt on the middle or side of the waist, and authenticated using measured acceleration signals. This requires using a smartphone case such as a holster for attaching the terminal to the waist. Users might find this unacceptable, because gait authentication then requires them to have the container with them. Consequently, we decided that the trouser front-pocket might be acceptable to users, because they can then have the terminal without using special tools, and we investigated performance improvement in this position. The study in [6] examined this position. This study aims to improve authentication performance in comparison to that previous study.

Table 1: Summary of gait-based authentication work

work	position	Sensor
Mäntyjärviet et al. [4]	belt	acceleration
Gafurov et al. [5]	hip	acceleration
Gafurov et al. [6]	ankle	acceleration
Gafurov et al. [7]	trouser pocket	acceleration
Gracian et al. [8]	belt	acceleration
Derawi et al. [9]	belt	acceleration
Soumik et al. [10]	eight-joints	rotation angle

2.2 Fusion of multiple sensors

Many acceleration-based approaches to gait-based authentication have been explored.

Mäntyjärviet et al. [4] proposed three authentication methods: fast Fourier transform, correlation, and statistical features. Gafurov et al. [5][6][7] studied methods based on acceleration, and made measurements by using acceleration sensors on various parts of subject's bodies. They used a template signal and multiple time-normalized signals, with the acceleration sensor placed in the trouser front pocket [7].

Gracian et al. [8] devised the feature of gait acceleration for user authentication. Derawi et al. [9] proposed a multi-sampling method that authenticated using multiple signals from both templates and inputs. Their method calculated distances of all combinations of templates and inputs with dynamic time warping (DTW). Soumik et al. [10] measured walking motions with eight angle sensors.

To the best of our knowledge, there are no studies on the fusion of multiple sensors placed in a trouser front pocket. To improve authentication accuracy, we propose a method of fused distances based on acceleration and angular velocity placed in a trouser front pocket.

3 PROPOSED METHOD

3.1 Gait recognition and quasi-periodic signal extraction

We attached a sensor unit whose x-, y-, and z-axis detected vertical, sideway, and forward-backward acceleration, respectively, in standing posture. The direction of each axis is shown in Figure 1. Each subject wore a sensor unit attached to a belt with hook and loop fastener. This unit was placed on the front of the left femur area.

Examples of three-axis acceleration and three-axis angular velocity are shown in Figures 2 and 3. During walking, the acceleration and gyro sensors measured similar waveforms repeatedly. These signals are quasi-periodic signals with no equalization of cycles and amplitudes.

The length of a gait cycle is two steps. The gait cycle consists of four periods, two double limb support periods, and two single limb support periods. We walk forward by repeating the four periods. If we extract the gait signals from different walking period for each subject, we may achieve good performance seemingly in authentication. To prevent influence on authentication accuracy by different waveform for each user, we decided to extract their quasi-periodic signals with the same order of the gait periods to all users. For this reason, we conducted a preliminary experiment to investigate the relation between walking motion and six-axis signals. Two force sensors synchronized with the sensor unit were attached to their left toe and heel. Examples of the acceleration along the x-axis and the signal of the force sensor are shown in Figure 4. The graph shows that the time when the acceleration becomes a local maximum is approximately equal to the time when the value of the heel force sensor begins to increase. This result indicates that the time of local maximum of acceleration is the heel landing time.

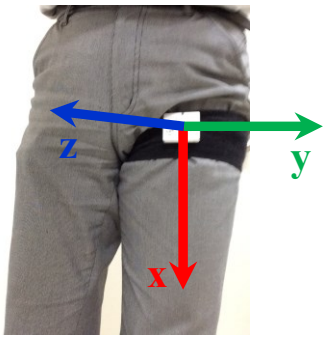


Figure 1: Directions of three axes.

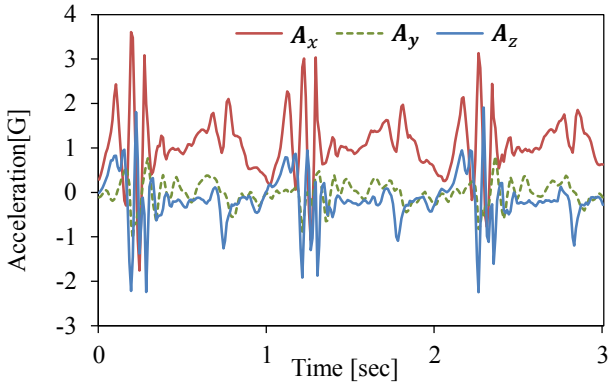


Figure 2: Gait signals from three-axis acceleration sensor.

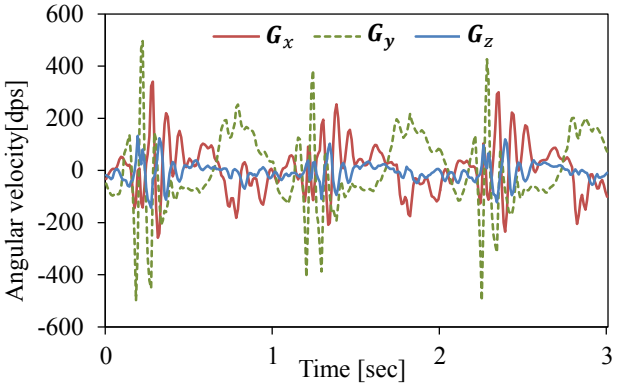


Figure 3: Gait signals from three-axis gyro sensor.

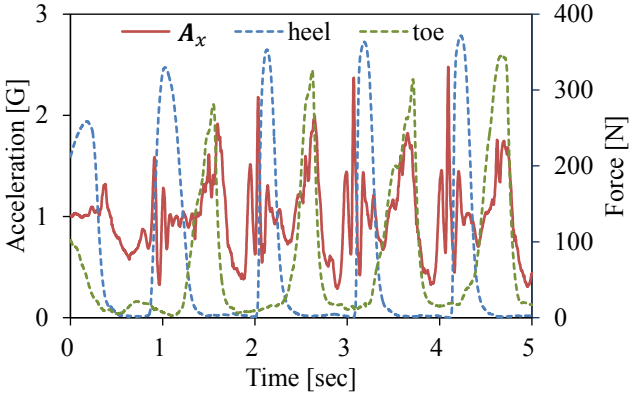


Figure 4: Example of the vertical acceleration signal and force signal.

3.1.1. Walking detection

In this study, we use a threshold in vertical acceleration to detect walking start time based on previous research [7]. Before beginning, all signals were smoothed using a Savitzky–Golay filter [11]. We look for the time t_s when the acceleration is greater than 1.2 G from the start of this quasi-periodic signal extraction method.

3.1.2. Quasi-periodic signal extraction

After walking detection, we extract quasi-periodic signals measuring the period between left-heel landing time. The extraction process with x-axis acceleration $\boldsymbol{A}_x$ is as follows:

1) We search for the maximum time T_0 within two seconds after t_s. We selected T_0 as the start time of cycle $\boldsymbol{C}_0$.
2) To find the end time of $\boldsymbol{C}_0$, we search for all times of local maxima from 0.7 to 1.3 s after T_0 from $\boldsymbol{A}_x$.
3) We extract subsets $\boldsymbol{s}_0$ that are 0.6 s of the signal. T_0 is the middle time of subset $\boldsymbol{s}_0$. In the same way, each $\boldsymbol{t}_1$ is the middle time of subsets $\boldsymbol{S}_1 = \{\boldsymbol{s}_{11}, \boldsymbol{s}_{12}, \boldsymbol{s}_{13} \cdots\}$, which are extracted as 0.6 s signals. We calculate values of the normalized cross correlation (NCC) among $\boldsymbol{s}_0$ and each $\boldsymbol{S}_1$. The middle time of NCC values is decided as the start time T_1 of the next cycle $\boldsymbol{C}_1$. Cycle $\boldsymbol{C}_0$ is between T_0 and T_1 . This is shown in Figure 5.
4) Next, we search for all times of local maxima from 0.7 to 1.3 s after T_1. We extract the subsets of signal from T_1 to each time of the local maxima. The time of minimum distance among $\boldsymbol{C}_0$ and each subset with DTW is decided as the start time T_2 of the next cycle $\boldsymbol{C}_2$. In this calculation, to eliminate the effect of differences in signal length, we divided each distance by the total length of $\boldsymbol{C}_0$ and each $\boldsymbol{S}_2$.
5) After the time T_n of minimum distance is calculated using DTW among $\boldsymbol{C}_{n-1}$ and $\boldsymbol{S}_n$, we begin searching for the next start time T_{n+1} by repeating step 4) .
6) When forward searching is completed, we repeat the process by searching backward at T_0.
7) When we observed the extracted signals, we found that those near the signals of starting to walk had a large distortion as compared with other signals. Based on the result of analysis, the variance of each signal with a large distortion is smaller than the variance of other signals. Hence, we searched for the first distorted signals whose variance was greater than the threshold 0.09. We assumed that the signals used for authentication were signals subsequent to it. Examples of the variance from extracted signals are shown in Figure 6. In this x-axis acceleration, we took the signals to be used for authentication as the cycles after C_0 . We recorded the starting times of extracted cycles, and extracted signals for the other two-axis acceleration and three-axis angular velocity using the same starting time.

Figure 7 shows two extracted signals from the same subject. In Figure 8, the two lines indicate extracted signals from different subjects.

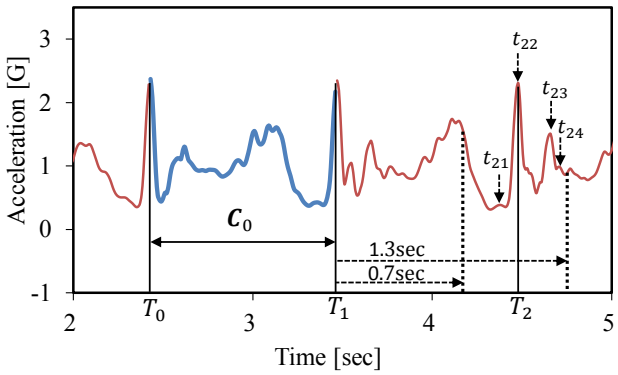


Figure 5: Example of extracted cycle $\boldsymbol{C}_0$ and local maximum.

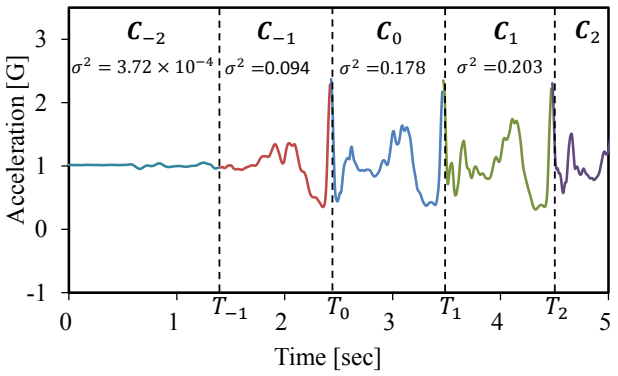


Figure 6: Example of extracted cycles and their variances.

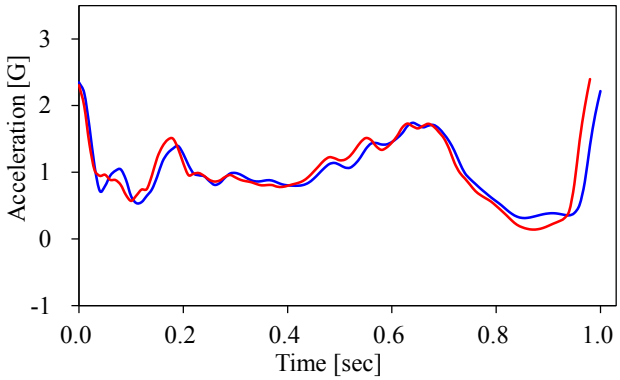


Figure 7: Extracted signals from same subject.

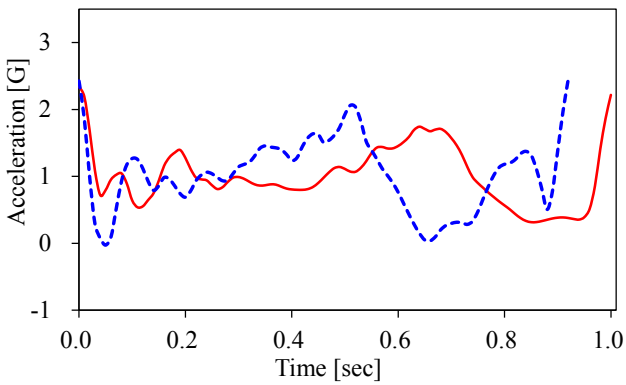


Figure 8: Extracted signals from two subjects.

3.2 Distance calculation methods

We selected DTW that is frequently used for calculating dissimilarity between time series data. Let $\boldsymbol{X} = \{x(i)|i = 1,2,\cdots,m\}$, $\boldsymbol{Y} = \{y(j)|j = 1,2,\cdots,n\}$ be time series data. The DTW distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$DTW(\boldsymbol{X},\boldsymbol{Y}) = f(m,n)$$
$$f(i,j) = min\begin{cases} f(i-1,j-1) + dist(x(i),y(j)) \\ f(i,j-1) + dist(x(i),y(j)) + GP \\ f(i-1,j) + dist(x(i),y(j)) + GP \end{cases}$$
$$f(0,0) = 0$$

where $DTW(\boldsymbol{X},\boldsymbol{Y})$ is the DTW distance, m and n are the number of lengths in signals $\boldsymbol{X}$ and $\boldsymbol{Y}$, and GP is a gap penalty in the case of non-linear extension. We adopted the different distance calculation method for each sensor. The distance calculation function is substituted into $dist(x(i),y(j))$ corresponding to the type of sensors. Next, to adapt the differences of signal length to differences of walking speed, Normalized distance $D(\boldsymbol{X},\boldsymbol{Y})$ is calculated as

$$D(\boldsymbol{X},\boldsymbol{Y}) = \frac{DTW(\boldsymbol{X},\boldsymbol{Y})}{m+n}$$

In the multi-sample case, we used the median as the distance. Let $\boldsymbol{Y} = \{\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_k,\cdots,\boldsymbol{Y}_p\}$ be multiple template signals. This distance was calculated as

$$D(\boldsymbol{X},\boldsymbol{Y}) = \underset{k}{\text{median}}\big(D(\boldsymbol{X},\boldsymbol{Y}_k)\big)$$

where $D(\boldsymbol{X},\boldsymbol{Y}_k)$ is the normalized distance between an input signal and k template signals of multiple template signals.

3.2.1. Angular velocity distance

It is known that angular velocity does not depend on distance from the center of rotation. We calculate the absolute distance between the input signal and template signals. Even if signals of the same subject are selected, they do not correspond to the amplitude value from a difference in walking speed. To reduce differences between signals of the same subject, we normalized the signals by dividing the amplitude of each time by specific values. We adopted the method of normalization that divides amplitude of signal by the root mean square (RMS). The reason for using RMS for normalization is that it provided the best accuracy among some normalized methods in a preliminary experiment.

Let $\boldsymbol{g}_{in}^{q} = (g_{in}^{q}(1), g_{in}^{q}(2), \cdots, g_{in}^{q}(i), \cdots, g_{in}^{q}(m))$be the q-axis input angular velocity signal, and let $\boldsymbol{g}_{t\,k}^{q}\left(g_{t\,k}^{q}(1), g_{t\,k}^{q}(2), \cdots, g_{t\,k}^{q}(j), \cdots, g_{t\,k}^{q}(n)\right)$ be the q-axis k template angular velocity signal. We calculate the difference of the composed angular velocity between the i^{th} amplitude of a q-axis input angular velocity signal and the j^{th} amplitude of a q-axis k template angular velocity signal by the absolute distance as

$$dist(g_{in}^{q}(i), g_{t\,k}^{q}(j)) = \left|g_{in}^{q}(i) - g_{t\,k}^{q}(j)\right|$$

3.2.2. Acceleration distance

When measuring circular motion, it is known that acceleration depends on the distance from the center of rotation. If different amplitude normalizations are applied to each axis acceleration, they are compressed at different ratios at the same time. As a result, when the normalized accelerations of the three axes at the same time were combined as a vector, the direction of the vector was changed before normalization. This problem was caused by comparing it with the values of acceleration. Hence, we compared it with the direction of three-axis acceleration between the input and the template acceleration signals [12]. Let $\boldsymbol{a}_{in}(i) = \left(a_{in}^{x}(i), a_{in}^{y}(i), a_{in}^{z}(i)\right)$ be the i^{th} input acceleration vector of an input signal, and let be $\boldsymbol{a}_{t\,k}(j) = \left(a_{t\,k}^{x}(j), a_{t\,k}^{y}(j), a_{t\,k}^{z}(j)\right)$ be the j^{th} template acceleration vector of a k template signal. We calculate the difference of direction between the i^{th} input acceleration vector and j^{th} k template acceleration vector as

$$dist\big(\boldsymbol{a}_{in}(i), \boldsymbol{a}_{t\,k}(j)\big) = \arccos\frac{\langle \boldsymbol{a}_{in}(i), \boldsymbol{a}_{t\,k}(j)\rangle}{\|\boldsymbol{a}_{in}(i)\|\|\boldsymbol{a}_{t\,k}(j)\|}$$

To compare this three-axis composite method with others, authentication accuracy of each axis acceleration was calculated based on same distance calculation method for angular velocity.

3.3 Distance fusion

To eliminate subject dependency, we subtracted the average distance from the distance before fusion. This average distance was calculated between a subject's template signal $\boldsymbol{Y}$ and the same subject's training data $\boldsymbol{\gamma}$ except his or her template signal $\boldsymbol{Y}$. The normalized distance is calculated by subtracting the average distance from the distance calculated by DTW between an input signal and the template signals as

$$D_s(\boldsymbol{X},\boldsymbol{Y}) = D(\boldsymbol{X},\boldsymbol{Y}) - \overline{D(\boldsymbol{\gamma},\boldsymbol{Y})}$$

Finally, we calculated the fused distances D_f as

$$D_f = f(D_s(\boldsymbol{a}_{in},\boldsymbol{a}_t), D_s(\boldsymbol{g}_{in}^{x},\boldsymbol{g}_t^{x}), D_s\big(\boldsymbol{g}_{in}^{y},\boldsymbol{g}_t^{y}\big), D_s(\boldsymbol{g}_{in}^{z},\boldsymbol{g}_t^{z}))$$

where $f(\,)$ is a function of fusion which combines the distances.

In this study, we consider four rules for fusing distances for authentication (1) Addition without weight coefficients (denoted as Sum), (2) Linear logistic regression (denoted as LLR), (3) Support vector machine (SVM) with linear kernel (Linear), (4) SVM with a radial basis function kernel (RBF)

In this study, we obtained too many negative instances as compared with positive instances. It is well known that SVM performs poorly in this case. Hence, we applied the synthetic minority over-sampling technique [13] to adjust the number of these instances.

4 EXPERIMENT

4.1 Dataset

Data was collected from 50 subjects, ranging in age from 18 to 21 years old. We instructed the subjects to walk at their normal walking speeds. When the measurement began, the subjects remained stationary for a few seconds. After that, they walked a specified distance once. The measurement course is a flat and straight indoor passageway. The subjects did not use a clock or metronome to measure their walking speed. We set the sampling frequency of the sensor unit to 1,000 Hz. To equalize the performance of the smartphone's sensors, we changed the sampling frequency from 1,000 to 100 Hz by thinning out.

4.2 Experimental setting

We obtained 30 signals of each axis acceleration and 30 signals of each axis angular velocity from every subject. We divided the signals into five groups and performed five-fold cross-validation. To generate a fusion model, we used four groups as training data, and one group as test data. We calculated the distances between all of the training signals of all subjects. The distances between the same subjects are positive instances, and the distances between different subjects are treated as negative instances. The overall accuracies were calculated with common thresholds to each classifier in each fusion rule.

Template signals used for calculating distance include six signals, because the number of template signals is equal to the number of template signals of the previous study [7]. The manner of selecting templates from training data was to select six sequential signals from 24 signals. However, when some of the sequential six signals were selected as test data by cross-validation, we selected the signals in sequence from the nearest start time in the training data.

4.3 Experimental result

We evaluated accuracy by equal error rate (EER). The EER is the value when the false acceptance rate (FAR) and the false rejection rate (FRR) are the same. For comparison purposes, we calculated EERs four combinations of each method, and previous work distance calculation method [7].

We summarized the EERs in Tables 2 and 3. By comparing with each combinations, we can find that both multi-sensor and multi-sample are effective for accuracy improvement. The minimum EER (the best result) was 1.0%, which was achieved by the proposed multi-sensor multi-sample method with two SVMs.

Figures 10 show the receiver operating characteristics (ROC) curves for each authentication combinations. From the ROC curves, proposed method whici is combination method with multi-sensor and multi-sample shows the best performance, because most of multi-sample with RBF line is plotted in the lower error rate area. The best EER from previous work method [6] to each axis signal for this dataset was 7.8%.

5 CONCLUSION

This paper describes an authentication method using multi-sampling and multi-sensors to improve the accuracy of gait-based authentication.

First, we observed the relation among the steps and six-axis signals in order to extract the quasi-periodic signals generated by walking motion of the same phase order in all subjects. These findings show that it is possible to divide into quasi-periodic signals by extracting x-axis acceleration from local maxima to local maxima.

We evaluated the proposed method with 50 subjects. The best EER performance was 1.0%, which was achieved by the multi-sensor multi-sample method using SVM. These results indicate that the combination of multi-sensor and multi-sample is useful for gait-based authentication. Furthermore, proposed method leads to better results than the conventional method.

Table 2: Uni-sensor EERs [%].

	Uni-sensor uni-sample authentication	Uni-sensor multi-sample authentication
$\boldsymbol{a}^x$	8.8	4.5
$\boldsymbol{a}^y$	5.3	2.2
$\boldsymbol{a}^z$	4.6	2.2
$\boldsymbol{g}^x$	6.6	2.4
$\boldsymbol{g}^y$	8.2	3.1
$\boldsymbol{g}^z$	7.4	3.0

Table 3: Multi-sensor EERs [%]

	Multi-sensor uni-sample authentication	Multi-sensor multi-sample authentication
Sum	1.7	1.2
LLR	1.5	1.1
Linear	1.5	1.0
RBF	1.4	1.0

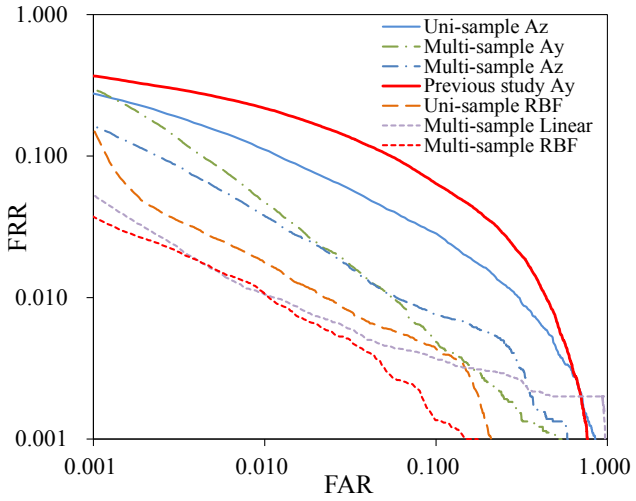


Figure 10: ROC curves of the best methods in each authentication combinations

REFERENCES

[1] A.K. Jain, A. Ross, and S. Prabhakar, An Introduction to Biometric Recognition, IEEE Trans. Circuits and Systems for Video, Vol.14, No.1, pp.4-20 (2004).

[2] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Ortega-Garcia, Dealing with sensor interoperability in multi-biometrics: The UPM experience at the Biosecure multimodal Evaluation 2007, Proc. of SPIE Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, pp.69440J-69440J (2008).

[3] X. Zhou and B. Bhanu, Feature Fusion of Side Face and Gait for Video-based Human Identification, Pattern Recognition, Vol.41, pp.778-795(2008).

[4] J. Mäntyjärvi, M. Lindholm, E. Vildjiounaite, S. Mäkelä, and H.A. Ailisto, Identifying Users of Portable Devices from Gait Pattern with Accelerometers, Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp.973-976 (2005).

[5] D. Gafurov, E. Snekkenes, and T.E. Buvarp, Robustness of Biometric Gait Authentication Against Impersonation Attack, Proc. of the 1st International Workshop on Information Security (IS'06), OnTheMove Federated Conferences (OTM'06), Vol.4277, pp.479-488 (2006).

[6] D. Gafurov, K. Helkala, and T. Sondrol, Biometric Gait Authentication Using Accelerometer Sensor, Jounal of Computers,Vol.1, No.7, pp.51-59(2006).

[7] D. Gafurov, E. Snekkenes, and P. Bours, Gait Authentication and Identification using Wearable Accelerometer sensor, Proc. of IEEE Workshop on Automatic Identification Advanced Technologies, pp.220-225(2007).

[8] G. Trivino, A. Alvarez-Alvarez, and G. Bailador, Application of the Computational Theory of Perceptions to Human Gait Pattern Recognition, Pattern Recognition,Vol.43, pp.2572-2581(2010).

[9] M.O. Derawi, P. Bours, and K. Holien, Improved Cycle Detection for Accelerometer Based Gait Authentication, Proc. of the 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing(IIH-MSP '10), pp.312-317(2010).

[10] S. Mondal, A. Nandy, P. Chakraborty, and G.C. Nandi, Gait Based Personal Identification System Using Rotation Sensor, Journal of Emerging Trends in Computing and Information Sciences,Vol.3, No.3, pp.395-402(2012).

[11] A. Savitzky and M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least-squares Procedures, Anal. Chem., Vol.36, No.8, pp.1627-1639 (1964).

[12] F. Okumura, A. Kubota, Y. Hatori, K. Matsuo, M. Hashimoto, and A. Koike, A Study on Biometric Authentication based on Arm Sweep Action with Acceleration Sensor, Proc. of International Symposium on Intelligent Signal Processing and Communications, 2006(ISPACS '06). International Symposium on, pp.219-222(2006).

[13] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, SMOTE:Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Vol.16, pp.321-357(2002).

Management Issues and Solution on Smart Meter Communication System

Naoto Miyauchi [*], Yoshiaki Terashima[**], Tadanori Mizuno[***]

[*] Transmission & Distribution Systems Center, Mitsubishi Electric Co., Japan
[**]Faculty of Science and Engineering, Soka University, Japan
[***]Faculty of Information Science, Aichi Institute of Technology, Japan
Miyauchi.Naoto@ab.MitsubishiElectric.co.jp[*], tyoshi@soka.ac.jp[**], mizuno@mizulab.net[***]

Abstract - Recently smart meters begin to be installed into some kind of residential consumer systems in Japan. The smart meter enables automatic and real-time system inspection by observing situation of meter, communication equipment and others. As the result, these systems can achieve stable system operations.

Especially it is expected to control electronic power generation and distribution more efficiently and robustly by using smart meters in fields of the electricity industry. We have developed the smart meter communication system which consists of smart meters and management computing functions. The smart meters can communicate with the management computing functions through 920MHz wireless, cellular and PLC (Power Line Communication) networks.

This paper explains some issues to develop the management computing functions in the smart meter communication system and proposes the system solution to realize retrieving of electrical power consumption certainly.

Keywords: SMART GRID, NETWORK ARCHITECTURE, NETWORK MANAGEMENT SYSTEM

1 INTRODUCTION

An electrical power system consists of an electric power plant, an electrical power grid and a lot of consumers. In the conventional electric power system, electric power was flowing from the electric power plant to the consumers in one way. In recent years, renewable power generation systems such as solar cell, wind power and others attract attention. It is possible for the consumers to generate the renewable electric power in their homes. So it is necessary to take into consideration the reverse power flow of the electricity from consumers to the electric power plant.

Authors have proposed the system architecture of the smart meter management system at the electric power system and have developed a trial production system [1].

This paper analyzes the issues of management aspects of the smart grid communication system and proposes the system solution to realize retrieving of electrical power consumption certainly.

2 SMART METER COMMUNICATION SYSTEM OVERVIEW

2.1 Overview of Smart Grid

A smart grid is a next-generation electricity grid. In the U.S., the smart grid is examined as part of a Green New Deal Policy, and development is promoted by the following technical field [2].

Wide-area situational awareness: Monitoring and display of power-system components and performance across interconnections and over large geographic areas in near real time. The goals of situational awareness are to understand and ultimately optimize the management of power-network components, behavior, and performance, as well as to anticipate, prevent, or respond to problems before disruptions can arise.

Demand response and consumer energy efficiency: Mechanisms and incentives for utilities, business, industrial, and residential customers to cut energy use during times of peak demand or when power reliability is at risk. Demand response is necessary for optimizing the balance of power supply and demand.

Energy storage: Means of storing energy, directly or indirectly. The significant bulk energy storage technology available today is pumped hydroelectric storage technology. New storage capabilities—especially for distributed storage—would benefit the entire grid, from generation to end use.

Electric transportation: Refers, primarily, to enabling large-scale integration of plug-in electric vehicles (PEVs). Electric transportation could significantly reduce U.S. dependence on foreign oil, increase use of renewable sources of energy, and dramatically reduce the nation's carbon footprint.

Cyber security: Encompasses measures to ensure the confidentiality, integrity and availability of the electronic information communication systems and the control systems necessary for the management, operation, and protection of the Smart Grid's energy, information technology, and telecommunications infrastructures.

Network communications: The Smart Grid domains and sub domains will use a variety of public and private communication networks, both wired and wireless. Given this variety of networking environments, the identification of performance metrics and core operational requirements of different applications, actors, and domains—in addition to the development, implementation, and maintenance of appropriate security and access controls—is critical to the

Smart Grid.

Advanced metering infrastructure (AMI): Currently, utilities are focusing on developing AMI to implement residential demand response and to serve as the chief mechanism for implementing dynamic pricing. It consists of the communications hardware and software and associated system and data management software that creates a two-way network between advanced meters and utility business systems, enabling collection and distribution of information to customers and other parties, such as the competitive retail supplier or the utility itself. AMI provides customers real-time (or near real-time) pricing of electricity, and it can help utilities achieve necessary load reductions.

Distribution grid management: Focuses on maximizing performance of feeders, transformers, and other components of networked distribution systems and integrating with transmission systems and customer operations. As Smart Grid capabilities, such as AMI and demand response, are developed, and as large numbers of distributed energy resources and plug-in electric vehicles (PEVs) are deployed, the automation of distribution systems becomes increasingly more important to the efficient and reliable operation of the overall power system. The anticipated benefits of distribution grid management include increased reliability, reductions in peak loads, and improved capabilities for managing distributed sources of renewable energy.

In Japan, the government installs a study group and is overhauling the issues of an electric power system and information and communication technology as follows [3].

- Renewable energy (unstable power supplies, such as sunlight and wind force)
- Influence on power transmission and an electricity grid
- Integrated method of the communication technique for electric power control (non-IP (Internet Protocol)), and the Internet technology

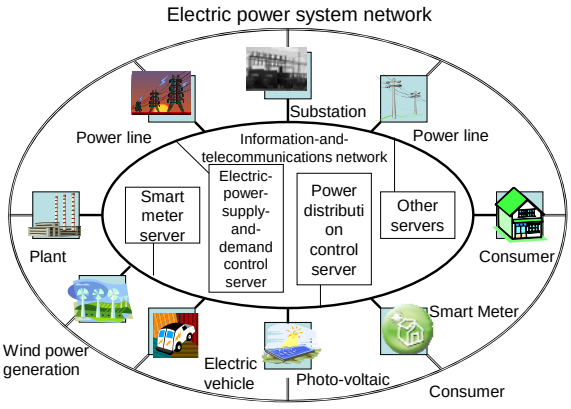


Figure 1. Smart Grid key map

The concept of a smart grid is shown in figure 1. In Figure 1, an electric power system consists of following equipment.

- Plant: It generates electricity using fire power, hydraulic power, atomic power, etc.
- Power line: The electric power generated in plant is turned to a substation, and electricity is transmitted.
- Substation: Transforming the electric power
- Power line: Electricity is supplied to a consumer.
- Consumer: The ordinary homes, the commercial establishment, the industrial institution which consume electric power.

The circle of the outside in a Figure 1 shows an electric power system network, and an inside circle shows the information-and-telecommunications network which supervises and controls power equipment.

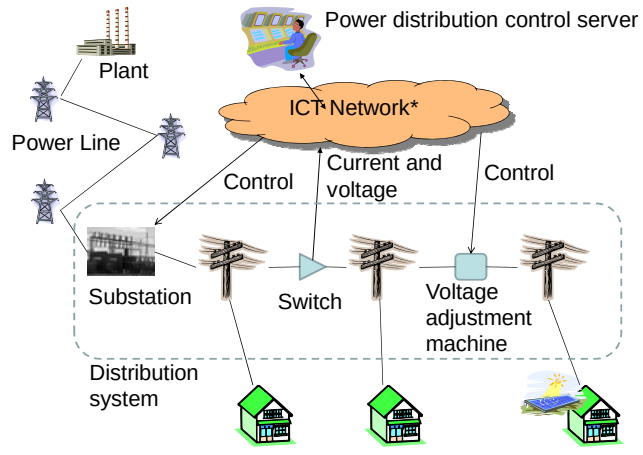
Server groups, such as an electric-power-supply-and-demand control server of a central portion, a power distribution control server, and a smart meter server, employ and control the whole smart grid.

2.2 Electric-Power-Supply-and-Demand Control

In the electric power company, frequency is kept constant by maintaining the demand of electric power, and the balance of supply.

Spread of power generation by renewable energy, such as sunlight and wind force, may lost the demand-and-supply balance of electric power greatly according to the weather.

Conventionally, in order to take the demand-and-supply balance of electric power, have absorbed change of demand-and-supply balance by thermal power generation by basing on the nuclear power generation of output regularity, and the pumped hydro power generation which buries the big demand difference during day and night, but If renewable energy increases, it will become impossible to store change of frequency in a rated value, and it will be expected that it becomes difficult to maintain the present electric power quality.



ICT* : Information and Communication Technology

Figure 2. The outline of power distribution control

In order to control the demand-and-supply balance of electric power, the supply-and-demand-control system in

consideration of power generation by renewable energy which carries out cooperation employment of a dynamo and the storage battery is needed.

2.3 Power Distribution Control

A power distribution control system controls to stop a power failure part at worst, also when the whole distribution system is supervised and an accident occurs, in order to supply electric power to a consumer stably on proper voltage (shown in Figure 2).

In the distribution system which connects a substation and a consumer, although wind power generation spreads through a subject and photo-voltaic is spreading home use through a subject, a large-lot user, Since it is easy to be subject to the influence of weather change, if power generation by these renewable energy spreads, the flow of the electric power of a distribution system may change suddenly in the unit of a part -- the reverse power flow turned to the substation from the consumer occurs -- and maintenance of proper voltage may become difficult only by conventional power distribution apparatus.

In order to solve these subjects, the flow of electric power is analyzed at high speed, voltage is predicted, and the voltage control system which supplies the electric power of proper voltage is needed.

2.4 Smart Meter

In the electric power company, in order to advance laborsaving of the electric power meter inspection-of-a-meter business installed in the consumer, the electric power automatic inspection-of-a-meter network system is planned.

An electric power automatic meter reading system consists of smart meters with a communication function (shown in Figure 3).

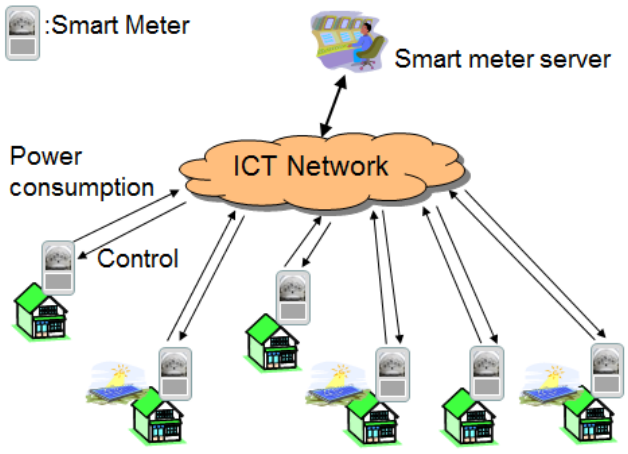


Figure 3. The outline of smart meter

By introducing a smart meter, it becomes possible to carry out the remote inspection of a meter of the power consumption of a consumer through a communication network.

By realization of the remote inspection of a meter, the timing of the inspection of a meter is sub divisible every 30 minutes from conventional 1 time per month, for example.

Also it becomes possible to grasp an amount demanded in real time.

It is expected that introduction of a smart meter enables it to perform electric-power-supply-and-demand control more correctly than before.

By introducing a smart meter, the following effect is expectable.

By collecting and controlling the production of electricity of the sunlight installed in a consumer, or wind power generation, electric-power-supply-and-demand control and power distribution control may be finely realizable.

Moreover, the mechanism in which energy saving is promoted may be able to be built by connecting a smart meter with the electrical machinery and apparatus in a consumer.

2.5 Outline of Trial Production System

In order to realize a smart grid, the telecommunications system which satisfies the following requirements is required.

- Interconnection can be carried out at high speed, diverting the telecommunications system for the existing electric power systems as much as possible.
- A secure system can be built so that failure of a subsystem may not affect a whole system.
- It prepares for the new service which will appear in the future, and has an open interface.

The trial production system was designed based on the requirements for the telecommunications system for realizing a smart grid. Figure 4 shows the schematic structure of a trial production system [4].

An electric-power-supply-and-demand server controls the apparatus in power generation plants, such as plant, and controls an electric power supply according to the electricity demand expected. In order to control apparatus, two kinds of networks (The RPR (Reverse Power Relay) transmission system network of a 1Gbps-Ehternet base, and a bus type controller network) were adopted.

A power distribution control server carries out the output surveillance of the electric power of a distribution system, measurement of voltage, and the photo-voltaic by the side of a consumer.

In order to supervise and control apparatus, the cable transmission network of the OFDM (Orthogonal Frequency Division Multiplexing) system was adopted.

Moreover, in order to carry out comparative evaluation, the optical fiber network of the GE-PON (Gigabit Ethernet-Passive Optical Network） system was also used together.

A smart meter server controls a switch while collecting electric energy from a smart meter.

The optical fiber network of the GE-PON system was adopted between the concentrators used as a server and the base station of a smart meter, and the 920MHz small electric

power radio mesh network was adopted between the concentrator and the smart meter.

Currently, a trial production is completed partially and these servers and networks are under evaluation.

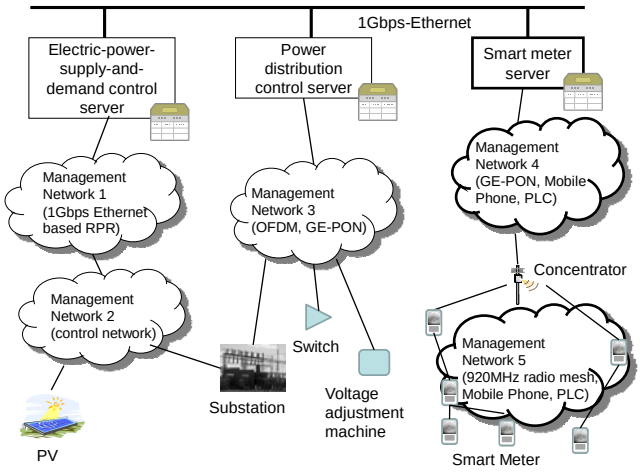


Figure 4. The schematic structure of a trial production system

Over 10 million electric power meter is installed in each electric power company in Japan, the communication network which connects with these electric power meter the smart meter server which collects and manages the measurement data of electric power meter is newly needed [5].

In order to reduce construction and maintenance costs of a communication network, the 920MHz radio mesh network was adopted, and the trial production system is designed.

A radio mesh network is a system transmitted to a concentrator, relaying subsequent data to the electric power meter in which the next adjoins the measurement data received from adjoining electric power meter, and had composition which stores 500 sets of electric power meter in one set of a concentrator.

Moreover, in order to avoid that the signal which two or more meter sends collides in order to judge transmitting timing autonomously, sharing the frequency to which each terminal was restricted, the transmitting timing control scheme was introduced.

3 SMART GRID MANAGEMENT SYSTEM

The schematic structure of the smart grid management system is shown in Figure 5. The Smart grid management system consists of an electric-power-supply-and-demand control management server, a power distribution control management server and a smart meter (SM) management server.

The electric-power-supply-and-demand control management server supervises and manages electric-power-supply-and-demand equipments of PCS (Power Conditioning System) and others through a RPR network and the other control networks.

The power distribution control management server supervises and manages power distribution control systems such as SVR (Supply Voltage Rejection) through an OFDM network, a GE-PON network and the other control networks.

The SM management server supervises and manages smart meters and concentrators through the GE-PON network and a specific power-saving wireless network.

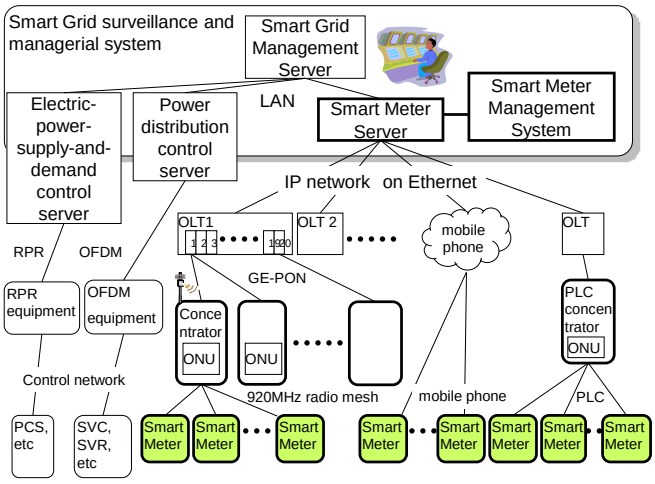


Figure 5. The schematic structure of smart grid management system

3.1 Management Protocol

SNMP (Simple Network Management Protocol) is used to manage the equipments which support IP communication.

And the management protocol is defined to support a local definition for equipments which IP communication is impossible. Also the protocol can transfer the non-IP protocol and the IP-protocol. As the result, communication equipment can use SNMP protocol as the management protocol.

3.2 Management Function

Generally, although FCAPS (Fault, Configuration, Accounting, Performance and Security) is needed as a controlling function of a telecommunications system, as management of a smart grid, it is thought that fault management, configuration management, a performance management, and a security management are required.

Since it was assumed in the future that many vendors and organizations participate in a smart grid, it was thought that a fee collection controlling function was also needed, but it carried out the outside of the range of a trial production system.

3.3 Management Information

About the equipment in which IP communication is possible, in order to use SNMP as a management protocol, management information also defines management information on the basis of MIB (Management Information Base) specified in IETF (Internet Engineering Task Force).

Also about the equipment in which IP communication is impossible, management information is defined on the basis of MIB as a definition of management information.

3.4 Managed Objects

When managing a smart grid, the equipment used as the candidate for management is as follows.

Smart grid management system supervises the equipment state and measurement value in these candidates for management.

Electric-power-supply-and-demand control: A photo-voltaic panel (PV), a power conditioner (PCS), various dynamos, a storage battery, electric-power-supply-and-demand server

Power distribution control system: Current transmission (CT), a transformer (VT), a protective relay, a stationary type reactive power compensating device (SVC), a pole transformer (SVR), the transfer device uncut [the electric current] off, a switch, power distribution control server

Automatic meter reading system: A watt hour meter (electric power meter), a concentrator, automatic inspection-of-a-meter server

HEMS (Home Energy Management System): A HEMS controller, lighting setup,

BEMS (Building Energy Management System): A lighting setup, an air conditioner, power equipment

FEMS (Factory Energy Management System): A production facility, a lighting setup, an air conditioner, power equipment

CEMS (Community Energy Management System) : Power generation equipment, electric car (EV)

When defining the candidate for management, in order to use SNMP as a management protocol, about the equipment in which IP communication is possible, management information also defined management information on the basis of MIB specified in IETF.

Also with the equipment in which IP communication is impossible, the extended definition was carried out on the basis of MIB as a definition of management information.

Smart grid management system performs the control actions (starting, a stop, reset, etc.) to these candidates for management while supervising the attribute information, including a state, an observed value, etc., and notice information in these candidates for management, including the notice of an obstacle, the notice of a change of state, etc.

4 SMART METER MANAGEMENT ISSUES AND THEIR SOLUTIONS

We have designed and developed the smart meter management system on the trial production system. The smart meter management system can observe situations of smart meters and communication equipment continuously. The smart mater management system gathers the following three pieces of management information to achieve the stable operation on the smart meter networks on the SNMP protocol.

A) **Equipment status;** which is status of hardware and software of smart meters

B) **Communication status;** which is status of communication between smart meter and smart meter server

C) **Retrieving ratio;** about every 30 minutes consuming electricity

Figure 6 shows relationship between management information and the smart meter communication systems.

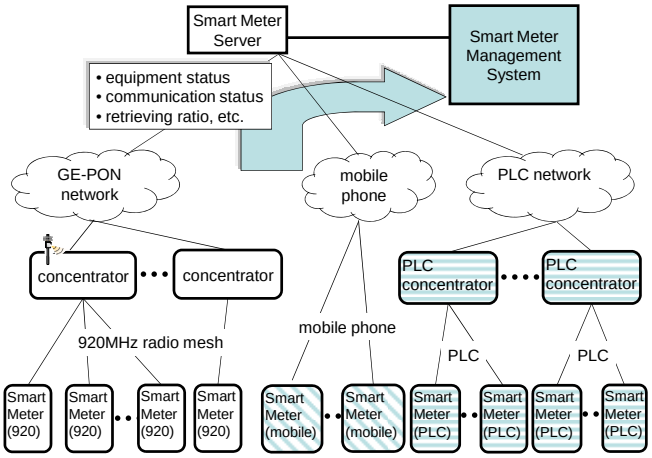


Figure 6. Management Information of smart meter communication systems

The smart meter management system has to retire electric power consumption from smart meters in fixed period which is currently 30 minutes. It is the most important to satisfy the fixed period during the retrieving electric power consumption certainly. So the smart meter management system is observing the management information about the smart meter networks continuously.

We propose a system solution to develop the smart mater management system to satisfy the fixed period in automatic and real-time system inspection. The system solution provides three methods to solve issues depend on Scalability, Connectivity and Smart meter setting of the smart meter management system. The following shows our proposed system solution in detail.

1. **Salability:** A number of smart meters are evaluated as from 2 million to 30 million in each Japanese electric company. Thus the smart meter management system must manage and retrieve the management information from huge number of smart meters, in appropriate fixed period certainly. To solve this issue, our smart meter management system has scale up architecture.

2. **Connectivity:** The smart meter management system must monitor connectivity between smart meters and

a smart meter server, in small communication traffic. To solve this issue, our management system retrieves smart meters events in every 30 minutes, and also retrieves equipment and communication status of smart meters once a day.

3. **Smart meter setting:** Smart meter must be set at appropriate geographical location, using appropriate communication mode (920MHz wireless, mobile phone, or PLC). To solve this issue, we have developed the smart meter setting simulator, and have planned about monthly meter setting, based on communication capability of each communication mode, and geographical meter location.

Figure 7 shows our proposed system solution based on Scalability, Connectivity and Smart meter setting. The smart meter management system is observing the management information as we noted earlier. The managers of the smart meter management system can identify trouble of communication equipments by analyzing the equipment status. And the managers can identify trouble of transmission for the retrieving of electric power consumption by analyzing the communication status. Because smart meters are set up at consumer's home site, the status of the smart meter communication may be changed with frequency. Especially it is serious issue in 920MHz wireless communication. The managers can confirm to retrieving of electric power consumption in fixed period that is predefined by analyzing the retrieving ratio.

The management information is transferred from the smart meter server to the smart meter management system through the smart meter server over the SNMP protocol. And some analysis functions for the management information are operating at the smart meter management system. The smart meter management system fulfills the role of manager function of the SNMP protocol. Smart meters, concentrators and communication equipment for GE-PON network, mobile phone and PLC network are managed through MIBs.

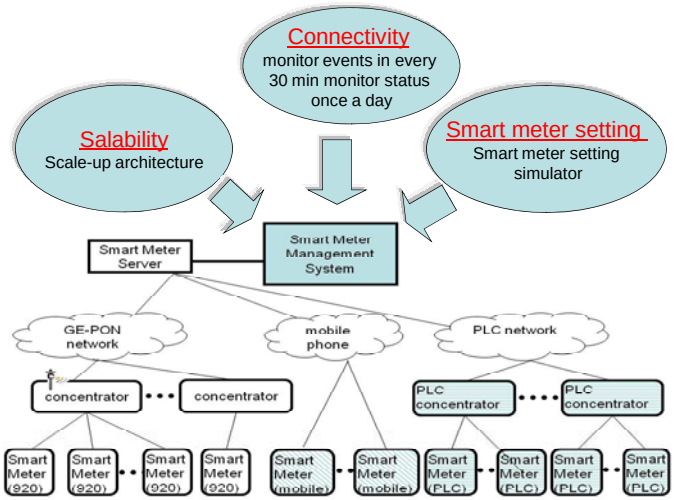


Figure 7. Management issues of smart meter management systems

5 CONCLUSION

This paper described overall of smart grid system which includes the smart meter communication system and proposed system solution for the smart meter management system. The proposed system solution can realize the stable retrieving electrical power consumption from smart meters in the fixed period. To satisfy the fixed period is the important to realize the power flow of the electricity from consumers to the electric power plant continuously in the smart grid communication system. We discussed about management issues of the smart meter management system implementation for the stable transmission of electric power consumption. And we proposed the system solution of the smart meter management methods for Scalability, Connectivity and Smart meter setting. Also we described the management system design based on the SNMP protocol.

Currently we have been developing and evaluating the smart mete communication system included the smart meter management system. The three issues that we analyze are getting to change for the worse depend on scale of the smart meter communication system. So we are planning to evaluate to retrieve electrical power consumption in various cases of different system scales.

REFERENCES

[1] Naoto Miyauchi, et.al, Smart grid management system – Design and trial development, International Workshop on Information (IWIN) (2012).

[2] National Institute of Standard Technology (NIST) Framework and Roadmap for Smart Grid Interoperability Standards (2010).

[3] The study group about the international standardization concerning a next-generation energy system, Toward the international standardization concerning a next-generation energy system. (2010).

[4] Naoto Miyauchi, et.al, Smart Grid Monitoring and Managing System, Information Processing Society of Japan (IPSJ) Special Interest Group (SIG) Technical Report (2011).

[5] Koichi Ishibashi, Smart network technology which realizes a smart grid, Business creation symposium of Japan and the Asia cooperation No.2 (2011).

Session 6:
Systems and Applications
(Chair : Tomoya Kitani)

Proposal for displaying discomfort information on the road targeting to the users of wheelchairs

Hiroshi Jogasaki[*], Yuta Ibuchi[**], Shinichiro Mori[***] , Yoshitaka Nakamura[*], and Osamu Takahashi[*]
[*]School of Systems Information Science, Future University Hakodate, Japan
[**]Graduate School of Systems Information Science, Future University Hakodate, Japan
[***] Fujitsu Laboratories Ltd.
G3113001@fun.ac.jp

Abstract Japan is entering into the age of decreasing birthrate and aging population. The numbers of users of wheelchairs is increasing and it is important to navigate users to identify the comfort level of the surface profile of the road. Also recently GPS (Global positioning system) sensors, acceleration sensors and gyro sensors and so on have been widely used and embedded in smartphones. People can correct many kinds of data from smartphone sensors and can share as knowledge using network. This article is addressing on the alert system to the users of wheelchairs who can sense the surface profile of road than any other people. We are proposing the alert system which has analyzed from discomfort point on the road using a three-axis acceleration sensor and a GPS sensor of smartphone to minimize the cost. Each user has different discomfort level. So we have proposed mapping solution of indicating each user's discomfort point on the road using interactive method on the smartphone to minimize the risk of the impact of the surface profile of the road.

Keywords: probe information system, navigation system for wheelchairs, smartphone, wheelchairs, discomfort level

1 INTRODUCTION

Japan is entering into the age of decreasing birthrate and aging population. Users of wheelchairs will be increasing. People cannot imagine how wheelchair uses are always facing discomfort because of the road condition. Even on the flat road, because of sensitivity of the wheelchairs, users feel discomfort level from the surface of the road. Typical wheelchairs are using tube less front wheel with small diameter. Major size is 7 inches. This is because front wheel is in charge of changing direction for wheelchairs [1]. That is the reason wheelchairs are so sensitive on the surface condition of the road surface. A solid tire with small diameter senses the difference of the surface of the road very much.

Visualization of the discomfort level of the surface of the road helps to navigate the users of wheelchairs to avoid discomfort course of the road. Discomfort level of the vibration on the wheelchair is mainly related to speed of the wheelchairs and weight of the user with the condition of the surface of the road. Sharp slopes also can be dangerous to the users of wheelchairs to control the speed of the wheelchairs. When users slow down just before the bad condition of the road, discomfort level can be mitigated. In case of the outdoor course, we can use GPS sensors on the smartphones to detect position on the road. On the same road, if the sidewalk is wide enough, it may be different condition even on the same road. It requires to detect 1m level of preciseness on the detected data of GPS sensors of smartphones.

There are so many types of wheelchairs in the market. Nowadays electric-powered wheelchairs are evolving and are becoming popular. But those are still expensive and not for many People. Major type in Schools and hospitals and department stores are human-assisted models. Those models are prepared whenever required people visited the site, site owners can provide moving ability. Use case of those types of wheelchairs are the people who are not familiar to the place are using and are trying road at the first time. We have picked up such a model for the experiments to certify the ability of our proposed system.

A sensor's technology on the smartphone is becoming popular because of the widespread use of MEMS (Micro Electric Mechanical System) technology. After the year of 2000, MEMS technology have used as many variety of sensors' Technologies like acceleration sensors on the automotive to detect crash for air back and gyro sensors to guide drivers for safety drives and protecting the hand vibration when pushing button of digital camera and so on. Many researchers are working on the advanced studies in this field of mobile sensing. Such mobile sensing are using the embedded technology with moving objects like cars, bikes, bicycles, humans. And also the number of users of smartphone is increasing dramatically and people can use many kinds of sensors like acceleration sensors and gyro sensors for logging data long period of time to gather information on the circumstances and can share the information within the people who knows the importance of such data. People can develop cost effective and convenient system using smartphones and a sensors' technology embedded in them. This paper is addressing three-axis acceleration sensors and GPS sensors on the smartphone to detect bad condition of the road with discomfort level of vibration on the wheelchairs.

Our study proposes a method for detecting bad condition of the surface level of the road which causes discomfort level of vibration to the humans. We have used human interactive input method to calculate each person's threshold of the discomfort level of the road. It is based on each person's discomfort level of the vibration from the input of each user and indicates potential discomfort positon on the road.

2 RELATED WORKS

There is three related works on navigation to the users of wheelchairs. One navigator has created on smartphone, and the other works are getting log from the commercial level three-axis accelerometers.

2.1 Indoor and outdoor navigation system for disabled

At the year of 2012, headed group by WATTANAVARANGKUL NATTAPOB have created the navigation system based on the static information like stairway and precipitous slope. Smartphone has been used to get the value from a direction sensor and communicate using Bluetooth to a wheel speed sensor. But it is not based on the condition like sharp drop in road level (Figure 1), or the sidewalk road under construction (Figure 2) or the surface of the old pavement on the road (Figure 3). Old pavement on the road causes discomfort level to the users of wheelchairs. That is why it cannot detect the discomfort level of the surface of the road [2].

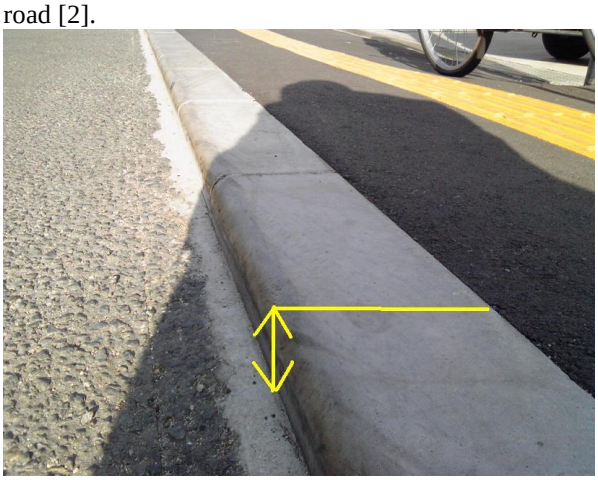


Figure 1.Sharp drop in the level of the road

Figure 2. Sidewalk which is under construction

Figure 3. The old pavement on the sidewalk

2.2 Unevenness evaluation of sidewalk pavement with vibration acceleration of wheel chair

In the study at the year 2004 by headed group by Miyoshi OKAMURA, they have proved that surface of the sidewalk with paver tile like Figure 5 causes unsatisfied vibration even bad effect physically to the human body. And also it has proven that accelerometer's value is proportional to speed. They found that type of surface like Figure 4 can be used by the users of wheelchairs during only 1 hour per day otherwise it causes side-effects. Similar type of sidewalk road can be seen like Figure 5 [3] [4] [5].

The facts they have found are as below.

1) Dominant frequency of the response of the acceleration value on the vibration of the wheelchair is almost same as integral multiplication of the space size between the joints.

2) Acceleration value is proportional to the speed of the wheelchair and acceleration value of vertical direction is most remarkable.

3) When the weight of the user becomes lighter, level of the vibration becomes larger.

4) Evaluate the comfort level of the vibration on the wheelchair by using the measurement of vibration acceleration value.

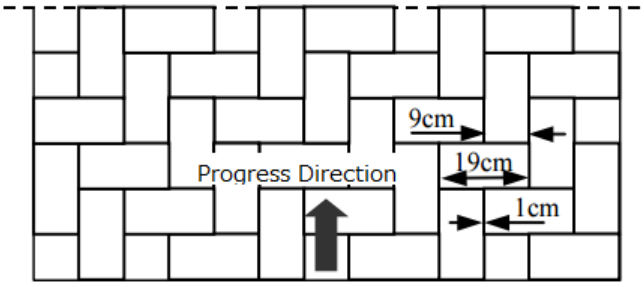


Figure 4. Surface condition on the sidewalk with paver tile and direction of movement

Figure 5. Sidewalk with paver tile

2.3 Spatiotemporal Life-Log Mining of Wheelchair Driving for Visualizing Accessibility of Roads

In the year 2013, headed group by Yusuke Iwasawa have tested using three-axis accelerometers to capture the surface condition of the road to show the results of classification and to visualize the results of tough/smooth surface detection. It can demonstrated as Figure 6.

But they were not using smartphone-based accelerometers and their objectives are to visualize accessibility of the road. They were not targeting the comfort level of the surface of the road [6] [7].

They have also studied on the comfort level of the users on the wheelchairs based on the value of VAL (Vibration Acceleration Level) theoretically. But we thought it needs human interaction to realize the potential discomfort level which is depending on the each user's sensibility.

2.4 Summary of related works

In Table 1, it show the advantages and disadvantages of the related works mentioned in this section.

The first work did not sense the surface on the road to detect the discomfort level of the vibration. Second work shows damage on vibration. But Second work and third work are using exclusive tools of accelerometers only measuring on this purpose and third work is targeting on the road not people's discomfort level.

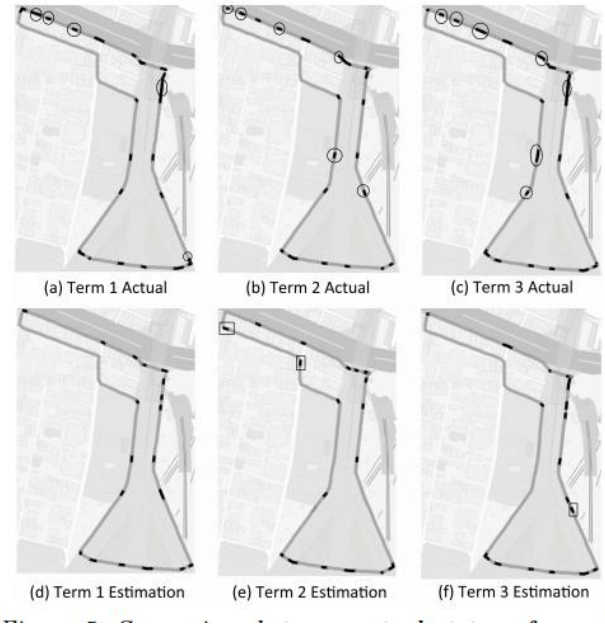


Figure 6. Comparison between actual status of ground surface and estimated status.

Table 1. Advantage and disadvantage of related works

	Concerning the surface condition on the road	Using Acceleration sensor on the smartphone
2.1. Navigation	No	No
2.2 Evaluation	Yes	No
2.3 Visualization	Yes	No

Because of such reasons, we propose a method of solving problem such as introductory cost and sensing the individual discomfort level of the surface of the road by smartphone sensors in this paper. The proposed method can proactively display the potential discomfort point of the road after user setup automatically the threshold value of the acceleration caused by the surface condition on the road.

3 PROPOSED METHOD

In this section, we explain a method of our proposing system how to approach to solve the existing problems in related works and the purpose of this study.

3.1 Purpose and approach

The related works have problems such as introductory cost, and not addressing the discomfort level of each individual level of the surface of the road.

Our proposed system is using a three-axis accelerometer and a GPS sensor on the smartphone to gather information at reasonable cost and using interactive method to address on human discomfort level of the surface of the road and its positioning through a GPS sensor. And also because discomfort levels are depending on the feeling of the people, it can address individual discomfort level of the value on Vibration.

This study aims to collect the acceleration values of the discomfort points of the road and to show the bad conditional points of the sidewalk which can be potential discomfort points for the users of the wheelchair on the map. Final target of this study is to create navigation system to avoid the discomfort point of the road in the future.

3.2 An overview of the proposed system

Proposed system consists of four parts as Figure 7.It is important to implement interactive input method to identify each person's discomfort level historically. First module is Data collection module on the smartphone which can collect the sensed data from three-axis accelerometers and GPS sensors on the smartphone. And second module is Data Process module on Server which can calculate the discomfort threshold value based on the gathered data. Third module is Data store module on Server receiving data from smartphone. Fourth module is Data display module in the smartphone which can show existing potential point of discomfort level on the road which is exceeding the level of each person's discomfort level into the map of smartphone.

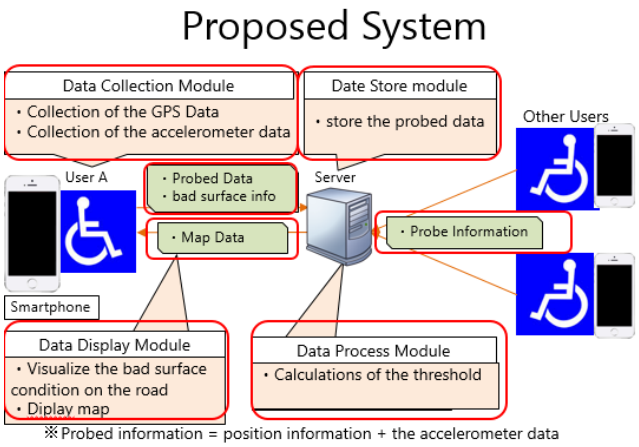


Figure 7. Proposed System

3.3 Visualizing the discomfort point of the road

In this section, we describe how our proposed method can visualize the discomfort points of the road in the system. Acceleration values of the Smartphone can be measured like Figure 8. From the timing of inputting the unique ID into the system, proposed system will store and calculate threshold of the discomfort levels of the users of the wheelchairs.

3.3.1. Start moving

We use iPhone4 as our smartphone which has fixed on the left arm of the wheelchairs to get the information. Data collection will start soon after user enter unique ID into the smartphone on the screen of Figure 9. Data gathering will continue during 60 seconds. Every 60 seconds, data will be sending to Server with the sets of the records of Table 2.

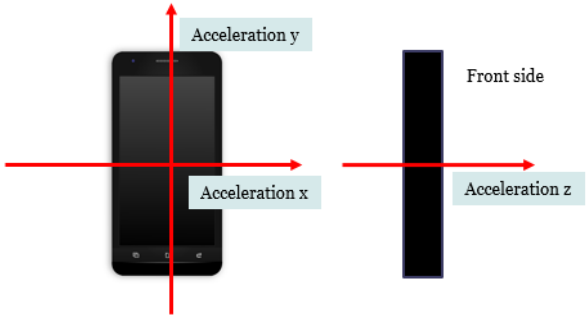


Figure 8.Acceleration value of Smartphone

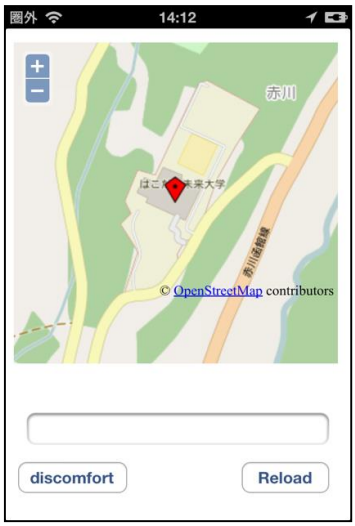


Figure 9. Data Collection User Interface on the smartphone

Table 2. Sending record to the server

UserID	ID of each user
Latitude	Value from GPS
Longitude	Value from GPS
Acceleration Y	Value from Accelerometer

3.3.2 Data collection

Data will be collecting in the smartphone every 1 second to have the value of latitude / longitude from a GPS sensor and the maximum/minimum acceleration y value from the three-axis accelerometer.

And user will push the discomfort button on the screen of the smartphone in case he feels discomfort level of vibration. If it is the first time to track discomfort level, that value will be stored in the DB accordingly as user's threshold value of Min/Max like Figure 10.

If the case is not the first time, it will be stored after calculating average value between the current value and historical value like Figure 11.

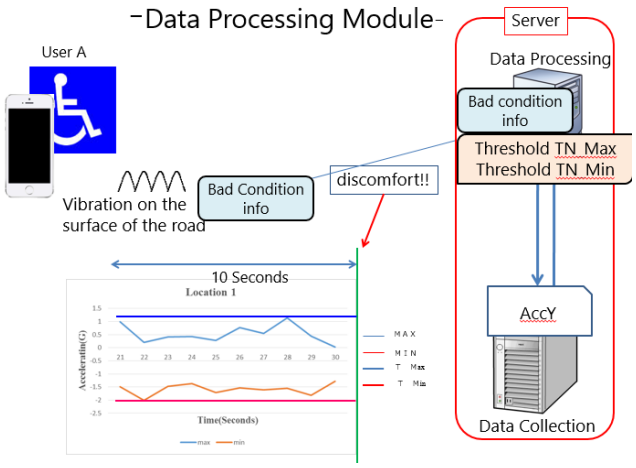


Figure 10. Data collection module in case of no data in DB

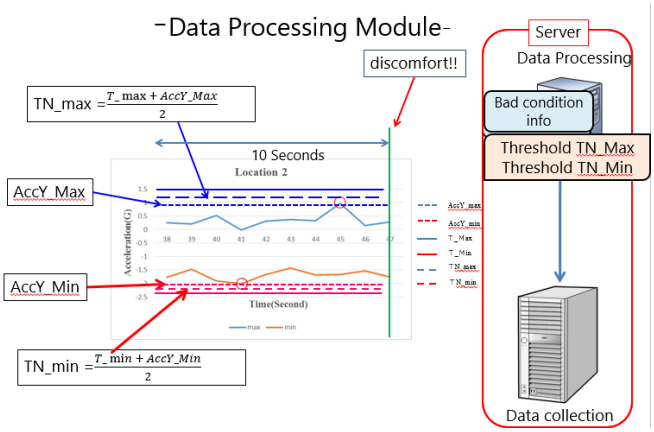


Figure 11. Data collection module with historical data

Values are stored like Table 3 to identify each user's discomfort profile on the DB.

Table 3. Stored record in the DB

UserID	ID of each user
Latitude	Value from GPS
Longitude	Value from GPS
Acceleration Y	Value from Accelerometer
Threshold max	Max threshold value of each user
Threshold min	Min threshold valu of each user
Comfort	information that feels discomfort

3.3.3 Data Display

After system records the value to DB, such probe information will be displayed on the screen of smartphone like Figure12. Every person's potential discomfort points can be seen on the map, if the person's threshold values are under the other person's threshold values. We have used OpenStreetMapAPI for our display API.

4 EXPERIMENTS AND DISCUSSIONS

In this section, we explain about preliminary experiments before creating proposed method and simulation we have done for confirming the effectiveness of the proposed method. And discuss about the results and concerning points.

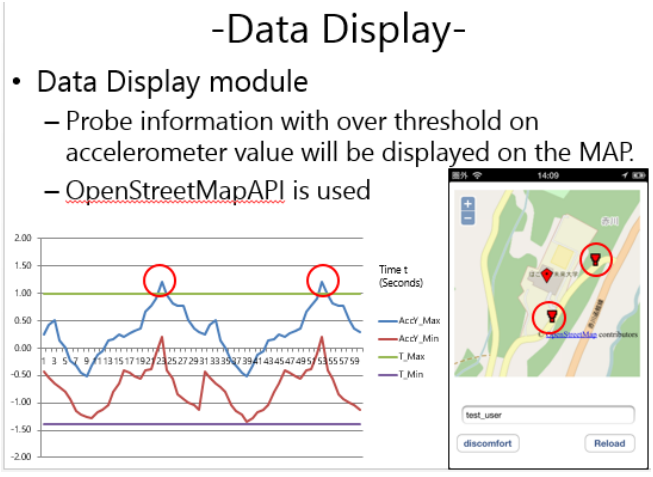


Figure12. Data display on smartphone

4.1 Preliminary experiment

Before creating proposed method, we have done the preliminary experiments to learn about the relationship between vibration on the wheelchairs and real condition of the surface of the road. We could find that the front tube-less tire is sensitive to the difference of the level of the surface on the road. For the assessment of the accuracy level to sense the surface profile of the road with vibration which can cause the discomfort level of the wheelchair users, we have tested as below.

(1) On the left arm of the wheelchair, we fixed on the smartphone. And collect data from the application of HASCLogger like Figure 13 on the testing board like Figure 14 with different in space between the several bars with the height of 2mm/4mm as Experiment 1.

(2) We have used major type of the wheelchair with front tire with 7inch, rear tire with 24 inch in diameter.

(3) We have used 4 patterns of difference of pattern of the testing board with 4mm/2mm height.

(4) Our test operated on each patterns 5 times with 5 persons. (4 patterns x 2 heights x 5 times x 5 persons = 200 times).

(5) Test subjects are 5 students with different weight (55kg, 58kg, 64kg, 65kg, 73kg)

(6) We have done walking tests from front door of the university to bus stop like Figure 18 as Experiment 2. With 5 persons round-trip.

In the Experiment 1, we have tested in the room how smartphone fixed on the left side arm of the wheelchair can detect the vibration from the surface of the road. Testing board on the floor has designed to be picked up the vibration from the ground by front tube-less tire of wheelchair. Testing board has been designed like Figure 14 to see how smartphone's accelerometer can detect the surface profile exactly from the value of the acceleration y from the vibration of the wheelchair. From the initial raw data, it cannot be distinguish exactly the signal from the surface of the road like Figure 15.

Figure 13. Experimental environment

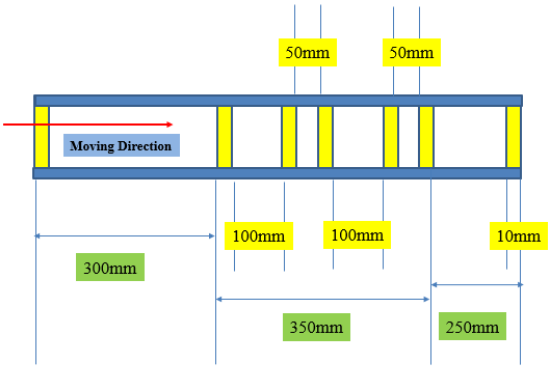


Figure 14. Example testing board with difference of the level

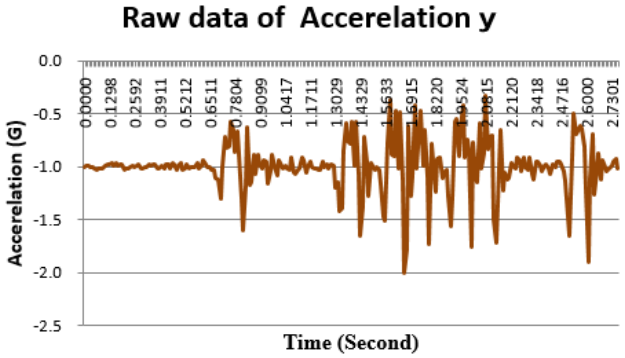


Figure 15. Raw data of Acceleration y

We have found that values of the accelerometer can be filtered noises by using 4 periods of moving average. In case of the raw data of acceleration related to the Figure 14 could be plotted like Figure 15.Many noises could be seen on the chart of the Acceleration. After using 4 periods of Moving Average value, Noise can be eliminated dramatically as Figure 16. Simplified graph can be seen as Figure 17. Out of 200 times test, almost 100% we could get exact difference of the level of the test board from the acceleration sensor of smartphone.

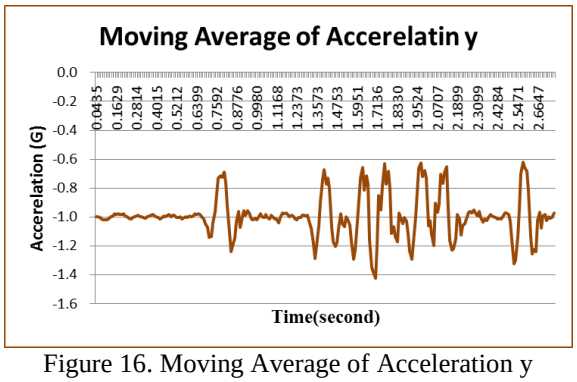


Figure 16. Moving Average of Acceleration y

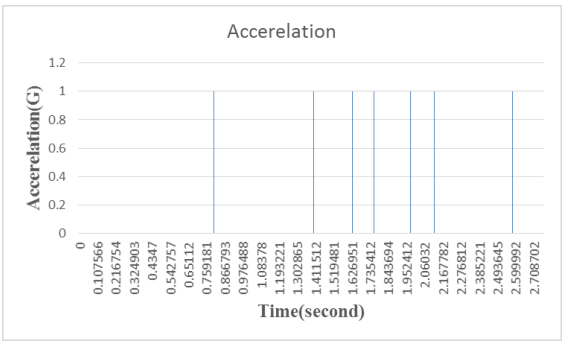


Figure 17. Simplified graph

In the experiment 2, testers have ride on the wheelchair at road of the outside of University. Wheelchairs with smartphone could gather the information of acceleration values and GPS values from the entrance to Bus-stop on the road like Figure 18. To figure out the real moving status like hearing the user's voice and seeing surface status of the road like smooth or rough or bumpy or so, test team took movie on the different smartphone by chasing testers behind the user of the wheelchair. Simulation

We have done the simulation on the data of which has gathered from the preliminary experiment 2 how proposed method can show the effectiveness.

4.2 Result of the simulation

From the entrance of the university to the bus-stop, we have got the logging Data on GPS and acceleration sensor value from the smartphone. Based on the assumption that user could feel three discomfort points such as after 27seconds and 45 seconds and 57 seconds from the starting point of the entrance of the university. In the chart on the Figure 19, Location 1 is after 27 seconds, Location 2 is after 45 seconds, and Location 3 is after 57 seconds. We could simulate the moving result using our proposed system to show discomfort point of the road on the map like Figure 19.

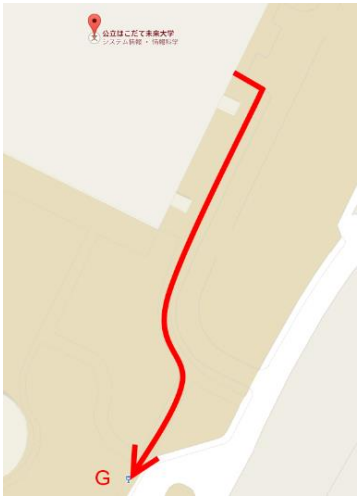


Figure 18. Experimental road from entrance to bus stop

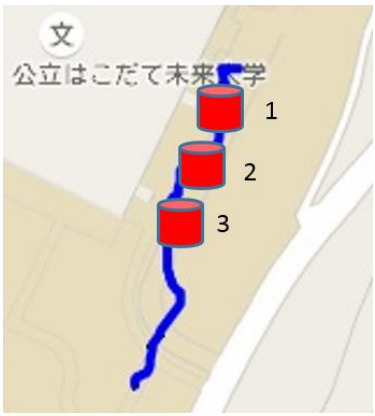


Figure 19 Experimental road from entrance to bus stop

These series of points have recognized as bumpy sidewalk like Figure 20 by seeing tracking movie on the smartphone.

Figure 20. Bumpy sidewalk

And also from gathered log data from the wheelchair, min/max data of acceleration can be plotted on the chart. First case is location 1 as Figure 21, second case is Location 2 as Figure 22, and third case is Location 3 as Figure 23. From these chart, it can pick up the man/mix value as threshold value of this person's discomfort level of the vibration on the wheelchair.

Figure 21. MAN/MIN value of Location 1

Figure 22. MAX/MIN value of Location 2

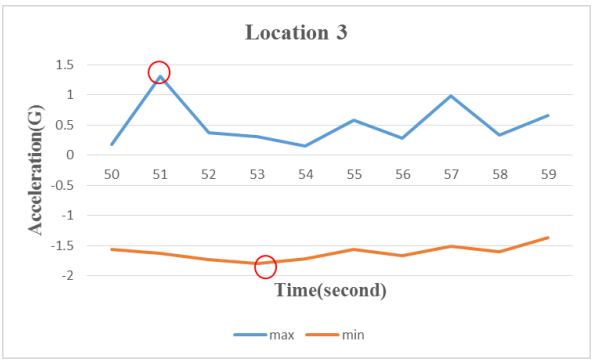


Figure 23. MAX/MIN value of Location 3

Smartphone stored accelerometer y values like Table 3.User could detect the discomfort level of the surface of the road on the smartphones' screen.

4.3 Discussion of experimental results

In this section, we discuss the experiment results.

We could show our proposed method how discomfort point can be visible to every user of wheelchairs. But there is two concerning points in the experiment data.

First point is related to speed of the wheelchair. If the wheelchair stops, vibration becomes nothing. The vibration levels of the wheelchairs are proportional to the speed of the wheelchair according to Okamura [3] [4] [5]. Too much speed causes unexpected sensitivity to the user of the wheelchair. There is law in Japan that wheelchair should run under the speed of 6km per hour which is equivalent to 1.667 meters per second. Because that acceleration values of Y are proportional to the speed of wheelchairs, when wheelchairs' speed are exceeding the value of 6km per hour, it can sense much bigger value of the acceleration value Y unexpectedly even the surface of the road is not bumpy. It can cause misleading alert to the users of wheelchairs. Points of discomfort should be plotted only when the speed of the wheelchairs are not exceeding the limit of the speed of the law. We should consider about that.

Second point is about the saturation value in minimum value. Because of the gravity, value of acceleration y is starting at the value of -1(G). Value of minimum is saturated on the value of -2(G) which is the minimum boarder value of smartphone. It is more suitable to use maximum value to identify the discomfort level of the surface of the road.

Table 3 Assumed stored value as threshold min/max

	Real Value		Threshold (after calcultion)	
	Acc Y min (G)	Acc Y max (G)	Acc Y min (G)	Acc Y max (G)
Location1	−2.0084475	1.14548875	−2.0084475	1.14548875
Location2	−2.007637	0.96374525	−2.00804225	1.054617
Location3	−1.79909875	1.038414	−1.9035705	1.0465155

5 CONCLUTION

There is existing problems in related works in the navigation to the users of wheelchairs. Points are reasonable cost for sensing acceleration data with GPS data and consideration about each person's discomfort level of the value of acceleration. We have proposed to visualize discomfort level of surface information on the road to the map on the smartphone. Every trigger has initiated by each user of wheelchairs. So wheelchair users can be helped by using this system. We have done the two pattern of experiment. First experiment was for basic experiment of which to find the effective way to track vibration of the wheelchairs on the road. Second experiment was for simulation on our proposal system how this system can display the discomfort point on the map.

After two experiments we have got the fact as follows.

1) We could understand that smartphone can use for the effective logging tool to identify discomfort levels of the sidewalk together with human interaction.

2) We have found that 4 periods of moving average can filter the noise of the signal from vibration of the wheelchairs.

3) Combination with GPS sensors and acceleration sensors can be effective to point out the discomfort point of the road.

4) After we get the each person's threshold, People can see the potential discomfort point of the road in the historical DB for their alert of the bumping state.

For the future work, we need to evaluate on the real road to get result on this method not by simulation. And to create navigation system to evaluate effectiveness of this method. And also to become more realistic navigation, we need to consider the slope identification by gyro sensors and scalability of servers and collectiveness of the probed data.

We have used GPS sensors on the smartphone. There is some possibility existing 10m of error range on GPS data in smartphone. So we should find the technology of sensitivity which can detect road 1 or road 2 exactly on the Figure 24.

Figure 24. Two sidewalks besides the road

REFERENCES

[1] Spinal outreach team and school of health and rehabilitation sciences university of Queensland, Manual wheelchairs – Information resource for service providers- Jan 2014

[2] Wattanavarangkul Nattapob and Wakahara Toshihiko, Indoor and outdoor navigation system for disabled, IPSJ SIG Technical Report (2012) (in Japanese)

[3] Miyoshi Okamura and Naohiro Fukada, Study on unevenness evaluation of sidewalk pavement with vibration acceleration of wheel chair, Japan Society of Civil Engineering, Pavement Engineering, Vol 9 pp17-24(DEC 2004) (in Japanese)

[4] Miyoshi Okamura, Effect of joint of tile pavement on vibration of wheelchair, Japan Society of Civil Engineering, Vol 64 No.1 pp237-246(March 2008)(in Japanese)

[5] Miyoshi Okamura, Study on performance indexes of sidewalk pavement by focusing on ride comfort of wheelchair, Japan Society of Civil Engineering, Pavement Engineering, Vol 14 pp189-194(DEC 2009) (in Japanese)

[6] Yusuke Iwasawa and Ikuko Eguchi Yairi, Life-Logging of Wheelchair Driving on Web Maps for Visualizing Potential Accidents and Incidents, The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 2012 , 3D2-R-13-9(in Japanese)

[7] Yusuke Iwasawa and Ikuko Eguchi Yairi, Spatiotemporal Life-Log Mining of Wheelchair Driving for Visualization Accessibility of Roads, The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 2013 , 1D3-5(in Japanese)

A Proposal of Lump-sum Update Method as Transaction in MongoDB

Tsukasa Kudo†, Masahiko Ishino‡, Kenji Saotome*, and Nobuhiro Kataoka**

†Faculty of Comprehensive Informatics, Shizuoka Institute of Science and Technology, Japan
‡Faculty of Information and Communications, Bunkyo University, Japan
* Hosei Business School of Innovation Management, Japan
** Interprise Laboratory, Japan
kudo@cs.sist.ac.jp

Abstract - Along with the progress of the cloud computing, it became necessary to deal with various and large quantity data in the distributed database environments. So, the various NoSQL databases have been proposed and put to practical use. However, as for the NoSQL databases, since it supports the distributed environment, the integrity of the database update is basically guaranteed only by the object unit. Therefore, there are serious restrictions to update the plural objects as a transaction. On the other hand, it is often necessary to perform the lump-sum or long-time update as a transaction in business systems. In this paper, we propose an update method to update the plural object in a lump-sum as a transaction in MongoDB, which is one of the NoSQL databases.

Keywords: database, NoSQL database, MongoDB, transaction processing, batch processing, concurrency control

1 INTRODUCTION

Nowadays, a large amount of data has been published, and it is utilized in various fields as big data. As a feature of big data, Volume (huge amount), Velocity (speed), Variety (wide diversity) have been pointed [6]. For example, large amounts of data, such as in the online shops and the video sharing sites, must be accessed efficiently by worldwide users, though it has more complex data structures than the conventional relational databases, including images and videos as well as texts.

To cope with this situation, various kinds of NoSQL (Not Only SQL) databases has been proposed and put to practical use [10]. As for the NoSQL database, to achieve the above-mentioned feature to the above problems, it is composed as the distributed database having a large number of servers, and to ensure the efficiency and reliability by redundancy such as replication and so on. Also, for example, MongoDB, which is one kind of the NoSQL database, is the document-oriented database and its structure is not defined by the schema. So, it is possible to add necessary attributes to its each data at any time and to manipulate various kinds of data flexibly [1].

On the other hand, unlike the relational database management system (hereinafter, "RDBMS"), it is not guaranteed to maintain the ACID property of the transaction processing in the case of the plural data manipulation. That is, it is generally maintained only on the individual update units called the atomic object. Also, as for the distributed environment, only the eventual consistency is guaranteed, that is, its consistency is not maintained until the completion of all the data manipulation including such as the synchronization of the replication

```
{ "_id" : 1, "name" : { "first" : "Tsukasa", "last" : "Kudo" },
"address" : "Hukuroi-shi, Shizuoka"}
```

Figure 1: Composition of MongoDB document.

[11]. They cause the serious restrictions on the data manipulation, for example, the intermediate result of the update having no consistency can be queried.

Here, even as for the RDBMS, there is a problem about maintaining the ACID properties in the case to update a large amount of data associated mutually in a lump-sum. That is, though the lock method is used to perform the above-mentioned manipulation concurrently with the other update, it causes the long latency of the latter. For this problem, we have proposed the temporal update method using the transaction time database that manages the history of the time series of the data, and shown that it is possible to maintain the ACID properties without this long latency even in the above-mentioned case [5]. In this paper, we propose an update method for MongoDB, which utilizes the concept of the temporal update, and show that the efficient lump-sum update maintaining the ACID properties can be realized even in the above-mentioned case. In other words, by this method, we can update the plural data in a lump-sum as a transaction, which was difficult to be executed by the conventional method.

The remainder of this paper is organized as follows. In Section 2, we show the problem of MongoDB about the concurrency control, and the abstract of the temporal update. In Section 3, we show our proposal for the lump-sum update method for MongoDB. In Section 4, we show the implementation and evaluations of this method, and show the considerations about this evaluation results in Section 5.

2 CONCURRENCY CONTROL OF MONGODB AND TEMPORAL UPDATE

2.1 Overview of MongoDB and Issue in Concurrency Control

MongoDB is a document-oriented NoSQL database, which data is the documents expressed by JSON (JavaScript Object Notation) format shown in Figure 1 [13]. The document is composed of the fields. For example, in this figure, {"_id": 1} is a field, of which identifier is "_id" and value is 1. Here, "_id" corresponds to the primary key of the relational

database. And, the field is able to have a nested structure, for example, the name field (name) in the figure is composed of the following fields: the first name field (first) and the last name field (last). Since the document structure of MongoDB is not defined by schema, any necessary fields can be added to any document at any time. So, each document is able to have different fields except "_id". Furthermore, since it is possible to store the various kinds of objects, such as images and videos to its fields, it can handle a variety of data compared to the RDBMS. Here, the set of document is called the collection. So, the collection and document correspond to the table and record in the relational database, though it is not strict.

As for the data manipulations, the following CRUD operations are provided as well as the RDBMS: insert, find (corresponding to select), update and remove (corresponding to delete). Furthermore, findAndModify command is also provided to execute both of the query and update exclusively, that is, they can be executed as the atomic operation. However, unlike SQL of the RDBMS, it does not provide the command to update the plural documents as a transaction. That is, there is a problem that the isolation and atomicity of the transaction cannot be maintained in the case where the plural documents are updated in a lump-sum.

For this issue, two phase commit protocol is shown [8]. In this method, for example, in the case of performing the account transfer from the account A to the account B of a bank, its processing ID is saved in the document of the account transfer management collection, which has the status about this processing: initial, pending, applied and done. These accounts are updated one after another in pending; meanwhile, this processing ID is saved to the documents of these accounts, and the updating accounts can be managed. Then, the status transit to applied. In the case of successful completion, this processing ID is deleted from the documents to exit. And, in the case of abnormal termination, the compensation transaction is performed to cancel the updates of the accounts and recover to a consistent state [8].

It is considered that this method is same as the saga for the RDBMS, which executes a mass update sequentially as divided plural update set. And, in the case of failure, the compensation transactions are executed to recover the data [7]. On the other hand, it has been shown that the isolation of the transaction is not maintained in the case of the concurrent execution with the other transactions. For example, in the case where the failure occurs in the transaction after the account A was updated and the result was read by the other transaction, the former transaction must be canceled. However, since the result of this transaction has been already queried by other transactions, it causes the problem in the actual system operations, such as the cascading aborts. Furthermore, in the case where the data related mutually is updated by the method like this, we showed that there is the problem that the consistency of data is not maintained [4].

2.2 Temporal Update Method for Relational Database

Currently, many databases of mission-critical business systems are built by the RDBMS, and a lump-sum updates of a large amount of data spanning a long period of time (hereinafter, "batch update") are often performed. For example, in the banking systems, there is a large amount of account transfer business, which is entrusted by the card companies and so on. Meanwhile, users update the database immediately (hereinafter, "online entry") for their deposits and withdrawals by the ATM. At present, the latter service is provided as a non-stop service. That is, the both of these processing have to be executed concurrently. Here, the lock method is generally used to maintain the ACID properties of the transactions while executing the batch update. However, since there is the conflict between the both processing, there is the problem that it makes the online entries wait for a long while.

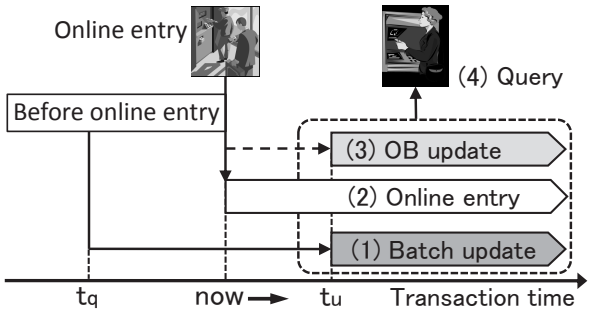


Figure 2: Data manipulation by temporal update.

For this problem, the authors have proposed the temporal update method shown in Figure 2. This method utilizes the concept of the transaction time database, which is a kind of temporal database that manages the time history of its data [12]. And, the transaction time expresses when some fact existed in the database, so its relation is expressed by $R(K, T_a, T_d, A)$. Here, K shows the primary key attribute of the data of above-mentioned fact; T_a shows the transaction time when the data was inserted into the database; T_d shows the transaction time when it was deleted from the database; A shows the other attribute. In other words, though the data is deleted logically from the database by setting the deletion time to T_d, it remains physically in the database. So, all the CRUD operations on the data of the database can be managed as a history of the time series. Here, until the data is deleted, the value now is set to T_d, which indicates the current time [14].

The feature of this method is that we avoid the conflicts between the batch updates and online entries by expanding the concept of the transaction time into the future. That is, as for the batch update, the data at the past time t_q is queried, and the processing result is stored at the future time t_u. On the other hand, as for the online entries, the data at the current time now is manipulated. Thus, the conflict between the both processing can be avoided without using the long locks. Here, the batch update processing must be applied to the result of the online entries performed between the time t_q and t_u. So, the former processing is applied individually to the result of the latter, and these processing results are stored into the database as the OB update in Figure 2.

As a result, three kinds of update result data is stored at the time t_u as shown in this figure: (1) the batch update, (2) the online entry and (3) the OB update. Therefore, the valid data has to be queried by the query processing, which is shown by

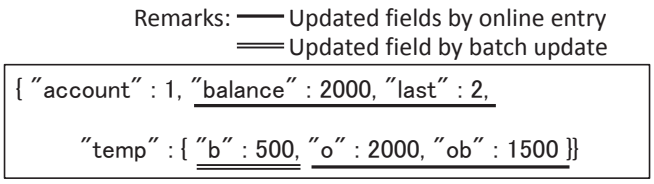


Figure 3: Data structure in proposal method.

(4) query. It is achieved by querying these data in the following order of priority: the OB update, online entry and batch update. To be concrete, in the case where both of the batch update and online entry are executed, the OB update result is queried; in the case of only the batch update, its result is queried; in the case of only the online entry, its result is queried; Therefore, we can query the same update result as in the case where the batch update is executed on the online entry results at the time t_u, without the long latency of the online entry by the lock method.

3 PROPOSAL OF LUMP-SUM UPDATE METHOD FOR MONGODB

In this section, we propose a lump-sum update method for MongoDB, which is based on the concept of the temporal update method mentioned in Section 2.2. Here, if we applied the temporal update intended for the RDBMS to the lump-sum update in MongoDB, multiple documents of the update result would be created for one fact of the real world: by the online entry, batch update and OB update. It means, for example, in order to execute the OB update with the online entry, the two documents represented by (2) and (3) in Figure 2 must be manipulated as one transaction. However, there is the restriction that the plural data cannot be updated as one transaction in MongoDB. So, if we apply this method to MongoDB just as for the RDBMS, the problem occurs: the consistency among data is not maintained. Meanwhile, since the document structure of MongoDB is not defined by the schema, its fields can be added flexibly. And, more importantly, the data manipulations on the different fields of the same document do not conflict.

For this reason, in this study, we propose the following lump-sum update method for MongoDB as shown in Figure 3. In addition, in the following, we omit to write "_id" of documents. In this method, all the results of the online entry, batch update and OB update are saved in the same document, and the valid field is queried in the same way as the temporal update for the RDBMS. This figure shows the document of the balance of the deposit account: the account number (account), balance (balance) and update number (last). Here, "last" corresponds to the time stamp, and it is increased by one for each update of the balance. In addition, it is used for the optimistic concurrency control as described later. Also, "temp" is the temporal field that is added temporarily during the execution of the batch update, and the update results are inserted to the corresponding field with the processing classification: "b" is for the batch update; "o" is for the online entry; "OB" is for the OB update. In addition, "balance" and "last" fields are also updated by the online entry. So as shown

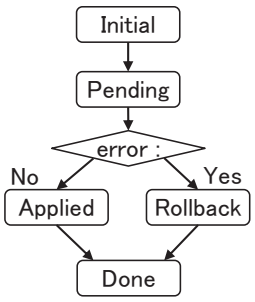


Figure 4: Processing stage transition in update

in Figure 3 by the underline and double underline, since the online entry, including the OB update, updates the different fields from the one of the batch update, there is no conflict between these update processing. Then, the valid update result can be queried similarly to the temporal updates, that is, by querying the update result data with the following priority: the OB update, online entry and batch update.

Also, in the temporal update, as shown in Figure 2, since the batch update starts at the time t_q and completes at t_u, the OB update must be executed, which accompanies the online entry during this period. Therefore, as for the proposal method, we define the following four processing stage like the two phase commit protocol in MongoDB as shown in Figure 4. And, the update processing is performed with transitioning them sequentially: "initial" shows that the stage is before batch update; "pending" shows it is during the batch update; "applied" shows batch update has completed, and the data of temporal field is being reflected to the regular field; "done" shows all the processing has completed. Incidentally, in the case where the failures occur in the batch update processing, the processing stage transitions from "pending" to "rollback". In the "rollback" stage, the batch update results are canceled, and the processing stage transitions to "done".

Here, the transition time from "initial" to "pending" corresponds to the above-mentioned t_q; the one from "pending" to "applied" corresponds to t_u. That is, the batch update is executed using the data as of the former transition time. Here, MongoDB is not the transaction time database. So, in the case where a data is updated by the online entry before the batch update, this batch update is performed on this result. However, since the result of the OB update accompanied with the online entry is queried based on the query priority, the query result is same to the temporal update method. And, at the latter transition time, the batch and OB update complete.

Figure 5 shows the data at the end of each processing stage. (1) shows the data at the end of "initial". Since it is prior to batch updates, temp field does not exist. Also, (2) shows "pending". Since the batch update has completed, the data has been set to temp field. In the case of this figure, the batch update debited 500 from the account. Meanwhile, the online entry deposited 1000 to the same account, and OB update debited 500 from this result. Then, all the results were stored in the temporal field. At this time, "balance" and "last" fields have been also updated by the online entry. Incidentally, since the value is set only to the fields corresponding to the executed updates, all the fields of temp are not always set.

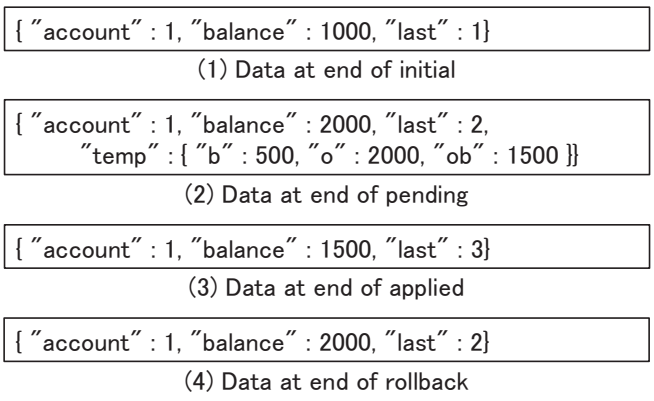


Figure 5: Data change in update

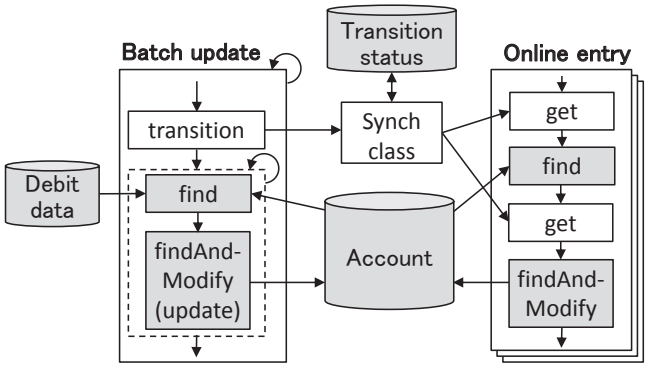


Figure 6: Software structure of prototype

While the processing stage is “applied”, the valid data is queried by the online entry transactions. In the case of this figure, the balance of 1500 in “ob” field is queried. Furthermore, in this stage, the valid data is reflected into “balance” field, then “temp” field is deleted. This is the processing for the next batch update. At the end of this stage, each field has the value shown in (3), and 1500 is set to “balance” field, which is the result of the OB update. In this way, the query results of the online entry do not change through this stage.

On the other hand, in the case where the batch update processing fails, the processing stage transitions to “rollback”. In this stage, only “balance” field is queried by the online entry continually; “temp” field is ignored. And, “temp” field is deleted without affecting the online entry. So, when the rollback has completed, “balance” field is not changed and this document become the state shown in (4).

In this way, the processing stage transitions to “done”, and we get the result (3) in the case of successful completion; we get (4) in the case of abnormal termination.

4 IMPLEMENTATION AND EVALUATIONS

4.1 Implementation of Prototype

To evaluate the proposed method, we constructed a prototype intending to manipulate the deposit accounts of the banking system. We use MongoDB Ver. 2.6.7 for the database; Java Ver. 1.6 for the programming language; MongoDB Java Driver Ver. 2.13 to access MongoDB from Java [9]. In addition, OS is Windows 7 (64bit). Figure 6 shows its software construction. The batch update and online entry programs are implemented by Thread class of Java to execute the both concurrently. Each program executes the following processes as shown in this figure: it query the data of the deposit account from the database (find); then, it updates the data of the database (findAndModify, update).

The batch update program executes the processing to debit from the deposit account collection (Account) in a lump-sum, based on the account and amount information stored in the debt data collection (Debit data). As shown in Figure 4, the processing stage transitions (transition) from “initial” to “pending”, then the batch update is executed. After its completion, the processing stage transitions to “applied”, and the data in “temp” field is reflected into “balance” field. Incidentally, in the case of abnormal termination, it transitions to “rollback”. Finally the processing stage transitions to “done”. The information of the processing stage is stored the transition status collection (Transition status), and it is accessed through “Synch” class from the batch update and online entry programs.

Meanwhile, the online entry executes the processing to deposit to each deposit account individually. As for this prototype, it was configured to perform deposits of certain amount of money from the plural terminals concurrently. Here, the online entry has to be accompanied by the OB update during the processing stage of the batch update is “applied”. So, its program was configured to query the processing stage by Synch class (get). And, to query this data efficiently by the program without accessing the database, it is saved in the instance of Sync class, However, in the case where the processing stage transitions from one stage to the next stage during the online entry executing, there is the possibility of the incorrect OB update execution. In other words, in the case where the transition occurs between the “find” and “findAndModify” in Figure 6, there may be the unnecessary OB update execution or the lack of it.

For this issue, We implemented Synch class using Synchronized keyword of Java, by which only one program can call it at the same time by the synchronization control. Then, we configured the online entry program to query the processing stage before not only “find” but also “findAndModify” as shown by the “get” in Figure 6. And, in the case where the transition occurs between them, the online entry program performs a retry. Furthermore, in order to prevent the transition between “get” prior to “findAndModify” and the completion of “findAndModify”, we configured Synch class to wait a certain time before transition that is requested by the batch update program. That is, the executing update of the online entry program can be completed before the transition by this way. Incidentally, while the processing stage is “applied”, not only “balance” but also the valid field has to be queried from this document. We implemented a class to manipulate the fields, and these manipulations were implemented by using the method of this class.

Since the online entries are executed from plural terminals concurrently, it is necessary to execute the concurrency con-

trol. So, we implemented the optimistic concurrency control by "last" field (update number) using findAndModify command, which is a method to perform the query and update of a document at the same time exclusively as mentioned in Section 2.1. And, in the case where the query condition matches to no data, the update is not performed and $null$ is returned as the query result. Therefore, we set the query condition of findAndModify command {"account":account number, "last": read updated number by "find" }, that is, the value of "last" is the result queried by "find" command just before. As a result, in the case where the target document was updated by the other program after the execution of this "find" command, no data matches this condition. And, the online entry program has to retry these processes from the beginning in this case.

Table 1 shows the target fields at each processing stage, which is queried and updated. As for the batch update, it is not executed when the processing stage is "initial" or "done"; it updates the different fields from the online entry when the stage is "pending" or "rollback". So, there is no conflict between the batch update and online entry. However, when the stage is "applied", the both update the same fields: "balance", "last" and "temp". That is, there is the conflict between them. Therefore, as for the batch update program, we also implemented the optimistic concurrency control using findAndModify command similarly to the online entry program. Incidentally, the batch update program queries "balance" when the stage is "pending", which is updated by the online entry program at the same time. That is, there is conflict between them. However, as we already mentioned in Section 3, the case where the online entry is executed, the OB update result becomes valid, which is created based on the execution result of the former. In other words, the batch update result is not used. Therefore, the concurrency control for this query is not required.

4.2 Evaluations of Concurrency Control

The proposal method does not lock the target documents through the duration of the batch update. That is, similar to the relational database, it has to be confirmed the inconsistencies by the concurrent execution of transactions do not occur. So, we performed the following three kinds of experiments to evaluate the concurrency control between the batch updates and online entry, using the prototype shown in Figure 6.

We performed the experiment in the case of successful completion of the batch update. The purpose of this experiment is to confirm that there is no lost update occurred by the illegal interface between the batch updates and online entries.

In this experiment, as shown in Figure 7, the number of the target deposit account is 60. And, its balance data is set prior to the experiment, which is calculated by the following equation as shown by the broken line.

$$balance = account\ number \times 1000 \qquad (1)$$

Here, the horizontal axis shows the account number of the deposit account; the vertical axis shows its balance. Then, the batch update program debits 20000 from the deposit accounts which account number is between 11 and 60. Here, the account, which balance is less than 20000 at this debiting time,

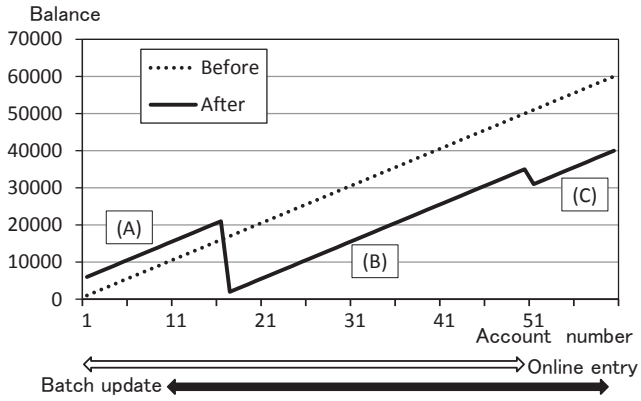


Figure 7: Result of case of successful completion

is excluded from this processing. In this experiment, since the batch update is successfully completed, the processing stage shown in Table 1 transitions from "pending" to "applied".

Meanwhile, the online entries are executed from five terminals concurrently, and each entry deposits 1000 to the deposit account which account number is between 1 to 50. Here, in order to avoid the conflict among the online entries, the different first update account number is assigned to each terminal: 1, 11, 21, 31, 41. Then, Each terminal updates the deposit account one after another. Here, after the program has processed the account which account number is 50, it processes the account which account number is 1. In this way, 50 deposit accounts are updated from each terminal.

The solid line in the Figure 7 shows this experimental result. The account indicated by (A) is not the target of the batch update or its balance was less than 20000. So, the batch update did not debit from it, and only the online entries deposited 5000. The account indicated by (B) is the target of the batch update, and the batch update or OB update debited 20000. Also, the online entries deposited 5000. So, the balance became 15000 reduction. The account indicated by (C) is not the target of the online entry, and only the batch update debited 20000. That is, even in the case where the batch update was executed concurrently with the online entry, no lost update occurs in both of the processes. As a result, we got the consistent update result.

5 CONSIDERATIONS

We consider whether the ACID properties of the transaction are maintained by the proposal method. As for the atomicity, the batch update completes as either of the following state: its update results after completion of processing are queried after the processing stage transits to "applied" as shown in Figure 7; it is canceled in the stage of "rollback". Therefore, the consistency was also maintained, that is, the collection transitions from a consistent state to another consistent state.

Next, as for the isolation, the batch update updates the different fields from the online entry, and the intermediate results of each processing are not queried by the other when the processing stage is "pending" as shown in Figure 3. Furthermore, in the both case of the successful completion and rollback, the

Table 1: Read and write fields in each processing stage

Processing stage	Batch update		Online entry	
	find	findAndModify	find	findAndModify
Initial	—	—	balance, last	balance, last
Pending	balance	temp.b	balance, last	balance, last, temp.o, temp.ob
Applied	last, temp	balance, last, temp (delete)	balance, last, temp	balance, last, temp (delete)
Rollback	—	temp (delete)	balance, last	balance, last
Done	—	—	balance, last	balance, last

batch update could be executed without affecting the online entries. So, the isolation is maintained. Lastly, as for the durability, the integration processing of the online entry and batch update results is executed when the processing stage is "applied". However, since the update results have been already reflected into the database, the durability is maintained by database management system of MongoDB. And, the query results do not change even if this process is interrupted.

Thus, the ACID properties of the transaction can be maintained by this method in MongoDB, even if the batch update is executed concurrently with the online entries. As shown in Section 3, this batch update corresponds to updating plural documents in a lump-sum. In other words, the update of plural documents in MongoDB can be executed as a transaction concurrently with the update of individual document.

6 CONCLUSION

Recent years, the utilizations of the NoSQL databases are spreading. However, there is the problem that the plural documents cannot be updated with maintaining the ACID properties of transactions. In this paper, we proposed the lump-sum update method for MongoDB, which is based on the concept of the temporal update method to execute the batch update as a single transaction in the relational databases. Concretely, in this method, the results of the following update are stored in temporal fields of the document and only the valid data is queried: the batch updates; the online entry; the OB update, which is applied the batch update individually to the online entry result.

We showed that the plural documents can be updated as a transaction by this method, even while the documents are being updated concurrently by the other transactions, that is, the online entries. Furthermore, we confirmed that this method achieves the above-mentioned function through the experiments using a prototype, which intended the deposit account.

Meanwhile, in the actual business systems, large number of transactions which update the plural data are executed concurrently. So, the future study will be focused on the concurrency control of such a update in the NoSQL databases.

This work was supported by JSPS KAKENHI Grant Number 15K00161. Also, the motivation of this study is an extension of a method, which has aimed to execute the batch update as a transaction in the relational databases and been registered as a patent [3], to the NoSQL database. We appreciate the members of Mitsubishi Electric Information Systems Corp. who supported to get this patent.

REFERENCES

[1] K. Banker, MongoDB in Action, Manning Pubns Co. (2011).

[2] J. Gray, A. Reuter, Transaction Processing: Concept and Techniques, San Francisco: Morgan Kaufmann (1992).

[3] Mitsubishi Electric Information Systems Corp., T. Kudo, Database System, Japan patent 4396988 (2009).

[4] T. Kudo, Y. Takeda, M. Ishino, K. Saotome, and N. Kataoka, Evaluation of Lump-sum Update Methods for Nonstop Service System, Int. J. of Informatics Society, Vol. 5, No. 1, pp. 21–28 (2013).

[5] T. Kudo, Y. Takeda, M. Ishino, K. Saotome, and N. Kataoka, An implementation of concurrency control between batch update and online entries, Procedia Computer Science, Vol. 35, pp. 1625–1634 (2014).

[6] D. Laney, 3D Data Management: Controlling Data Volume, Velocity and Variety, META Group, 2012, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[7] H. Garcia-Molina, and K. Salem, SAGAS, Proc. the 1987 ACM SIGMOD Int. Conf. on Management of data, pp. 249–259, 1987.

[8] MongoDB Inc., The MongoDB 3.0 Manual, http://docs.mongodb.org/manual/.

[9] MongoDB Inc., MongoDB API Documentation for Java, http://api.mongodb.org/java/.

[10] E. Redmond, and J.R. Wilson, Seven Databases in Seven Weeks: A guide to Modern Databases and The NoSQL Movement, Pragmatic Bookshelf (2012).

[11] P.J. Sadalage, and M. Fowler, NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, Addison-Wesley Professional (2012).

[12] R. Snodgrass and I. Ahn, Temporal Databases, IEEE COMPUTER, Vol. 19, No. 9, pp. 35–42 (1986).

[13] S.S. Sriparasa, JavaScript and JESON Essentials, Packt Pub. Ltd. (2013).

[14] B. Stantic, J. Thornton and A. Sattar, A Novel Approach to Model NOW in Temporal Databases, Proc. 10th Int. Symposium on Temporal Representation and Reasoning and Fourth Int. Conf. on Temporal Logic, pp. 174–180 (2003).

Parallel Multiple Counter-Examples Guided Abstraction Loop to Timed Automaton

Kozo Okano†, Takeshi Nagaoka‡, Toshiaki Tanaka‡, Toshifusa Sekizawa§, and Shinji Kusumoto‡

†Faculty of Engineering, Shinshu University, Japan
‡Graduate School of Information Science and Techlology, Osaka University, Japan
§College of Engineering, Nihon University, Japan
okano@cs.shinshu-u.ac.jp,
sekizawa@cs.ce.nihon-u.ac.jp,
kusumoto@ist.osaka-u.ac.jp

Abstract - Model checking techniques prove that a given system satisfies given specifications by searching exhaustively a finite transition system, which represents the system's whole behavior. If the system becomes large, it is impossible to explore the whole states in reasonable time. This is called the state explosion problem. One of the solutions to avoid the state explosion problem is using a model abstraction technique. Especially, Counter-Example Guided Abstraction Refinement (CEGAR) is considered as a promising technique. We have already proposed a concrete CEGAR loop for a timed automaton. This iteration loop refines the model in fine granularity level. It avoids the state explosion, however, the number of loops increases. This paper proposes a revised technique where multiple counter-examples are simultaneously applied in the refinement step of CEGAR. This device reduces the number of iteration loops. Experimental results show the improvement.

Keywords: CEGAR, Timed Automaton, Model Checking

1 INTRODUCTION

Recently, information systems play an important roles in social activities, thus software reliability becomes important. Model checking techniques[7] prove that a given system satisfies given a specification by searching exhaustively a finite transition system which represents the system's whole behavior. As systems become larger and more complicated, however, it is difficult to prove the reliability of the systems by model checking, because they need searching for whole states completely. For a large system, it is impossible to explore the whole states in reasonable time. This is called the state explosion problem.

One of the solutions to avoid the state explosion problem is a model abstraction technique. Especially, Counter-Example Guided Abstraction Refinement (CEGAR)[6] is considered as a promising technique.

In verification of real-time systems, a timed automaton is used[3], which can represent behavior of a real-time system. For a timed automaton, a real-valued clock constraint is assigned to each state of finite automaton (called a location). Therefore, it has an infinite state space which is represented in a product of discrete state space made by locations and continuous state space made by clock variables. In a traditional model checking for a timed automaton, using the property that we can treat the state space of clock variables as a finite set of regions; we can perform the model checking on a timed automaton. The size of the model, however, increases exponentially with clock variables; thus, an abstraction technique is needed.

Paper [16] firstly shows a concrete CEGAR loop for timed automata based on predicate abstraction techniques. It uses two abstraction models, over-approximation and under-approximation, while our previous approach [17] constructs an abstraction model based on only over-approximation. Their approaches similar in a sense that a location to be divided into two state while abstraction. This iteration loop refines the model in fine granularity level. It avoids the state explosion, however, the iteration grows.

This paper proposes a revised technique where multiple counter-examples are simultaneously applied. This device reduces the number of iteration loops.

Other related works include papers [11], [8], [9], [5], [2], [12], and [14].

For example, He et al. [11] have proposed a time abstraction technique and CEAGR loop with time abstraction technique as well as a compositional technique, which reduces state explosion occurring when we produce a product automaton from a network of timed automata. Papers [5] and [2] deal with hybrid automata and provide CEAGR for the model. None of these approaches, however, deals with refinement with multiple counter-examples.

This paper is organized as follows. Section 2 presents introductory material related to timed automata. Section 3 presents a short review of our previous proposed CEGAR for timed automata. Section 4 will provides our proposed multiple counter-examples abstraction refinement loop. Section 5 provides experimental results and discussions. The final section concludes the paper.

2 PRELIMINARIES

Here, we give a definition of a timed automaton and its related notions.

Definition 2.1 (Differential Inequalities on C)**.** *Syntax and semantics of a differential inequality E on a finite set C of clocks is given as follows:*
$E ::= x - y \sim a \mid x \sim a$,
where $x, y \in C$, a is a literal of a real number constant,

and $\sim \in \{\leq, \geq, <, >\}$. *Semantics of a differential inequality* $x \sim a$ *is true iff the evaluation of clock* x *is* $\sim$ *to* a. *That of* $x - y \sim a$ *is given in a similar way.*

Definition 2.2 (Clock Constraints on C)**.** *A set of clock constraints* $c(C)$ *on a finite set* C *of clocks is defined as follows: A differential inequality* in_1 *on* C *is an element of* $c(C)$. *Let* in_1 *and* in_2 *be elements of* $c(C)$, $in_1 \wedge in_2$ *is also an element of* $c(C)$.

Definition 2.3 (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 finite set of clocks;*
$I \subset (L \rightarrow c(C))$: *a mapping from a location to a clock constraint, called a location invariant; and*
$T \subset L \times A \times c(C) \times \mathscr{R} \times L$, *where* $c(C)$ *is a set of clock constraints, called a guard; and* $\mathscr{R} = 2^C$: *a set of clocks to reset.*

We denote $(l_1, a, g, r, l_2) \in T$ by $l_1 \overset{a,g,r}{\rightarrow} l_2$.

Dynamic of a timed automaton can be expressed via a set of locations and their evaluations. Changes of one state to a new state can be as a result of firing of an action or elapse of time.

Definition 2.4 (Clock Evaluation)**.** *Clock evaluation* $\nu (\in \mathbb{R}^{|C|}_{\geq 0})$ *for clock set* C *is a* $\mid C \mid$*-tuple of real values. An* i*-th tuple of* ν *corrsponds to the value of its associated clock. For a real value* d, $\nu + d$ *stands for* $(\nu^0 + d, \nu^1 + d, \ldots, \nu^{|C|} + d)$, *where* ν^i *stands for* i*-th tuple of* ν.

For a set of clock r *used in a transition,* $r(\nu)$ *is a new clock evaluation* ν', *where for every* $x \in r$, *its corresponding tuple is 0 else* ν^i.

For a guard g *used in a transition,* $g(\nu)$ *is evaluated as true or false in a usaul manner. Also is* $I(\nu)$ *for an invariant*I.

Definition 2.5 (State of a timed automaton)**.** *For a given timed automaton* $\mathscr{A} = (A, L, l_0, C, I, T)$, *let* $S = L \times V$ *be a set of whole states of* $\mathscr{A}$. *The initial state of* $\mathscr{A}$ *can be given as* $(l_0, 0^C) \in S$, *where* 0^c *stands for that each clock is evaluated as 0.*
For a transition $l_1 \overset{a,g,r}{\rightarrow} l_2$, *the following two transitions are semantically defined. The first one is called an action transition, while the latter one is called a delay transition.*

$$\frac{g(\nu), I(l_2)(r(\nu))}{(l_1, \nu) \overset{a}{\Rightarrow} (l_2, r(\nu))}, \qquad \frac{\forall d' \leq d \quad I(l_1)(\nu + d')}{(l_1, \nu) \overset{d}{\Rightarrow} (l_1, \nu + d)}$$

Transition $s \overset{\alpha}{\Rightarrow} s'$ is defined in Definition 2.6.

Semantics of a timed automaton can be interpreted as a labeled transition system.

Definition 2.6 (Semantics of a timed automaton)**.** *For a timed automaton* $\mathscr{A} = (A, L, l_0, C, I, T)$, *an infinite transition system is defined according to the semantics of* $\mathscr{A}$, *where the model begins with the initial state. By* $\mathscr{T}(\mathscr{A}) = (S, s_0, \overset{\alpha}{\Rightarrow})$, *the semantic model of* $\mathscr{A}$ *is denoted. Here,* S *is a sub set of* $L \times \nu$ *and* α *is an element of* $A \cup \mathbb{R}_{\geq 0}$. $\mathbb{R}_{\geq 0}$ *is a set of whole non-negative real numbers.*

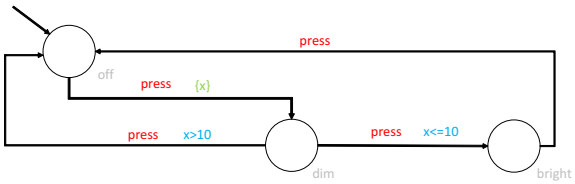


Figure 1: An Example Timed Automaton Representing Mug-light

In this paper, a state on a location l means an arbitrary semantic state (l, ν) such that ν satisfies l's invariant.

Definition 2.7 (run of a timed automaton)**.** *For a timed automaton* $\mathscr{A}$, *a run* σ *is finite or infinite sequence of transitions of* $\mathscr{T}(\mathscr{A})$ *as follows.*
$\sigma = (l_0, \nu_0) \overset{\alpha_1}{\Rightarrow} (l_1, \nu_1) \overset{\alpha_2}{\Rightarrow} (l_2, \nu_2) \overset{\alpha_3}{\Rightarrow}, \ldots$,
where $\alpha \in A \cup \mathbb{R}_{\geq 0}$.

Figure 1 is an example of a timed automaton, $\mathscr{AL} = (\{press\}, \{off, dim, bright\}, off, \{x\}, \emptyset, T)$, where $T = \{off \overset{press,-,\{x\}}{\rightarrow} dim, dim \overset{press,x\leq 10,-}{\rightarrow} bright,$
$dim \overset{press,x>10,-}{\rightarrow} off, bright \overset{press,-,-}{\rightarrow} off\}$.
It simply represents behavior of a mug-light with two bright modes. The brightness of the light is changed by the timing of pressing the single button of the mug-light.
One of possible runs of the $\mathscr{AL}$ is
$(off, (0)) \overset{0.5}{\Rightarrow} (off, (0.5)) \overset{press}{\Rightarrow} (dim, (0)) \overset{9.8}{\Rightarrow} (dim, (9.8))$
$\overset{press}{\Rightarrow} (bright, (9.8)) \cdots$.
For further detail about time automata, refer to [3].

3 CEGAR FOR TIMED AUTOMATA

3.1 Basic Algorithm

This section provides the base algorithm on abstraction refinement technique for the timed automata given in [16] and [17]. As mentioned above the algorithm in [16] and that of [17] is similar in abstract level. However, this paper proposes an extended method of [17], therefore, we describe the base algorithm based on [17].

Definition 3.1 (Abstraction assumption)**.** *The following condition is called Abstraction assumption.* $\forall i > 0 : (M_i \models p \rightarrow M_0 \models p)$, *where* M_i *is* i*-th abstract model. Model* M_0 *is the original model.*

The abstraction assumption should hold during CEGAR loop.

CEGAR loop[6] consists of the following four steps, namely *Initial abstraction*, *Model checking*, *Simulation*, and *Refinement*.

1. **Initial abstraction**
   An original model M_0 and a property p are given as input, and we abstract the original model M_0 and obtain an initial abstract model M_1.

We abstract the model preserving the abstraction assumption.

2. **Model checking**
   We perform model checking on the abstract model M_i. If a model checker outputs $M_i \models p$, then we can conclude that $M_0 \models p$ by the abstraction assumption. Then, we stop the loop. Otherwise, *i.e.*, the model checker outputs $M_i \not\models p$. Also a counter-example $\hat{\rho}_i$ is generated. We have to check every counter-example in P_i on the original model M_0, where P_i is a set of concretized runs on M_0, each of which is obtained from $\hat{\rho}_i$ by applying inverse of abstraction function h.

3. **Simulation**
   We check every concretized run in P_i on the original model M_0. If one of them is executable on M_0, then we conclude that $M_0 \not\models p$, because the found run is a real counter-example on M_0 and the property p. If none of them is executable on M_0, we have to refine M_i so that model checking on M_{i+1} does not produce the counter-example $\hat{\rho}_i$.

   We should notice that checking every run in P_i on M_0 can be performed symbolically using symbolical presentation on P_i or $\hat{\rho}_i$. We say that $\hat{\rho}_i$ is spurious when none in P_i is executable on M_0.

4. **Refinement**
   If $\hat{\rho}_i$ is spurious, then we refine M_i so that model checking on M_{i+1} does not produce the counter-example $\hat{\rho}_i$. The M_{i+1} is obtained automatically using $\hat{\rho}_i$. We repeat the loop by go to Model checking with M_{i+1}.

In our previous work, we give a concrete algorithm of CEGAR for a timed automaton. In the work, we only consider the reachability property as p. Thus, we check that $AG\neg l_e$, where l_e is an error location. The error location is a location where we think the control never reach.

The following subsections describe the details of each step.

3.2 Initial Abstraction

In Initial abstraction, we remove all of clock attributes from the given timed automaton[17].

Definition 3.2 (Abstraction Function h)**.** *For a timed automaton $\mathscr{A}$ and its semantic model $\mathscr{T}(\mathscr{A}) = (S, s_0, \Rightarrow)$, an abstraction function $h : S \to \hat{S}$ is defined as follows:*

$$h((l, \nu)) = l.$$

The inverse function $h^{-1} : \hat{S} \to 2^S$ of h is also defined as $h^{-1}(\hat{s}) = (l, D_{I(l)})$ where $\hat{s} = l$ and $D_{I(l)}$ is a region satisfying $I(l)$ representing by DBM.

Definition 3.3 (Abstract Model)**.** *An abstract model $\hat{M} = (\hat{S}, \hat{s}_0, \hat{\Rightarrow})$ of a given timed automaton $\mathscr{A}$ with its semantic model $\mathscr{T}(\mathscr{A}) = (S, s_0, \Rightarrow)$ is defined as follows:*

- $\hat{S} = L$;
- $\hat{s_0} = h(s_0)$; *and*
- $\hat{\Rightarrow} = \{(h(s_1), a, h(s_2)) \mid s_1 \overset{a}{\Rightarrow} s_2)\}$.

The i-th abstract model $\hat{M}_i = (\hat{S}_i, \hat{s}_{i,0}, \hat{\Rightarrow}_i)$ is obtained from the i-th timed automaton $\mathscr{A}_i = (A_i, L_i, l_{i,0}, C_i, I_i, T_i)$ by Definition 3.3.

Definition 3.4 (Abstract Counter-Example)**.** *A counter-example on $\hat{M} = (\hat{S}, \hat{s}_0, \hat{\Rightarrow})$ is a sequence (run) of states of $\hat{S}$ and labels. An abstract counter-example $\hat{\rho}$ of length n is represented in $\hat{\rho} = \langle \hat{s}_0 \overset{a_1}{\to} \hat{s}_1 \overset{a_2}{\to} \hat{s}_2 \overset{a_3}{\to} \cdots \overset{a_{n-1}}{\to} \hat{s}_{n-1} \overset{a_n}{\to} \hat{s}_n \rangle$. A set P of run sequences on $\mathscr{T}(\mathscr{A})$ obtained by concretizing a counter-example $\hat{\rho}$ is also defined as follows using the inverse function h^{-1}:*

$$P = \{\langle s_0 \overset{d_0}{\to} s'_0 \overset{a_1}{\to} s_1 \overset{d_1}{\to} s'_1 \overset{a_2}{\to} s_2 \overset{d_2}{\to} \cdots \overset{a_n}{\to} s_n \rangle \mid \bigwedge_{i=0}^{n-1} (s_i \in h^{-1}(\hat{s}_i) \wedge d_i \in \mathbb{R}_{\geq 0} \wedge s_i \overset{d_i}{\Rightarrow} s'_i \wedge s'_i \overset{a_i}{\Rightarrow} s_{i+1})\}.$$

We assume that a counter-example is a finite run [17]. We restrict the property to check as reachability, this assumption is reasonable. For a case of loop structures, see [17].

3.3 Model Checking

Abstract model $\hat{M}_i$ is a just automaton, therefore, we can use several model checkers at this step. We use UPPAAL to model check.

3.4 Simulation

Using the DBM library provided by UPPAAL team, we have developed a simulation program. Using DBM, checking every run in P_i on M_0 can be performed symbolically.

3.5 Refinement

The $(i+1)$-th abstract model $\hat{M}_{i+1}$ is obtained from a timed automaton $\mathscr{A}_{i+1}$ using the abstraction function h. The $(i+1)$-th timed automaton $\mathscr{A}_{i+1}$ is obtained from the i-th timed automaton $\mathscr{A}_i$ and a set of counter-example P_i.

Figure 2 shows the algorithm of Refinement, Algorithm 1.

Algorithm 1 is applied for a counter-example $\hat{\rho}(= \pi)$ and generates a refined timed automaton. It uses functions, Duplication(), RemoveTransition(), and DuplicateInitialLocation(). Functions Duplication(), RemoveTransition() and DuplicateInitialLocation() are functions to duplicate locations and transitions, to remove unnecessary transactions, and to duplicate the initial location, respectively. For the definition of these functions, please refer [17]. We omitted the detail of the sub functions. However, in Example 1, we give an example of refinement.

Paper [4] shows that clock condition in a form of $x - y < c$ cannot be dealt with. We assume that the following assumptions in the paper.

Assumption 1.
1. *We only check reachability: $AG\neg l_e$ for model checking.*
2. *The target timed automaton is diagonal-free, which means that the timed automaton does not contain clock condition in a form of $x - y < c$[4].*
3. *We assume that a counter example is a finite run.*

Refinement
Inputs $\mathscr{A}_i, \pi$

/* $\pi = \langle l_0 \xrightarrow{a_1,g_1,r_1} l_1 \xrightarrow{a_2,g_2,r_2} \cdots \xrightarrow{a_n,g_n,r_n} l_n(l_n = e)\rangle$ */
/* $succ\_list = \langle (l_0, D_0), (l_1, D_1), \cdots, (l_k, D_k)\rangle$, where (l_j, D_j) represents the j-th reachable state set along with π, and l_k is the last location reachable from the initial state. */
$succ\_list := tr(\pi)$
/* function tr () obtains $succ\_list$ form π */
$\mathscr{A}_{i+1} := \mathscr{A}_i$
for $j := succ\_list.length$ downto 1 **do**
  $e_j := (l_{j-1}, a_{j-1}, g_{j-1}, r_{j-1}, l_j)$
  $\mathscr{A}_{i+1} := Duplication(\mathscr{A}_{i+1}, succ\_list_j, e_j)$
    /* Duplication of the Location and Transitions */
  if $IsRemovable(\mathscr{A}_{i+1}, succ\_list_j, e_j)$ **then**
    $\mathscr{A}_{i+1} := RemoveTransition(\mathscr{A}_{i+1}, e_j)$
      /* Removal of Transitions */
    break
  else if $j = 1$ **then**
    $\mathscr{A}_{i+1} := DuplicateInitialLocation(\mathscr{A}_{i+1}, (l_0, D_0))$
      /* Duplicate the initial location and transitions from the initial location */
  end if
end for
return $\mathscr{A}_{i+1}$

Figure 2: Algorithm 1: Refinement Algorithm for a counter-example

Hereafter, we assume that Assumption 1 always holds in this paper.

Definition 3.5 (Badstate)**.** *For a given abstract model $\hat{M}$ and a counter-example $\hat{\rho}$, we simulate starting from initial state of $\hat{\rho}$. If we find a first state that doesn't satisfy the time constraints and there are no fireable actions anymore at the state on the abstract model, then we called the state as a badstate.*

Using Algorithm 1, Algorithm 2 in Fig. 3 applies each counter-example $\hat{\rho}$ in a given P_i. The result is sequentially reflected in the given timed automaton $\mathscr{A}_i$. If a badstate of a counter-example cannot be resolved, the counter-example is not reflected. In other words, for such a counter-example, Algorithm 1 is not applied. The next counter-example in P_i is chosen and the process is repeated.

Example 1. *Figure 4 depicts how Refinement transforms an i-th timed automaton to a refined $(i+1)$-th timed automaton. We can see that extra locations and transitions are added. On the $i+1$-th timed automaton, from the initial location A^1 we cannot reach the error location C.*

4 OUR NEW REVISED CEGAR LOOP

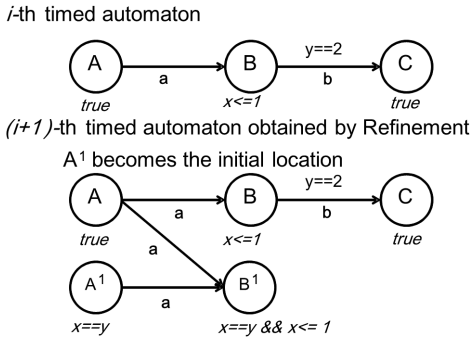
Our revised CEGAR loop differs in Model Checking, Simulation, and Refinement from the previous one.

Here, we describe each of them.

Refinement of CEs
Inputs $\mathscr{A}_i, P_i$

/* $P_i = \langle \rho_0, \rho_1, \cdots, \rho_k)\rangle$ */
$\mathscr{A}_{i+1} := \mathscr{A}_i$
for $j := P.length$ downto 1 **do**
  if $Refinement(\mathscr{A}_{i+1}, \rho_j)$ resolves a bad state of ρ_j **then**
    $\mathscr{A}_{i+1} := Refinement(\mathscr{A}_{i+1}, \rho_j)$
  end if
end for
return $\mathscr{A}_{i+1}$

Figure 3: Algorithm 2: Refinement



C is the error location

Figure 4: An Example of the Refinement

4.1 Model Checking

Normally, a model checker produces at most one counter-example. In our algorithm, we use master-worker configuration. Each worker performs model checking and generates a counter-example which we expect to be different to others. We describe how each worker generates a counter-example which we expect to be different to others, in Section 5.

4.2 Simulation

If one of concretized counter-examples can be executed on $\mathscr{A} = M_0$, then we conclude that $\mathscr{A} \not\models AG\neg l_e$. Otherwise we perform Refinement for every concretized counter-example simultaneously.

4.3 Refinement

The master gathers counter-examples from the workers, and performs Refinement algorithm shown in Fig. 3. End of the section, we will describe the modified version of Algorithm 2'.

The problem is to ensure not to produce an abstract model violating the assumption by applying more than one counter-example. First, Definition 4.1 provides the definition of the overlap of counter-examples.

Definition 4.1 (Overlap of the Counter-Examples)**.** *For counter-examples $\hat{\rho}_1$ and $\hat{\rho}_2$, an overlap of counter-example $\hat{\rho}_1$ and*

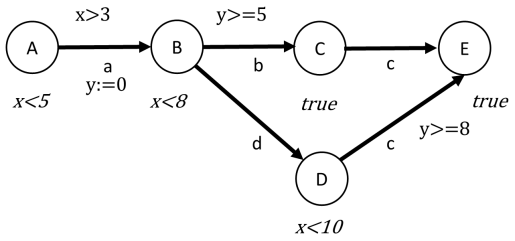


Figure 5: An Example Timed Automaton

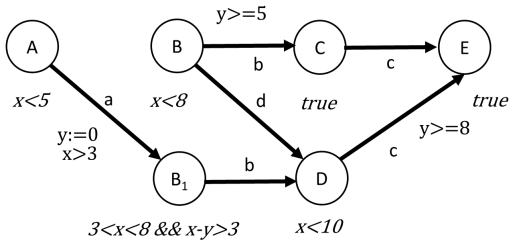


Figure 6: Timed Automaton Refined with B-C Transition

$\hat{\rho}_2$ is defined as that at least one same transition is in both of counter-examples $\hat{\rho}_1$ and $\hat{\rho}_2$.

Definition 4.2 (Correct Refinement)**.** *For a given set of counter-examples P and an abstract model $\hat{M}_i$, we say $\hat{M}_{i+1}$, which is produced by Algorithm 2, is a correct refinement of $\hat{M}_i$ when $\hat{M}_{i+1}$ preserves Abstraction assumption.*

In usual, we would obtain multiple abstract models $\hat{M}_{i+1}$ depending on the order of applying multiple CEs from $\hat{M}_i$. Theorem 1 proves that we can perform model checking nevertheless the order of applying multiple CEs varies.

Hereafter, we think timed automaton $\mathscr{A}_i$ and its abstract model $\hat{M}_i$ is any i-th timed automaton and abstract model obtained by i-times iteration of Algorithm 2, respectively.

Theorem 1. *For a given set of counter-examples P and an abstract model $\hat{M}_i$, $\hat{M}_{i+1}$ is a correct refinement of $\hat{M}_i$.*

Theorem 2. *Termination of CEGAR loop CEAGR loop using Algorithm 2 terminates.*

Theprem 1 and **2** support the correctness of our proposed method. Here, we omit the proofs.

The refined abstract models might be different depending on what order the counter-examples are applied. Also the refined abstract models might be the same.

Example 2. *For example, let consider a timed automaton in Fig. 5. Due to the clock constraints, neither a transition from B to C nor from D to E is firable.*

There are two counter-examples: $A \to B \to C \to E$ and $A \to B \to D \to E$.

First, let's consider the case the first counter-example is applied. Algorithm 1 generates a copy B_1 from location B, and generates a transition from B_1 to D as well as a transition from A to B_1. Finally it removes transition from A to B (Fig. 6). Next it applies the second counter-example. It generates a new location D_1 and removes a transition from B_1 to D (Fig. 7).

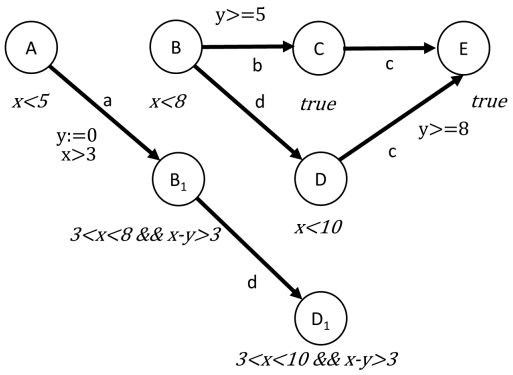


Figure 7: Timed Automaton Refined with D-E Transition

Refinement of CEs revised
Inputs $\mathscr{A}_i, P$

```
/* P = ⟨ρ0, ρ1, ··· , ρk)⟩ */
A_{i+1} := A_i
for j := P.length downto 1 do
    A_{i+1} := Refinement(A_{i+1}, ρ_j)
end for
return A_{i+1}
```

Figure 8: Algorithm 2': Refinement Algorithm (Multiple Paths)

Next let's consider the case the second counter-example is firstly applied. It generates a new location D_1 then generates also B_1. Finally, it removes a transition from A to B. The result is the same as Fig. 7. The application of the first counter-example does not affect the shape of the timed automaton.

Therefore, this example produces the same refinement. Note that we cannot reach location E from A in Fig. 7.

4.4 Revised Refinement Algorithm

For efficiency, we introduce a modified algorithm, Algorithm 2' shown in Fig. 8.

The differences between Algorithms 2 and 2' are the following two points.

- The set P_i of counter-examples is $\bigcup_{j \in \mathbb{K}} P_{ij}$ in Algorithm 2', where P_{ij} is a set of concretized j-th worker's counter-example in i-th iteration, and $\mathbb{K}$ is a set of workers.
- Algorithm 2' does not check the resolution of a bad state.

The major difference is the Algorithm 2' does not check the resolution of a bad state, which improves the efficiency. However, it means that there exists a counter-example which refines the abstract model but the counter-example is still pseudo in the refined abstract model. However, such a counter-example

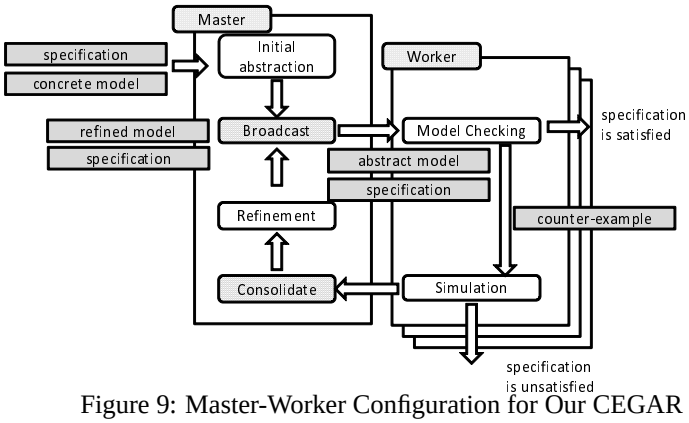


Figure 9: Master-Worker Configuration for Our CEGAR

will be detected in the next or future CEGAR loops. Therefore the modified Algorithm 2' also works fine.

5 EXPERIMENTS

5.1 Overview

We have performed experiments using two typical examples. One is Fischer's mutual exclusion protocol. Several p processes with the same shape of an automaton share a critical section. Mutual exclusion is established in a protocol using clock variables. Therefore it is a typical symmetric structure.

Another one is Gear Controller [15]. It is a model consisting of an engine, a gearbox, a human interface, a gear controller, and a clutch. It is a parallel system of hetero six components.

Before applying our tool, we need to obtain a single timed automaton presentation of Fischer's protocol (and Gear controller) since our proposed method cannot deal with a network of timed automata, which is used in UPPAAL verifier in general.

We performed the experiments under the following environment.
Master
CPU: Intel(R) Core™ 2 Duo
CPU L7700 1.80GHz
MM: 2.00GB OS: Ubuntu 10.0.4
Workers (14 cpus)
CPU : Dual Core AMD Opteron™
Processor 2210 HE 1.80GHz
MM: 6.00GB OS: CentOS 5.4

Figure 9 depicts the overview of the whole system. We use RMI framework on Java for communication between the master and workers. Each worker performs Model Checking and Simulation for its assigned abstract model.

We compare two strategies, the fastest trace and the shortest trace.

The fastest trace uses multiple counter-examples which the model checker UPPAAL finds with fastest trace option. The model checking is performed at each worker. With the fastest trace option, we can expect that each process of UPPAAL generates a different counter-example to others due to randomness of selection on next actions.

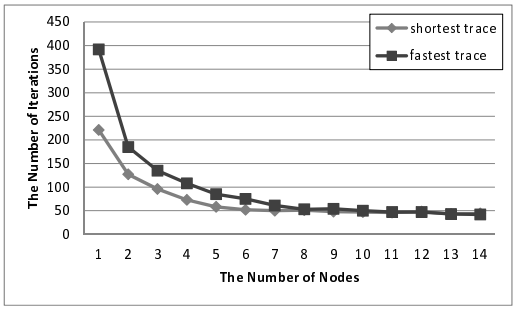


Figure 10: The Number of Iterations : Fischer's protocol

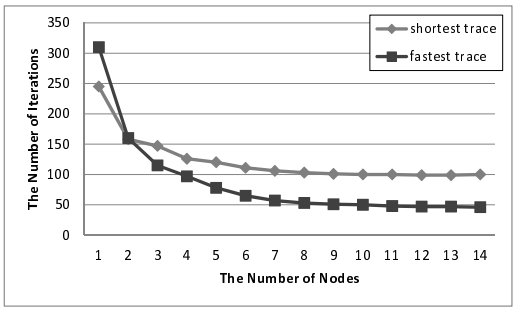


Figure 11: The Number of Iterations : Gear Controller

The shortest traces uses multiple counter-examples which the model checker UPPAAL finds with shortest trace option. Also with the shortest trace option, we expect that each process of UPPAAL generates a different counter-example to others due to randomness of selection.

The results are averages of five trials of the same configurations.

5.2 Results

Figures 10 and 11 show the results of the number of iteration. In both of Fischer's protocol and Gear Controller, the number of iteration decreases according to the number of workers. The shortest trace for Gear Controller has little effect.

Figures 12 and 13 show the results of the CPU times. The performance is improved according to the number of work-

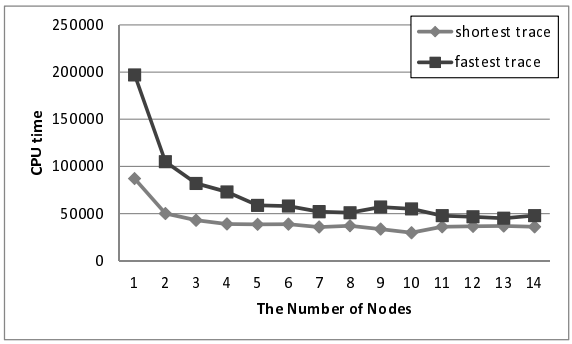


Figure 12: Execution Times : Fischer's protocol

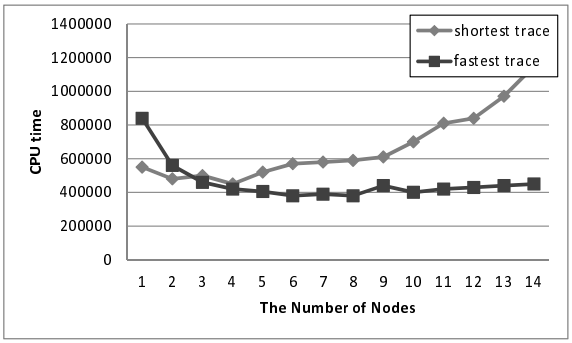


Figure 13: Execution Times : Gear Controller

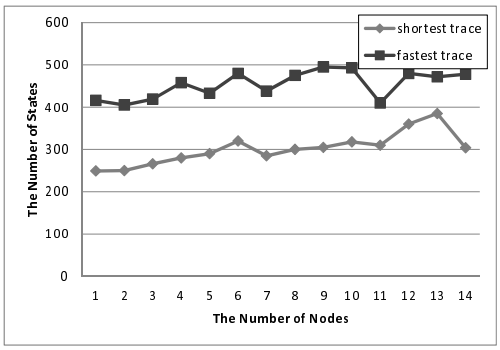


Figure 14: The Number of States : Fischer's protocol

ers, in Fischer's protocol while Gear Controller shows worse behaviors. The fastest trace also loses its acceleration but the shortest trace requires more time from four workers.

We can see that the numbers of iteration are improved in both of the cases, while CPU times are not.

This observation supports that our proposed method is potentially effective, however, we also consider the reason why CPU time is not improved.

We think two possibilities on the results.

One is the following hypothesis: Refinement with multiple counter-examples certainly refines parts of the automaton, however, which are not essential part of verification for property p. Thus, the refinement increases the size of the automaton, which increases CPU time.

The other one is the following hypothesis: the same counter-examples are generated. If some of workers generate the same counter-examples, then the efficiency will be worse. Such a phenomenon occurs because the random selections do not always guarantee that every counter-example is different to others.

Based on the above observations, we have performed the following additional experiments. First we have evaluated the number of states. If it increases according to the number of workers, then we can conclude that unnecessary states are generated.

Second we have also evaluated the ratio of unique counter-examples, which is a good index for the second hypothesis.

Figures 14 and 15 show the number of states. Fisher's protocol has gradual increase, while fastest trace of Gear Controller has strong increase.

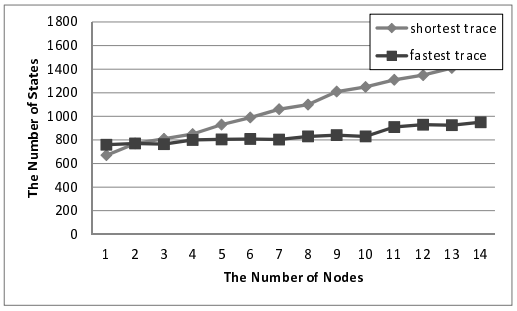


Figure 15: The Number of States : Gear Controller

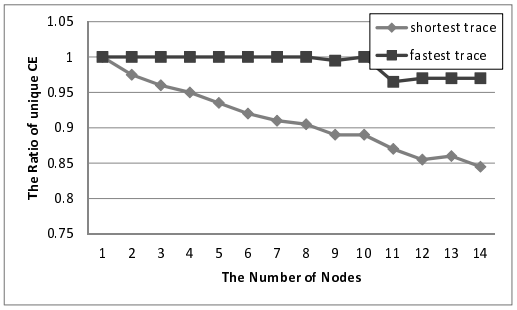


Figure 16: The Ratio of the Same Counter-Example : Fischer's protocol

Figures 16 and 17 show the ratio of unique counter-examples. If the ratio is equal to 1.0 then it means that every counter-example is different to each other. The shortest traces show that increase of the same counter-examples according to the number of workers.

The results support both of the hypotheses. In order to avoid such a problem, priority among the counter-examples is considered. Using the priority, we can control level of the refinement by filtering counter-examples used. We think, however, that there is no silver bullet, in other words the priority cannot be determined statically and in advance. As an approximate solution, we adopt threshold on the length of counter-examples. The idea is that we only use shorter counter-examples than threshold by the length of the shortest counter-example.

In order to avoid duplication of counter-examples, we think k-shortest path algorithm is worth to try. The algorithm is provided by Eppstein[10] and Jiménez [13].

Since UPPAAL uses more sophisticated data structure than DBM which we use and it also uses partial order reduction technique whereas we don't use any further improvements. Therefore we show the comparison between naïve approach and our approach in order to show the improvements.

We think that the experiments show our approach reduces the number of iteration, which also will improve the size of states of abstraction models.

The proposed works better than naïve CEGAR loop does. It is because the propose method can deal with more large system than the naïve CEGAR, in some cases. The CPU time is also improved. It implies that the main idea that we simultaneously apply the multiple counter-examples will improve

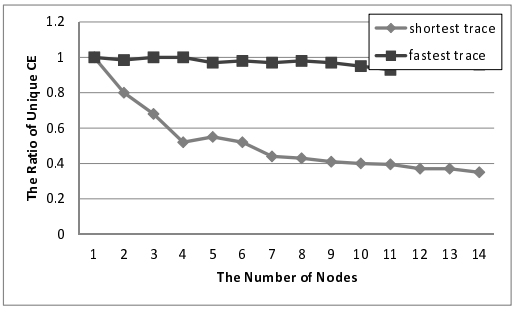


Figure 17: The Ratio of the Same Counter-Example : Gear Controller

the performance because it reduces the number of iteration. We also have to find further improvements such as detecting redundant counter-examples and reducing applies of counter-examples which do not contribute to refinement.

As a conclusion we can say that the main idea that we simultaneously apply the multiple counter-examples will improve the performance, however, there is some room to improve the performance.

6 CONCLUSION

This paper proposed a CEGAR loop for timed automata where multiple counter-examples are simultaneously applied. This device reduces the number of iteration loops strongly. The experiments show the promising results. Also we have obtained a candidate criterion for more effective multiple CEGAR.

Future work will be finding effective criteria for filtering better multiple counter-examples. Extension of the class of the property is also considered. For example, we want to try to provide CEGAR loop for some subset TCTL[1].

Acknowledgments

This work is partially being conducted as Grant-in-Aid for Scientific Research S (25220003) and also C (26330092).

REFERENCES

[1] R. Alur, C. Courcoubetis, and D. L. Dill. Model-checking for real-time systems. In *Proc. of the 5th Annual Symposium on Logic in Computer Science*, pages 414–425. IEEE, 1990.

[2] R. Alur, T. Dang, and F. Ivancic. Counter-example guided predicate abstraction of hybrid systems. In *Prc. of Tools and Algorithms for the Construction and Analysis of Systems TACAS 2003*, pages 208–223, 2003.

[3] J. Bengtsson and W. Yi. Timed automata: Semantics, algorithms and tools. In *Lecture Notes on Concurrency and Petri Nets*, volume 3098, pages 87–124, 2004.

[4] Patricia Bouyer, Francois Laroussinie, and Pierre-Alain Reynier. Diagonal constraints in timed automata: Forward analysis of timed systems. In *FORMATS'05*, volume 3829 of *LNCS*, pages 112–126, 2005.

[5] E. M. Clarke, A. Fehnker, Z. Han, J. Ouaknine, O. Stursberg, and M. Theobald. Abstraction and counterexample-guided refinement in model checking of hybrid systems. *Int. Journal of Foundations of Computer Science*, 14(4):583–604, 2003.

[6] E. M. Clarke, O. Grumberg, S. Jha, Y. Lu, and V. Helmut. Counterexample-guided abstraction refinement for symbolic model checking. *Journal of the ACM*, 50(5):752–794, 2003.

[7] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. MIT Press, 2000.

[8] A. E. Dalsgaard, R. R. Hansen, K. Y. Joergensen, K. G. Larsen, M. C. Olesen, P. Olsen, and J. Srba. opaal: A lattice model checker. In *Proceedings of the 3rd NASA Formal Methods Symposium (NFM'11)*, volume 6617 of *LNCS*, pages 487–493. Springer-Verlag, 2011.

[9] H. Dierks, S. Kupferschmid, and K. G. Larsen. Automatic abstraction refinement for timed automata. In *Proc. of the 5th Int. Conf. on Formal Modelling and Analysis of Timed Systems*, volume 4763, pages 114–129, 2007.

[10] D. Eppstein. Finding the k shortest paths. In *35th Annual Symposium on Foundations of Computer Science*, pages 154–165, 1994.

[11] F. He, H. Zhu, W. N. N. Hung, X. Song, and M. Gu. Compositional abstraction refinement for timed systems. In *Proc. 2010 Fourth International Symposium on Theoretical Aspects of Software Engineering*, pages 168–176, 2010.

[12] H. Hermanns, B. Wachter, and L. Zhang. Probabilistic cegar. In *Computer Aided Verification, Lecture Notes in Computer Science*, volume 5123, pages 162–175. Springer, 2008.

[13] V. M. Jimènez and A. Marzal. Computing the k shortest paths: A new algorithm and anexperimental comparison. In *AE 1999*, volume 1668 of *LNCS*, pages 15–29, 1999.

[14] S. Kemper and A. Platzer. Sat-based abstraction refinement for real-time systems. In *Proc. of the Third Int. Workshop on Formal Aspects of Component Software*, volume 182, pages 107–122, 2006.

[15] M. Lindahl, P. Pettersson, and W. Yi. Formal design and analysis of a gear controller: An industrial case study using uppaal. In *Lecture Notes in Computer Science*, volume 1384, pages 289–297, 1998.

[16] M. Oliver Mollera, Harald Rueß, and Maria Soreab. Predicate abstraction for dense real-time systems. *Electronic Notes in Theoretical Computer Science*, 65:218–237, 2002.

[17] T. Nagaoka, K. Okano, and S. Kusumoto. An abstraction refinement technique for timed automata based on counterexample-guided abstraction refinement loop. *IEICE Transactions on Information and Systems*, E93-D(5):994–1005, 2010.

Input/output Control Method for Serial Communication in the NC Equipment for Machine Tools

Akihiro Yamashita*, Hiroshi Mineno* , and Tadanori Mizuno**

*Graduate School of Science and Technology, Shizuoka University, Japan
**Faculty of Information Science, Aichi Institute of Technology, Japan
Yamashita.Akihiro@ma.mee.co.jp, mineno@inf.shizuoka.ac.jp, mizuno@mizulab.net

Abstract - New systems for manufacturing, such as the one presented by Industry 4.0 Smart Manufacturing, are being proposed to connect every component to the Internet to incorporate information from systems inside and outside the company, internal production systems, various kinds of control devices, and sensors.

In relation to the network connection and the distributed input/output control method, which are key technologies for IoT (Internet of Things), we report in this paper on a pioneering basic technology, which is the input/output control method for serial communication in the NC (Numerical Control) equipment for machine tools we conceptualized and adopted in our products in the 1990s. The basic idea is to modularize the components of the NC equipment, connect them through a network, and achieve distribution of the components to build a flexible system suitable for the purpose. The cumulative number of products shipped with this I/O control method incorporated exceeded 5,000,000.

Keywords: Industry 4.0, network, Internet of Things, distributed control, Numerical Control equipment

1 INTRODUCTION

New systems for manufacturing, such as the one presented by Industry 4.0 Smart Manufacturing, are being proposed to connect every component to the Internet to incorporate information from systems inside and outside the company, internal production systems, various kinds of control devices, and sensors. On the production sites, the application of such new systems also attracts attention. Efforts are made to incorporate all the relevant information into the manufacturing process to facilitate grasping of the current status or each step of the manufacturing process until completion. The relevant information covers the following: order information provided by the sales department and incorporated into the production system, parts orders made from the purchasing system, manufacturing/production instructions to the production sites, control information for machining or assembly, and in-process product identification information based on sensor technologies. In the field of parts machining, NC machine tools play a central role in IoT (Internet of Things), and the use of them makes it possible to establish a link between the parts machining information based on the manufacturing data (including part programs) and the sensor information obtained from the input/output (I/O) control.

In relation to the network connection and the distributed input/output control method, which are key technologies for IoT, we report in this paper on a pioneering basic technology, which is the I/O control method for serial communication in the NC equipment for machine tools we conceptualized and adopted in our products in the 1990s. The basic idea for this control method is to modularize the components of the NC equipment, connect them through a network, and achieve distribution of the components to build a flexible system suitable for the purpose. The cumulative number of products shipped with this I/O control method incorporated exceeded 5,000,000. It can be said that our idea is the precursor of today's NC system..

2 RELATED TECHNOLOGIES

2.1 Configuration and Functions of the NC Equipment

Figure 1 shows the configuration of the NC equipment. The NC system consists of the display and operation unit, control unit, drive units, servo motors, and spindle motor. The display and operation unit is used to create part programs, operate the machine, and display the part programs, machining status, and machine status. The control unit is used to analyze part programs, output the machine movement distance to the drive units as a movement command, and control the machine movement. The control unit is also used for sequence control for machining.

The drive units are used to control the servo motors for the tool nose path control and to control the spindle motor for rotating the tool to achieve the cutting work.

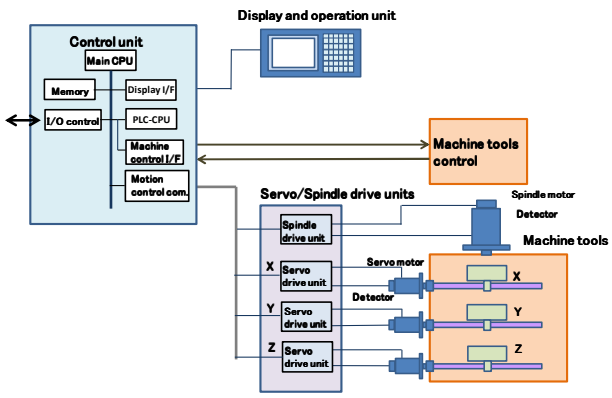


Figure 1: Configuration of the NC equipment

2.2 Path Control in the NC Equipment

The NC equipment executes part programs for driving the multi-axis machine tools (for example, X axis, Y axis, Z axis, etc. and the spindle) to move the tool nose position or move a part and cut the part

The NC control unit analyzes the part programs, adds the tool information to the analysis result, and calculates the tool path data. Then, the control unit calculates the movement distance per unit time for each control axis by an interpolating operation.

The position/speed commands calculated in the NC control unit are sent to the drive units. The drive units control the position/speed and the current for driving motors based on the position/speed command values sent by the control unit and the feedback information sent by the detector attached to each motor.

2.3 Sequence Control in the NC System

Besides the drive control, the NC equipment has a sequence control function to control the following auxiliary functions for the machining operation: monitoring of the machine status, commanding starting/stopping of the machining operation, replacement of the tool, turning ON/OFF of the cooling water supply, etc.

In the sequence control (I/O control function), input signals are used to monitor the machine status or to obtain the sensor information, and output signals are used to control the actuators.

The NC equipment controls the machine tools operations by calculating the movement command values at a fixed time interval and sending them to the drive units, implementing the sequence control of the machining procedure, and indicating status information on the display. In the sequence control, the control unit executes sequence programs. In response to the sensor information input signals sent in synchronization with the control cycle in the NC equipment, the machine control signals are output. It is essential to ensure real-time sequence control processing with an accuracy of 0.1 msec because the processing is synchronized with the positioning control cycle in the NC system.

2.4 Conventional Parallel I/O Control Method and Issues

The conventional I/O signal control is performed using a parallel I/O function of the control unit CPU. In this method, the machine status is checked using the parallel input signal from the control unit main CPU (ON/OFF control, +24 V/0 V), and the machine operations are controlled using the parallel output signal (ON/OFF control, +24 V/0 V). For this type of I/O control, several tens to one hundred signal wires are connected inside the power panel and externally for the sensor information input and the actuator control. Figure 2 shows the schematic block diagram.

In the parallel I/O signal method, each signal requires a wire connection between the controller and the actuator. Thus, the method involves problems such as the degradation

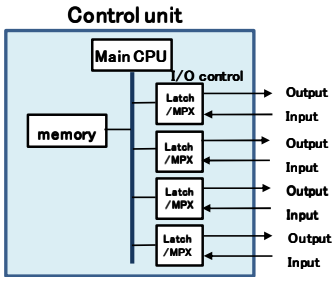


Figure 2: Parallel input/output.

of workability caused by the increase in the number of signals and wire bundles, and the lowering of reliability due to noises, etc. caused by the longer wire length.

Also, in the parallel I/O control method, data is directly transmitted between the control unit and the machines or actuators. Therefore, the control unit has to have numerous circuits to support the I/O control points according to the machine specifications. The control unit cabinet capacity cannot be selected flexibly without any restraints when the control unit has to contain the I/O control circuits.

2.5 Conventional I/O Control Method for Serial Communication and Issues

The NC control unit and the I/O control part are connected using serial communication. By transmitting the I/O signals for controlling the machine tools through communication, it is possible to suppress the problems of the above-mentioned wire connection or the lowering of reliability due to noises, etc.

In the 1990s, CPUs were in most cases mounted on both the input and output devices to establish serial communication. Figure 3 shows an example of I/O control for a regular serial communication. The remote I/O unit has a large circuit because CPU and memory dedicated to communication, communication control circuit, and I/O control interface circuit are required. Therefore, the method incurs higher costs as compared to the parallel I/O method. In addition, software processing is required on the remote I/O unit side.

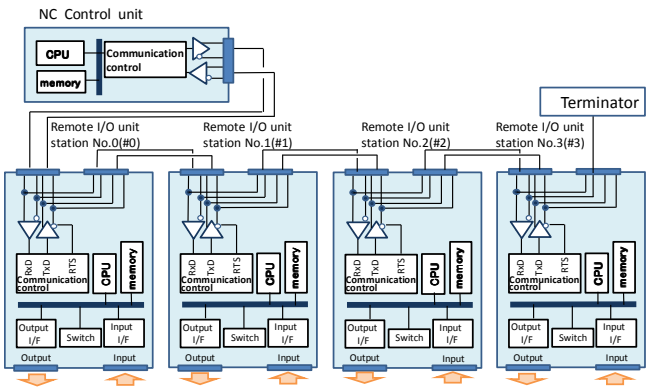


Figure 3: I/O control for serial communication.

On the other hand, communication software processing is required on the NC unit side. When communication software processing is added to the NC system in order to complete the real-time processing within a certain time limit, a high-performance CPU is required for the NC unit, which results in increased cost.

3 PROPOSED METHOD

3.1 Distributed Remote I/O Control Method

For the system where the control unit and multiple distributed remote I/O units are connected using serial communication, the distributed remote I/O control method introduced in this paper enables downsizing of the system, improves reliability, lowers costs, and enhances the safety by reducing the number of signal communication wires, eliminating the CPU from the remote I/O unit, and adopting a hardware mechanism for the fail safe function. Figure 7 shows a configuration of the NC system using the distributed remote I/O control method.

In Figure 4, the machine control interface of the control unit is used as the distributed remote I/O communication part. The I/O control circuits are separated from the control unit and contained in the distributed remote I/O units. The machine requiring numerous I/O control points can be supported by increasing the number of distributed remote I/O units. Consequently, the control unit cabinet capacity is no more dependent on the number of I/O control points.

3.2 Distributed Remote I/O Communication Procedures

A half-duplex method is adopted for serial communication between the control unit and the distributed remote I/O units. As compared to a full-duplex method, the number of communication wires can be reduced to one, and the wire rod, connector, and cable manufacturing costs can be reduced. In order to establish highly reliable data communication with the half-duplex method, a dedicated time-dividing communication procedure is defined to fit the characteristics of the NC equipment.

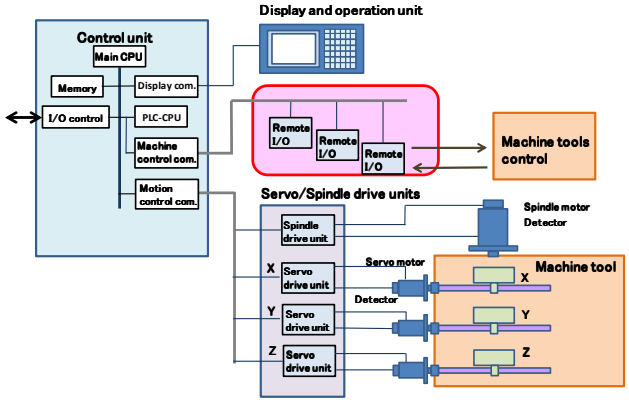


Figure 4: NC equipment configuration using the distributed remote I/O control method.

- Adoption of the EIA-485 differential system for the communication physical layer (data communication circuit)
- Adoption of the HDLC communication method
- In order to establish half-duplex communication, the communication cycle order is determined, and data transmission is enabled only during the transmission period.

In order to perform two-way serial communication between the NC control unit and multiple distributed remote I/O units, the NC control unit performs a time dividing communication with each unit. The following two modes are defined for communication: the offline status communication mode, and the online communication mode (normal I/O mode). The two modes can be distinguished by the difference in the frame header pattern.

Figure 5 shows the communication processing flow of the NC control unit which controls the distributed remote I/O units. At power-on, communication starts automatically in the offline status mode between the communication control part of the NC control unit and the communication control part of each distributed remote I/O unit. The status information of the remote I/O units is set in the communication control part of the NC control unit.

Based on the status information of the distributed remote I/O units stored in its communication control part, the NC control unit generates output data frames in the online communication mode and sends the frames to the distributed remote I/O units. The NC control unit then receives data frames sent from the distributed remote I/O units. When no error is found with the data frames, the NC control unit processes the input data.

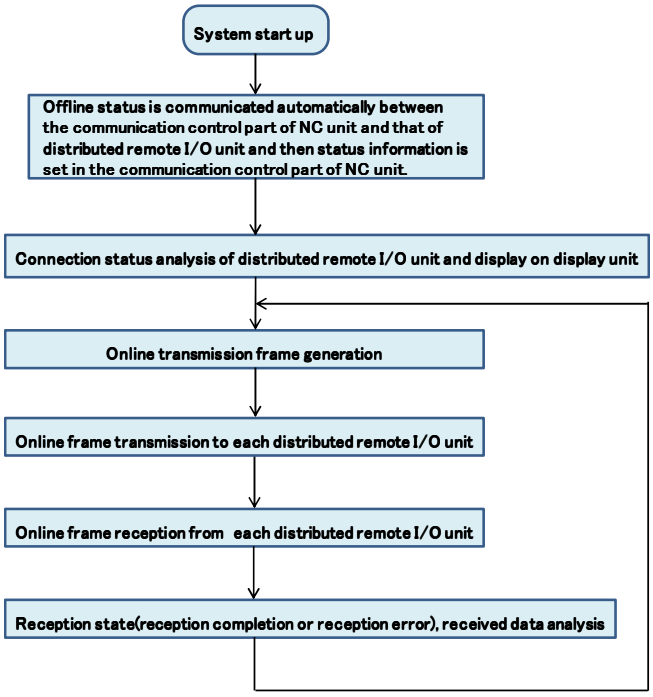


Figure 5: Communication processing flow chart on the NC control unit side

When the count for errors found with frames sent from the distributed remote I/O units exceeds a predetermined value, the NC control unit determines that a system error has occurred and stops the system..

4 IMPLEMENTATION SCHEME

4.1 Circuitry to Implement the Distributed Rmote I/O Control Method

Figure 6 shows the system and circuit configuration of the NC equipment for the distributed remote I/O control method.

In Figure 6, I/O data is transmitted from the communication control part of the NC control unit to the distributed remote I/O units using serial communication. To simplify the hardware circuit, serial communication processing is performed in the hardware circuit without using CPUs. Each hardware circuit is integrated into a one-chip LSI in the communication control part either on the NC control unit or the distributed remote I/O unit side. Also, half-duplex communication is established by controlling a communication control signal (RTS signal) to achieve reduction in the number of communication signal wires.

4.2 Implementation of the Distributed Remote I/O Control Method

Figure 7 is the timing chart for transferring of communication frames between the NC control unit and each distributed remote I/O unit, and shows the data frame configuration.

The NC control unit sequentially sends data frames to multiple distributed remote I/O units in a time dividing manner, and each remote I/O unit sends a data frame to the NC control unit after a predetermined time period. When the maximum number of stations is defined in advance for connecting distributed remote I/O units, the NC control unit can sequentially send data to each distributed remote I/O unit, and receive the data sent from the remote I/O unit. In this control method, the NC control unit can complete the data communication with the distributed remote I/O units within a one-cycle interval. By repeating this procedure, the NC control unit can perform the real-time processing of the data I/O with a fixed cycle.

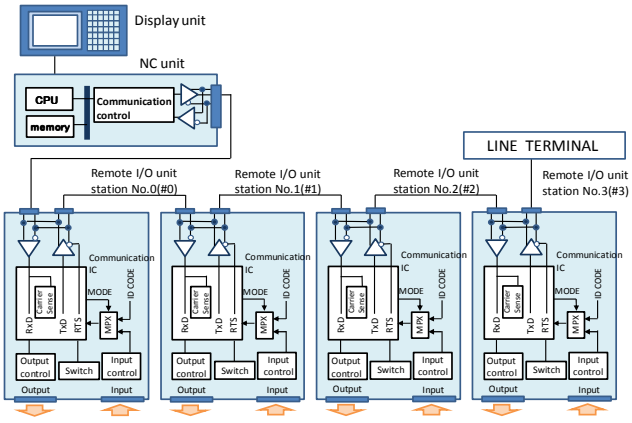


Figure 6: Distributed remote I/O circuit configuration.

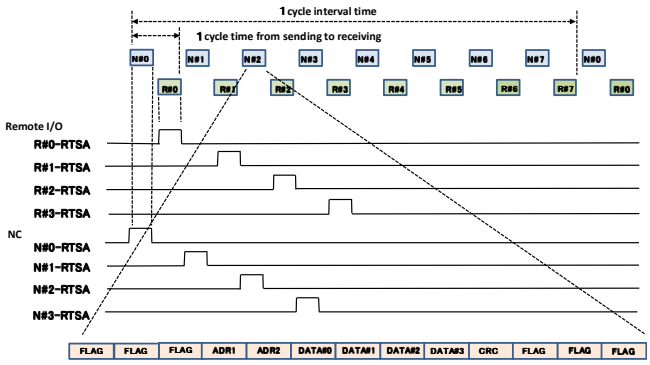


Figure 7: Distributed remote I/O control timing chart..

Figure 8 shows I/O circuits of a remote I/O unit. When the NC unit sends data in a data frame to a distributed remote I/O unit, the data is then output from the I/O unit as its output signal. Then, when the distributed remote I/O unit takes in an input signal, the I/O unit sends the data in a data frame to the NC control unit, and the data is then used as the input data in the NC control unit. When each bit of the I/O data has its own meaning, the normal I/O control is executed.

4.3 Implementation of the Distributed Remote I/O Communication Procedure

4.3.1 Offline Communication Procedure

Figure 9 is the processing flow chart of the distributed remote I/O unit.

(1) After initialization at power-on, the distributed remote I/O unit reads its own station number from the hardware setting (setting switch).

(2) After initialization at power-on, the communication control part of the NC control unit sequentially requests each distributed remote I/O unit in the offline status communication mode in a time dividing manner to send the type and setting information of the remote I/O unit.

(3) The offline status communication mode is the initial setting for the distributed remote I/O units.

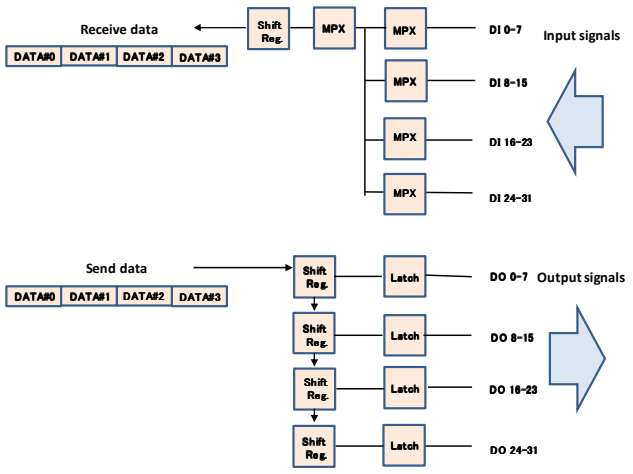


Figure 8: Input/output of the distributed remote I/O data.

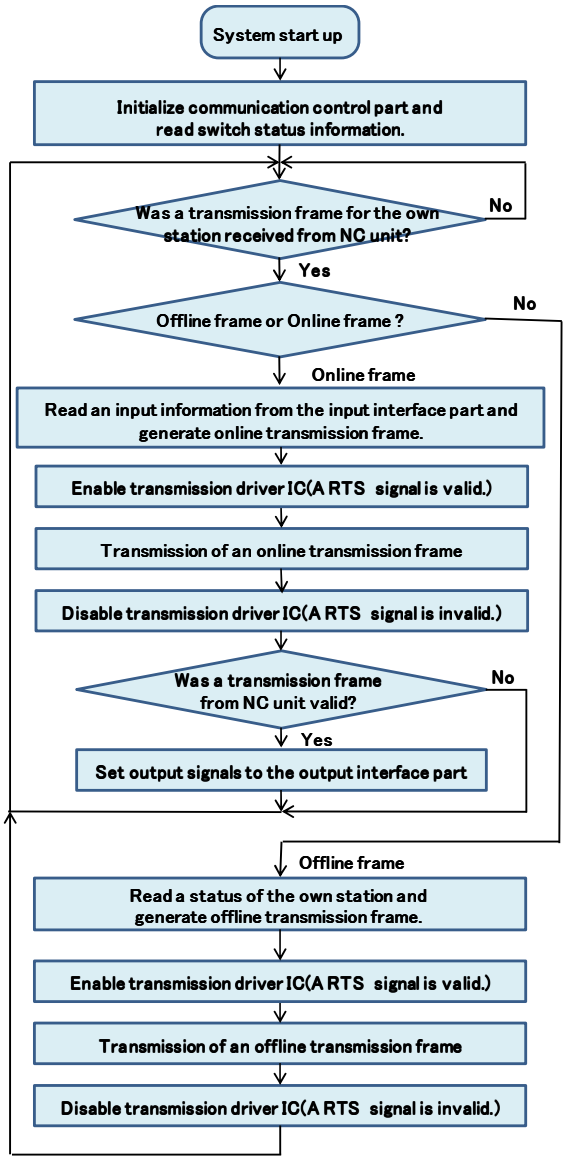


Figure 9: Distributed remote I/O processing flow chart.

(4) The distributed remote I/O units distinguish the offline status communication mode from the online communication mode, which is the normal I/O mode, by the difference of the header pattern (address data part) of the data frame sent from the NC control unit.

(5) In the offline status communication mode, when the distributed remote I/O unit detects its own station number in the data frame header pattern and receives the data addressed to its own station, the remote I/O unit determines the communication control signal RTS as valid after a predetermined time period, and sends data (remote I/O unit type and setting information) to the NC control unit.

(6) The communication control part of the NC control unit sequentially sends requests to obtain the type and setting information from the remote I/O unit in a time-dividing manner, and obtains the information on the number of stations and the unit type for all the remote I/O units connected with the one communication line.

(7) When the NC control unit obtains the hardware information of the remote I/O unit, the NC control unit switches the mode from the offline status communication mode to the online communication mode.

4.3.2 Online Communication Procedure

(1) In the online communication mode, the NC control unit periodically sends data to each distributed remote I/O unit. Each remote I/O unit checks the header pattern of the data sent from the NC control unit, and receives the data only when the header pattern corresponding to its setting switch exists.

(2) When the remote I/O unit detects an error in the received data frame, the remote I/O unit does not update the output signal. The remote I/O unit changes the header pattern of the data frame and sends the data frame to the NC control unit to let the NC control unit recognize the data frame error.

(3) The status signal of the machine tools or actuators is input to the distributed remote I/O unit, and the remote I/O unit sends the data frame to the NC control unit after a predetermined time period after receiving the data mentioned in (1).

(4) By repeating steps (1), (2), and (3) above, the NC control unit performs sequence control for the machine tools.

4.4 Implementation of Safety and Reliability

In the distributed remote I/O control system, communication processing is performed with a fixed cycle in a CPU-less hardware configuration, and the following operations are performed to ensure communication reliability and safety of machine tools.

(1) When the remote I/O unit does not detect any data frame sent from the NC control unit for a predetermined time period or longer, the remote I/O unit automatically resets its own output.

(2) When the remote I/O unit detects an error in the received data frame, the remote I/O unit does not update the output signal. The remote I/O unit changes the header pattern of the data frame and sends the data frame to the NC control unit to let the NC control unit recognize the data frame error.

(3) The NC control unit stops the system when it does not receive the data frame sent from the remote I/O unit.

(4) The NC control unit stops the system when the count for errors found with frames sent from the remote I/O units exceeds a predetermined value.

5 EVALUATION

5.1 Downsizing of the Control Unit and the Remote I/O Units

In the distributed remote I/O control method, a half-duplex communication method is adopted to simplify the communication wiring/circuit configuration. In addition, each hardware circuit is integrated into an LSI in the

communication part either on the NC control unit or the distributed remote I/O unit side, achieving downsizing. Figure 10 shows a comparison of cabinet capacity by NC generation

By adopting the distributed remote I/O control system, the I/O control circuits can be separated from the control unit. Therefore, the NC control unit can be significantly downsized owing to the high integration of the hardware circuits. The control unit capacity is 20% smaller than the conventional one. Instead, the I/O control circuits are contained in the remote I/O units. Owing to the integration of the additional circuits into an LSI in the communication control part, the remote I/O unit capacity is 50% smaller than the conventional communication control part.

By adopting the distributed remote I/O control method in the products, the system capacity of the control unit and the remote I/O unit is 25% smaller than the conventional one, which contributes to the downsizing of the cabinet.

5.2 Flexible System Configuration

The number of I/O signal control points depends on the machine tools configuration. In the conventional NC control unit, the number of I/O control points of each NC control unit is fixed for its hardware configuration. By adopting the compact-sized distributed remote I/O units in the products and configuring the system using multiple units as required, it was made possible to configure the I/O control part of the NC equipment flexibly to support the number of I/O control points required by the machine tools. Also, it is possible to provide a variety of functions. Other than for normal I/O signal control, it is possible to use distributed remote I/O units for the data output to the display, the pulse generator function, or the analog voltage output and input to and from the peripheral devices. Thus, the NC control unit supports the I/O type or the number of points of the machine to be controlled.

The conventional NC equipment requires wire connections from one power panel to all machine components in the system. By using distributed remote I/O units, the distributed arrangement of the NC control unit, driving amplifiers, and remote I/O units is made possible according to the configuration of the machines.

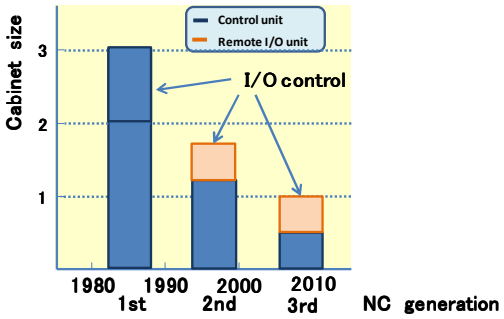


Figure 10: Comparison of cabinet capacity by NC generation

5.3 Reliability and Safety of the System

5.3.1 Reliability at Power-on

We evaluated whether the NC control unit recognizes distributed remote I/O units connected to the NC control unit, and how the system operates when misrecognition occurs after initialization at power-on.

After initialization at power-on, the NC control unit requested information on the type and settings of the remote I/O units in the offline status communication mode. After this hardware information was obtained, the NC control unit switched the mode to the online mode. In the above operation, when the actual machine information was not consistent with the NC control unit setting information, inconsistency was regarded as an alarm and the NC equipment operation did not start. As a result, we confirmed that the actual system configuration is checked without fail when the system mode transfers to the normal online mode from the initial status after power-on, and the system is highly reliable in preventing malfunction when a connected device is misrecognized.

5.3.2 Reliability and Safety at Fault Conditions

In terms of communication functions, we evaluated how the reliability and safety of the system can be assured in case of communication failures between the control unit and the distributed remote I/O units.

When a communication break or cable disconnection occurred, no data frame was sent from the remote I/O units, the NC control unit determined that a communication break or cable disconnection was occurring, and stopped the system. When noise was present and the count for errors in the data frame sent by the remote I/O module exceeded a predetermined value, the NC control unit also stopped the system. We confirmed that the NC control unit assures system safety by monitoring the data frame sent from the remote I/O units.

When an incident such as a cable disconnection or break occurred, the remote I/O unit reset the machine control signal output since it did not detect a data frame sent from the NC control unit for a predetermined time period. We confirmed that the machine control signal can be reset even when the NC control unit enters an error status and stops the system. Also, when the remote I/O unit detected an error in the received data frame, the remote I/O unit held and did not update the machine control signal output.

We confirmed that the above measures ensure the safety of the highly reliable system against abnormal stops of the NC control unit or communication failures such as communication breaks.

5.4 System Cost Reduction

5.4.1 Simplification of the Hardware

In the distributed remote I/O control system, a half-duplex method using a single signal wire is adopted for

communication. Therefore, costs involved with serial communication can be reduced for wire rods, connectors, etc. Also, the communication part is configured using hardware circuits without CPU. The circuit cost can be reduced by using communication ICs we developed for LSI. In the 1990s, the use of ASIC including CPU or user circuits was not practical yet. It was beneficial to configure the communication control circuits on the hardware because less parts costs were required for CPUs, memories, etc.

5.4.2 Simplification of the Software Processing

In the online communication mode, the NC control unit writes the control command and parameters in the send buffer of the communication part. After a predetermined time period, the NC control unit reads the receive buffer of the communication part. It was possible to simplify the software processing because the information of the external devices can be easily read through the distributed remote I/O units without much concern for the serial data communication.

5.5 Widespread Use of the Distributed Remote I/O

The NC equipment using the distributed remote I/O control method started to be shipped to the market in the late 1990s. The hardware circuits of the communication part, a core of the distributed remote I/O control, were integrated into a one-chip LSI in the communication part. Two types of LSIs were developed: one with a master function is equipped on the NC control unit, and the other with a slave function is equipped on the distributed remote I/O units.

Owing to the subsequent development of semiconductor technologies, the LSI of the communication control part is also used as the intellectual property (IP) core of FPGA or ASIC. Figure 11 shows the annual shipment amount of the LSIs for the communication part with the master or slave function and the IP cores. The annual shipment amount in 1997 when the shipment started is used as reference. The distributed remote I/O control method is now widely available in the market. The most recent annual shipment amount of products is more than 10 times of that of the first year of shipment, and the cumulative number exceeded 5,000,000.

6 CONCLUSION

We reported on the distributed remote I/O control method in the NC equipment. By utilizing network, we modularized the components of the conventional NC equipment according to their functions, and furthermore, optimized the individual modules. Thus, we achieved distribution of the component modules and improved flexibility of system configuration while maintaining high reliability and safety of the system. When serial communication started to be widely available in the 1990s, the basic idea of downsizing and distribution of the modules by a network herein was a pioneering solution. The concept of the idea is the origin and the inheritance for the today's NC equipment. We anticipate further contributions of the method reported herein to the development of manufacturing technologies for IoT.

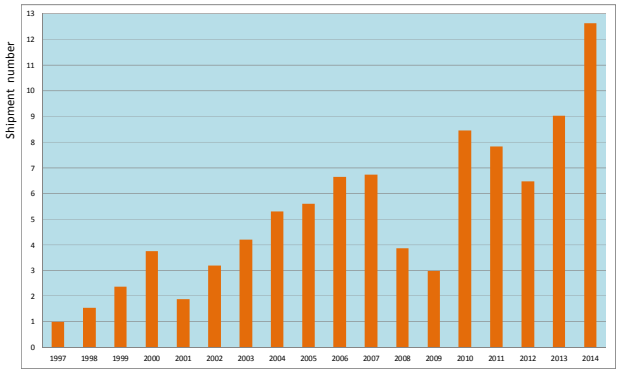


Figure 11: Cumulative amount of communication part LSIs/IPs shipped

REFERENCES

[1] Fei Tao, Ying Zuo, Li Da Xu, and Lin Zhang, IoT-Based Intelligent Perception and Access of Manufacturing Resource Toward Cloud Manufacturing, IEEE Tran. Industrial Informatics, Vol.10, No.2, pp.1547 – 1557(2014)

[2] Fei Tao, Ying Cheng, Li Da Xu Lin Zhang, and Bo Hu Li, CCIoT-CMfg: Cloud Computing and Internet of Things-Based Cloud Manufacturing Service System, IEEE Tran. Industrial Informatics, Vol.10, No.2, pp.1435 – 1442(2014)

[3] Lixin Du, Chunsun Duan, Shijun Liu, and Wei He, Research on service bus for Distributed Real-time Control Systems, IEEE/ITAIC , 2011 6th IEEE Joint International,Vol1, pp.401-405 (2011).

[4] Sauer, O., Developments and trends in shopfloor-related ICT systems, IEEE/IEEM, 2014 IEEE International Conference on, pp. 1352 - 1356 (2014).

[5] Jiafu Wan, Hu Cai, and Keliang Zhou, Industrie 4.0: Enabling technologies, IEEE/ICIT, 2014 International Conference on, pp. 135 - 140 (2015).

[6] Shrouf, F.; Ordieres, J.; Miragliotta, G., Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm, IEEE/IEEM, 2014 IEEE International Conference on, pp. 697 - 701 (2014).

[7] Decotignie, J.-D., Gregoire, J.-C., Integrating the numerical controller and the FMS, IEEE/IECON, 15th Annual Conference of IEEE , Vol.3. , pp. 675 - 680 (1989).

Keynote Speech 3:
Mr. Shinichi Baba
(Telecommunications Research Laboratory,
Toshiba Europe Research)

TOSHIBA
Leading Innovation

Toward Smart Community,
- Future Communications for an evolving energy, healthcare and other infrastructure systems

September 2015

S Baba and M Sooriyabandara
Telecommunications Research Laboratory
Toshiba Research Europe Ltd.

Index

Introduction

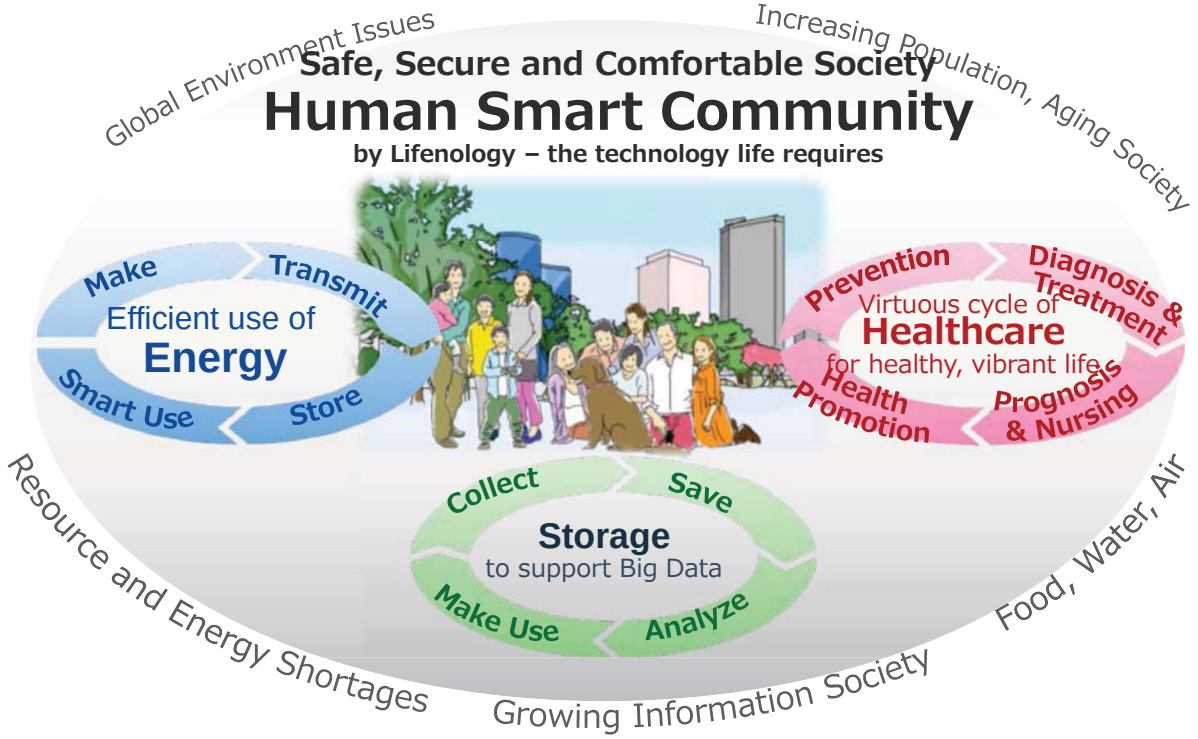
- **Challenges in Europe (and advanced countries…)**
  - Elderly society, Healthcare
  - Greenhouse effect, Energy Demand & Supply
  - Clean & Smart transportation
  - Climate change & Disasters
  - Foods
- **Solution: Smart community**
  - System and Infrastructure are empowered by ICT
  - More efficient and comfortable service

Priority 3. Societal Challenges funding
(€ million, 2014-2020)

Health, demographic change and wellbeing	7 472
Food security, sustainable agriculture and forestry, marine and maritime and inland water research and the Bioeconomy	3 851
Secure, clean and efficient energy *	5 931
Smart, green and integrated transport	6 339
Climate action, environment, resource efficiency and raw materials	3 081
Inclusive, innovative and reflective societies	1 310
Secure societies	1 695
Science with and for society	462
Spreading excellence and widening participation	816

* Additional funding for nuclear safety and security from the Euratom Treaty activities

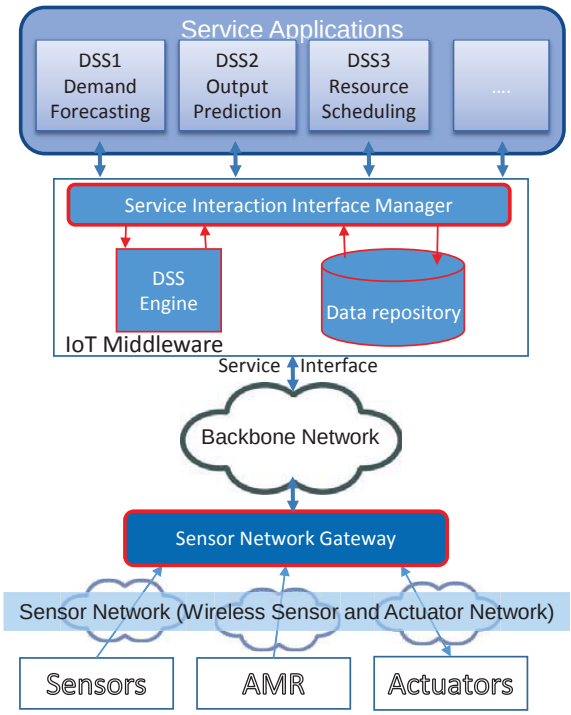
Future Vision for Smart Community



Smart system architecture

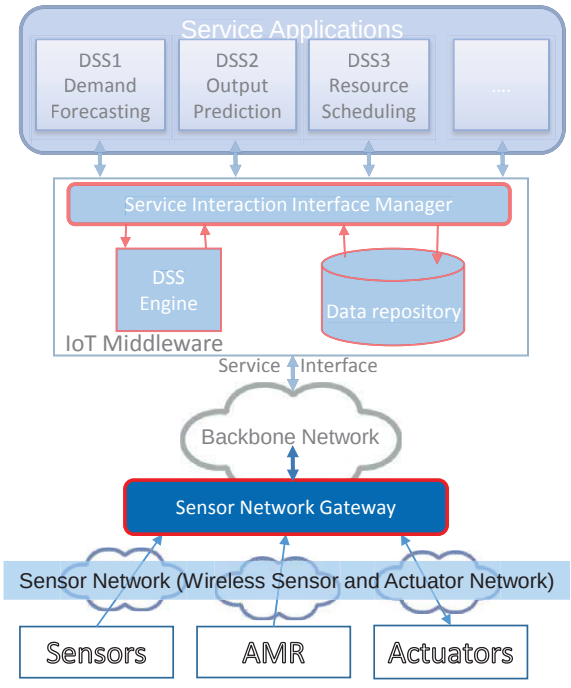
- **Service application**
- **Data repository**
  - Integration of sensor data
  - Unified data access scheme
  - Data access management
- **Sensor network**
  - Many communication standards
  - Wired and Wireless
  - Gateway provides flexibility and interoperability
- **Sensor and actuators**
  - Radically increasing in number
  - Existing vs New deployment

DSS: Decision Support Systems
IoT: Internet of Things
AMR: Automatic Meter Reading



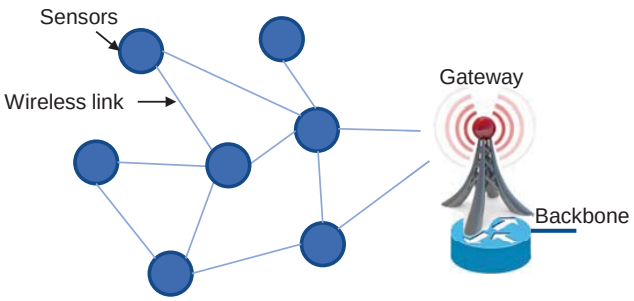
Challenges to Communication system

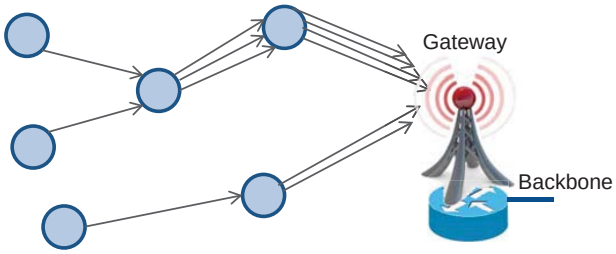
- **Wireless Everywhere**
  - Wireless control
    - High reliability
    - Energy efficiency
- **Network function optimisation**
  - Redefine
    - Content and information
  - Relocate
    - Distributed Architectures
  - Virtualization
    - Computing in data centre
    - Networking
- **Integrated approach and system of systems**
  - Cross-domain interactions
  - Convergence at Horizontal layer



Sensor network

- **Wireless Mesh network**
  - Sensor with radio capability forms a network
  - Automatic discovery of neighbor sensor
- **Gateway issue**
  - Focal point of data
  - Communication congestion increases around the gateway
    - Data: delay and drop
    - Sensors: more energy consumption for relaying data

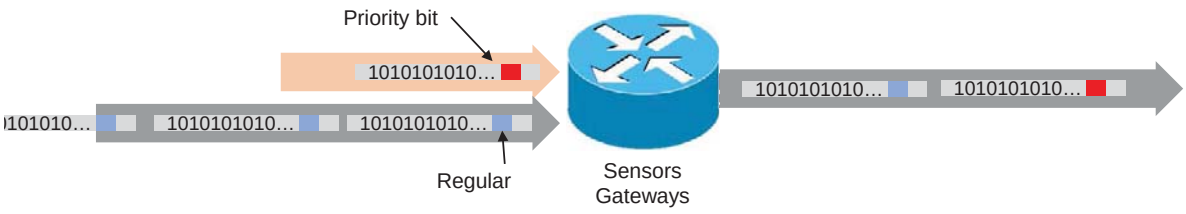




Improve Reliability

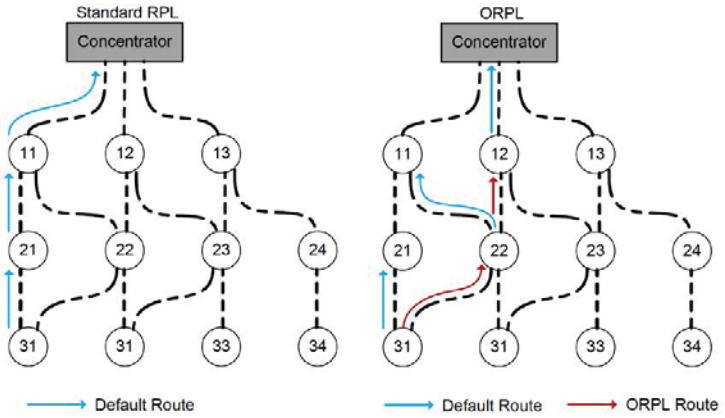
Quality of Service

- **Prioritized data is processed instantaneously**
  - Priority bit with data
  - Prioritized transmission in a radio system
  - Higher reliability (lower chance of drop)
  - Lower latency
- **Challenges**
  - Resource management
  - Resource limitation



Improve Reliability

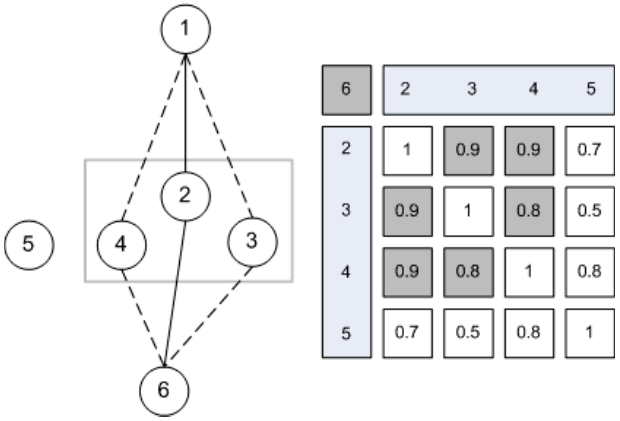
Opportunistic routing



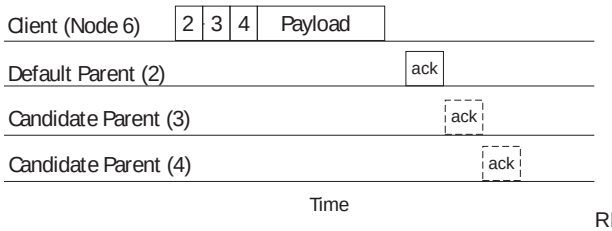
- **Enables sensors to select multiple candidate relays (other sensors)**
  - Improve resilience from a drop of wireless link to one relay
- **Key Questions**
  - How to select the best relay that reduces the transmission cost?
  - How to coordinate the relay set so as to reduce/eliminate repeated transmissions from the relaying nodes and therefore the overhead?

Example of Opportunistic routing

Opportunistic RPL



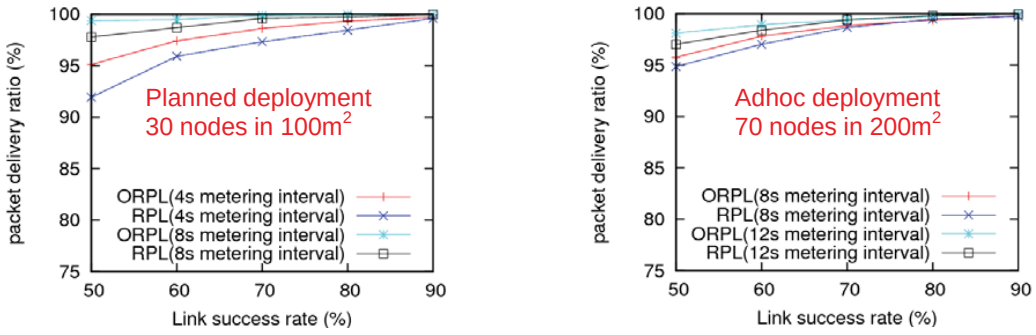
	2	3	4	5
2	1	0.9	0.9	0.7
3	0.9	1	0.8	0.5
4	0.9	0.8	1	0.8
5	0.7	0.5	0.8	1



- **Each node reports neighbourhood map via RPL scheme**
- **Node chooses candidate parent set based on this info.**
- **Frame sent to all candidate parents**
- **If transmission to default parent fails, the others will try to relay it**

RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks

Performance



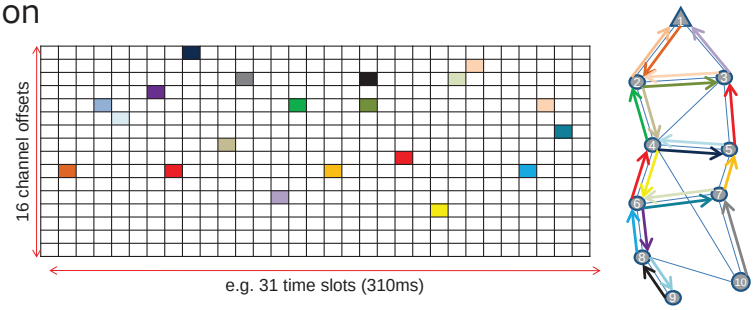
- **ORPL performs better than RPL in terms of Packet Delivery Ratio**
  - Under unsaturated conditions,
    - packet drops mainly due to channel errors
    - Poorer the link, more the link layer losses
    - ORPL provides diversity and thus improves performance

Improve Reliability

Time-Slotted Channel Hopping (TSCH)

- **Industrial applications such as factory automation, disaster defence, and security surveillance etc.**
  - WirelessHART
  - ISA100.11a
  - IEEE 802.15.4e

- **The latest IEEE 802.15.4e (TSCH) MAC represents the next generate on of high reliable and low power MAC protocol.**
  - Time slotted access and Node synchronization
  - Multi-channel communication
  - Channel hopping

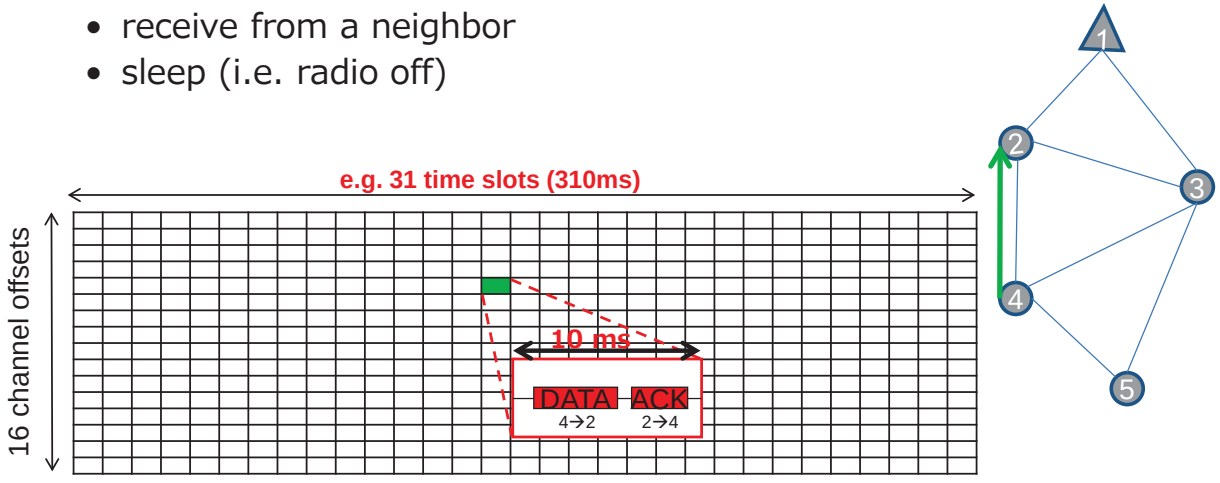


Some Application Areas for 802.15.4e

- ***Control loops*** in a wireless process control network, in which high reliability and a fully deterministic behavior are required
- ***Umbrella networks*** transporting data from different independent clients, and for which an operator needs flow isolation and traffic shaping
- ***Energy harvesting*** **networks,** which require an extremely low and predictable average power consumption
- ***Widespread monitoring*** such as corrosion monitoring or pipe leak detection, which requires a large number of sensors slow periodic reporting rates and open loop operation.

TSCH Schedule

- **All nodes in a TSCH network always keep synchronized**
- **Time is divided in slots, grouped in a *slotframe (i.e. CELL),* which continuously repeats over time (slotframe length tunable).**
- **A TSCH schedule indicates what to do in each cell:**
  - transmit to a neighbor
  - receive from a neighbor
  - sleep (i.e. radio off)



Channel Hopping

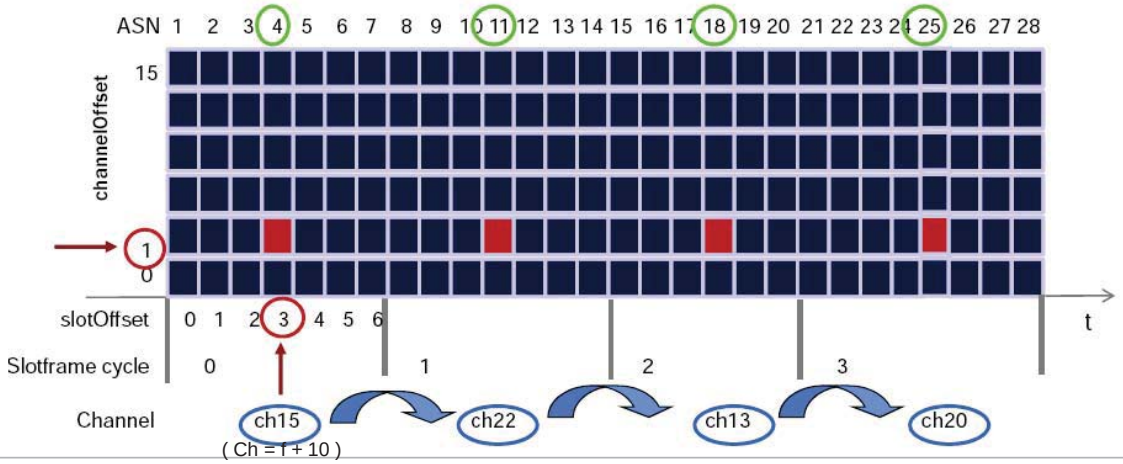
Table I. FrequencyTranslation

k	ASN	$chOf$	f
0	4	1	5
1	11	1	12
2	18	1	3
3	25	1	10

- **Channel offset is translated in an operating frequency f**

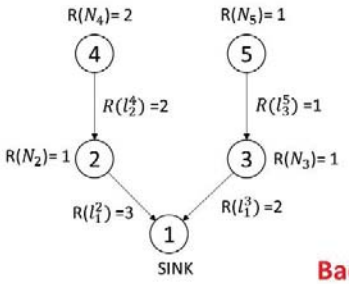
$$f = F\{(ASN + channelOffset)\ mod\ n_{ch}\}$$

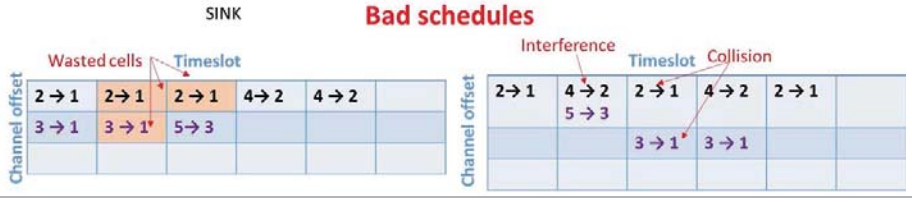
  - *ASN*: total # of slots that elapsed since the network was deployed
  - *ASN=(k·S+t) where S is the slotframe size, k the slotframe cycle*
  - n_{ch} *: number of used channels*
  - *F* is implemented as a look-up-table containing the sets of available channels



Scheduling

- **IEEE 802.15.4e defines the mechanism of how the TSCH schedule operates in the network**
- **Schedule design**
  - Centralized vs Distributed
  - Real-time vs Pre-calculation
  - Active vs Sleep





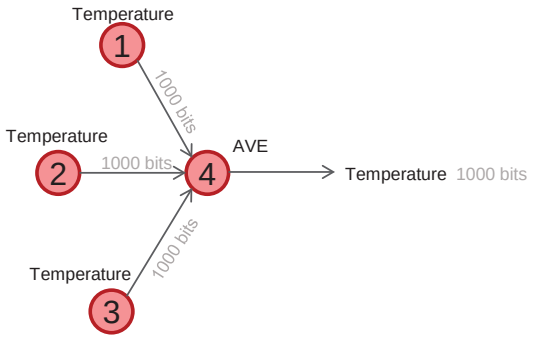
Content Centric Communication

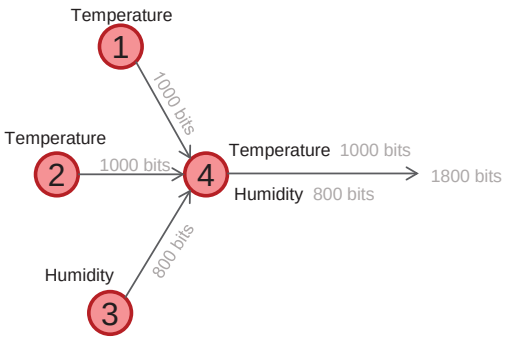
- **Communication management technologies to improve reliability in the sensor network**
- **But, issues still exist under vast IoT data communications**
  - Amount of sensors is increasing rapidly
  - Data from IoT networks including images and videos
- **Correlated data and image generated from sensors**

Content Centric

Content Centric Communication

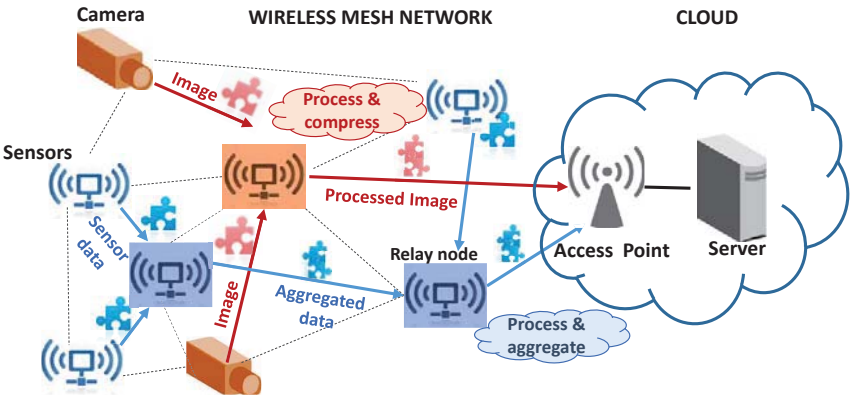
- **Data aggregation model**
  1. Only correlated data generated by the same application can be aggregated.
  2. Depends on data accuracy requirement, different aggregation ratios can be applied.





Content Centric Routing

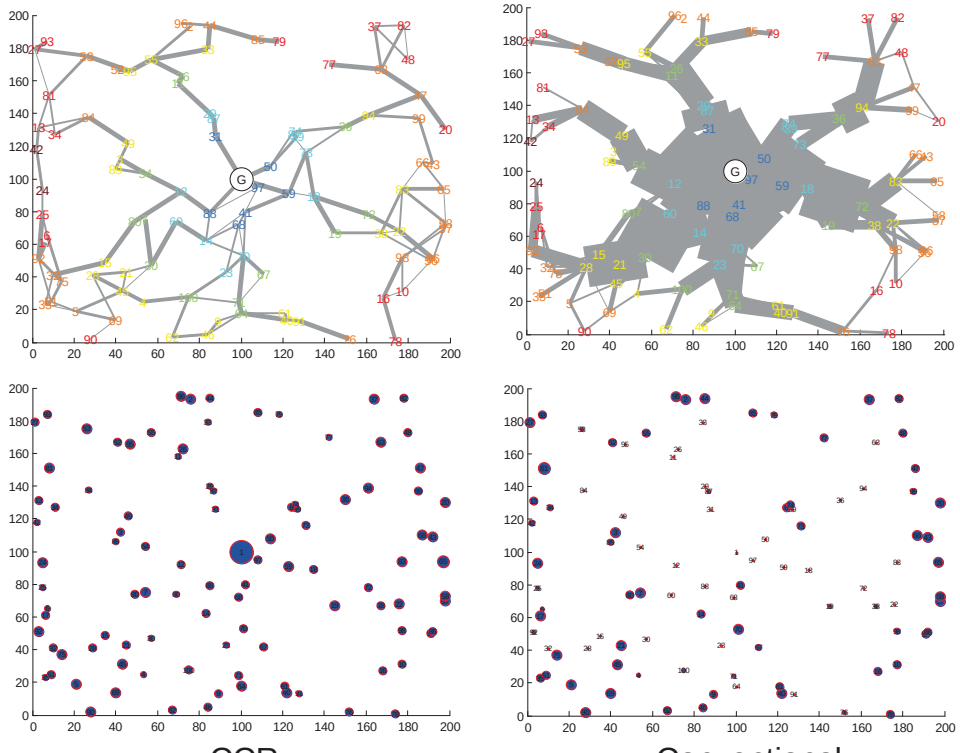
- **A content-centric and load balancing aware dynamic data forwarding**
- **Independent routing decisions are made by each node using only local information with the aim of**
  - Reducing the communication traffic by aggregating similar type of data, hence increasing the processing gain.
  - Balancing the energy-consumption among neighbouring nodes to extend the network lifetime.



Results : traffic and residual energy maps

Traffic map:

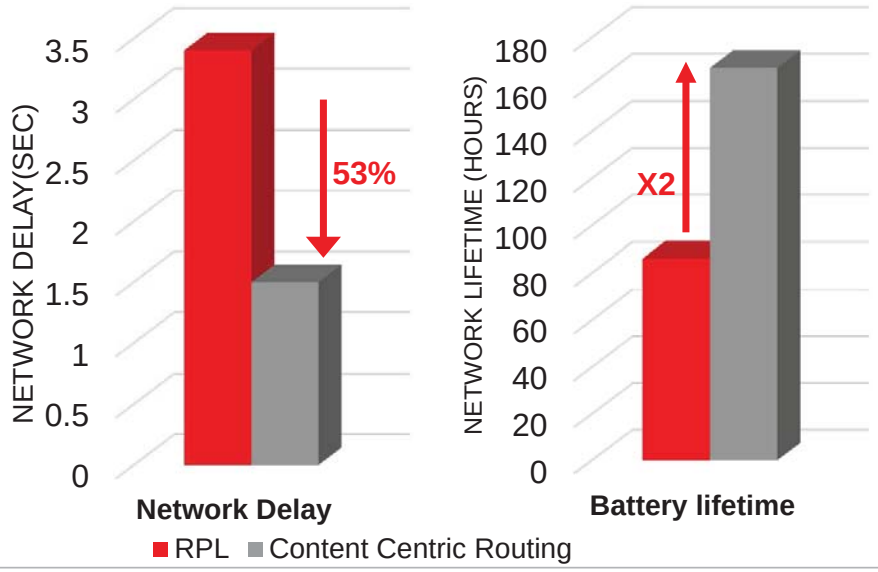
Energy map:



CCR Conventional

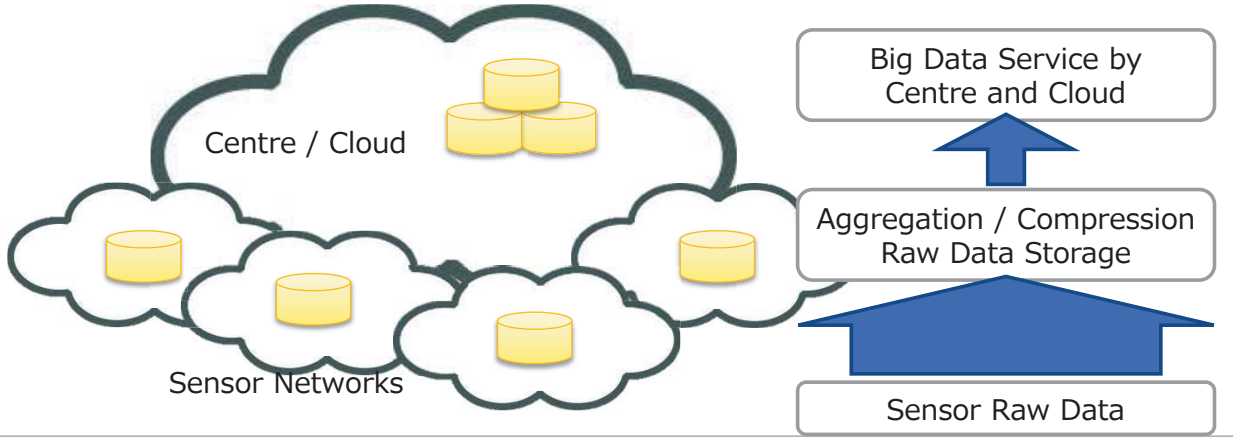
Results :

- **Distributed computing can avoid making redundant transmissions.**
- **The developed algorithm shows *100% lifetime improvement***



ICT integration

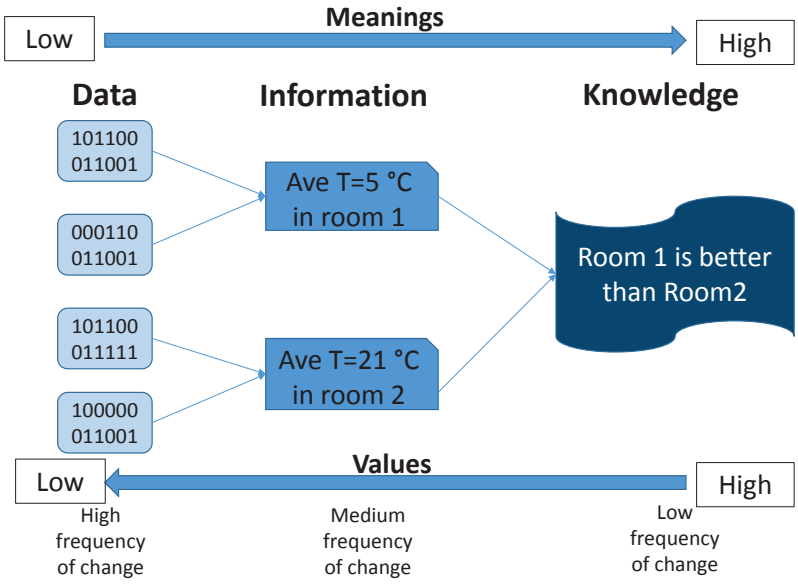
- **Efficiency optimization among Communication, Processing and Storage**
- **Further processing and storage capabilities in the sensor network**
  - More sophisticated data aggregation and compression in the sensor network
  - More efficiency and reliability for communication
- **Data for Centre and Cloud services**
  - Aggregated data, which saves storage and communication cost
  - Raw and precise data, by request, if it is necessary



ICT integration

Knowledge Centric

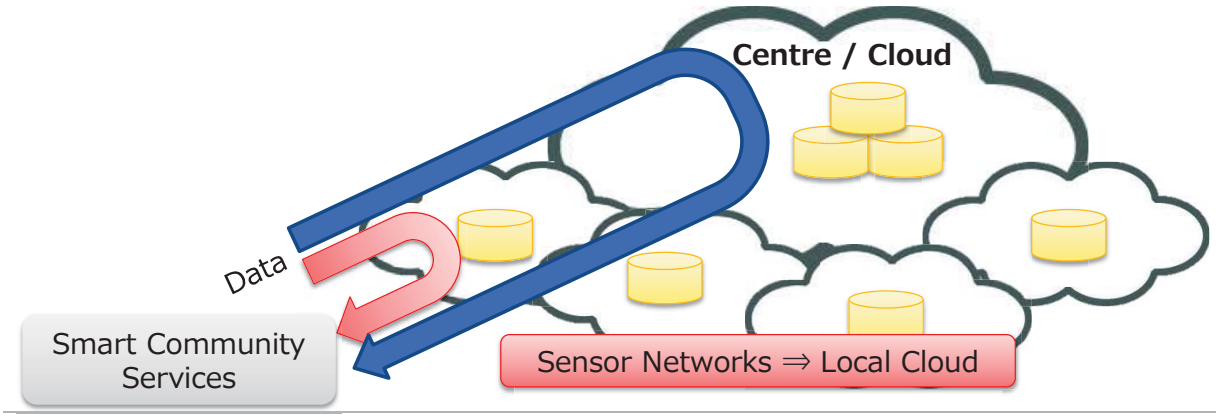
- **Knowledge extraction in the sensor network**



ICT integration

Mobile edge computing

- **Evolution of Sensor Network to a Local Cloud**
  - Two or more ways of providing a service
  - Quicker response by a service at the local cloud
  - For Centre / Cloud:
    - Protect from data and communication explosions

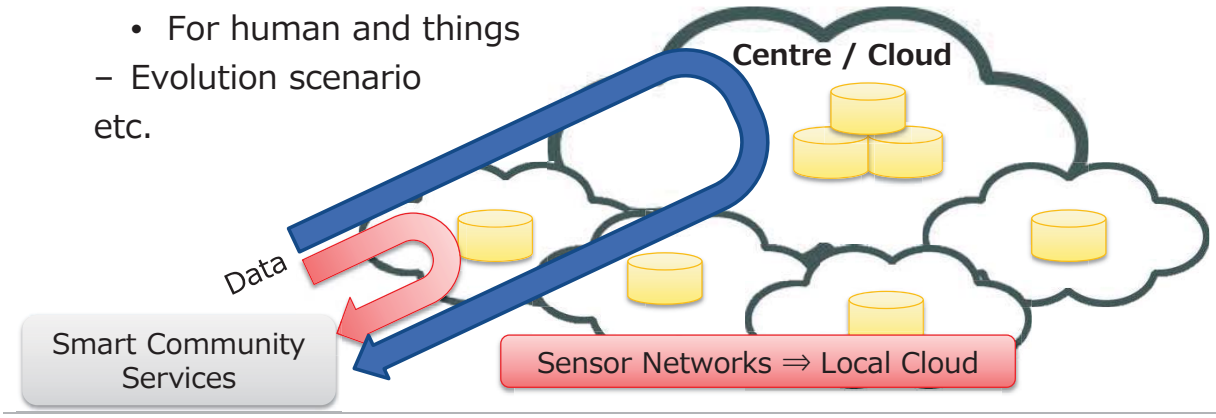


ICT integration

Mobile edge computing

- **Challenges**
  - Unified design in architecture
    - Service allocation
    - Data allocation
    - Service centric communication
  - Multiple service support
    - For human and things
  - Evolution scenario

  etc.



Summary

- **For future smart community system, the sensor network shall be improved further in,**
  - Capacity,
  - Latency,
  - Reliability,
  - Power efficiency,
  - Security,

- **Information and Communication technology integration is a key challenge for the sensor network evolution, and the smart community**

References
1. http://futurecity.glasgow.gov.uk/
2. http://europa.eu/rapid/press-release_MEMO-13-1122_en.htm
3. S. Gormus, et. al., Springer Wireless Networks Vol.20, Issue 8, pp.2147-2164, Nov 2014
4. Y. Jin, et. al., IEEE WCNC 2014, pp.3028-3087, 2014
5. https://techradar.cisco.com/trends/Fog-Computing


TOSHIBA
Leading Innovation

Keynote Speech 4:
Dr. Kazuya Kojima
(Kanagawa Institute of Technology)

Development of Real-Time Network Content Creation Technology using Character Animation

Kazuya KOJIMA

Kanagawa Institute of Technology

Our Research

- Human Information Processing
  - Measurement of Human Motion
  - Analysis of Human Motion
- Research and development of image contents
  - Image Processing Software
- Research and development of character contents
  - Character CG Animation

Motion Capture Technology

Motion Capture Technology

- Optical Motion Capture System
  - Analog Camera → Digital Camera
  - 300 thousand pixels → 12 mega pixels
  - Indoors → Outdoors

- Measurement of various Human motion is possible
  - Full body capture
  - Facial capture
  - Finger capture

→ Performance Capture

Current situation and issues of MOCAP

- Motion capture area
  - Big Area, Remote place
- Various devices
  - Input Devices : Analog Signal, Camera, EMR, Sensor…
  - Output Devices : PC, HMD, Tablet…

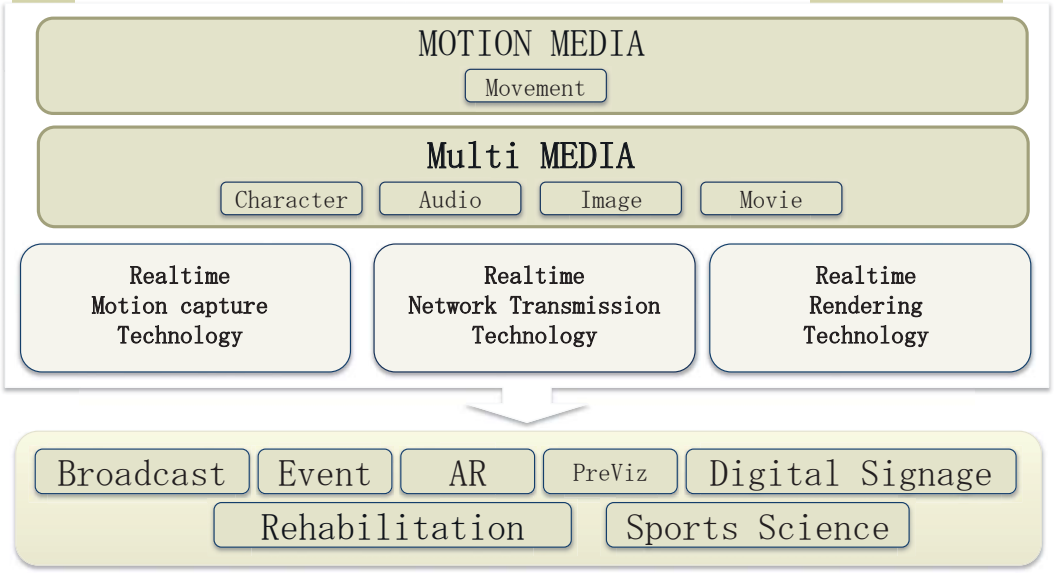
The processing of the computer has been complicated by various usages.

Solution of computer process power → Parallel distributed system

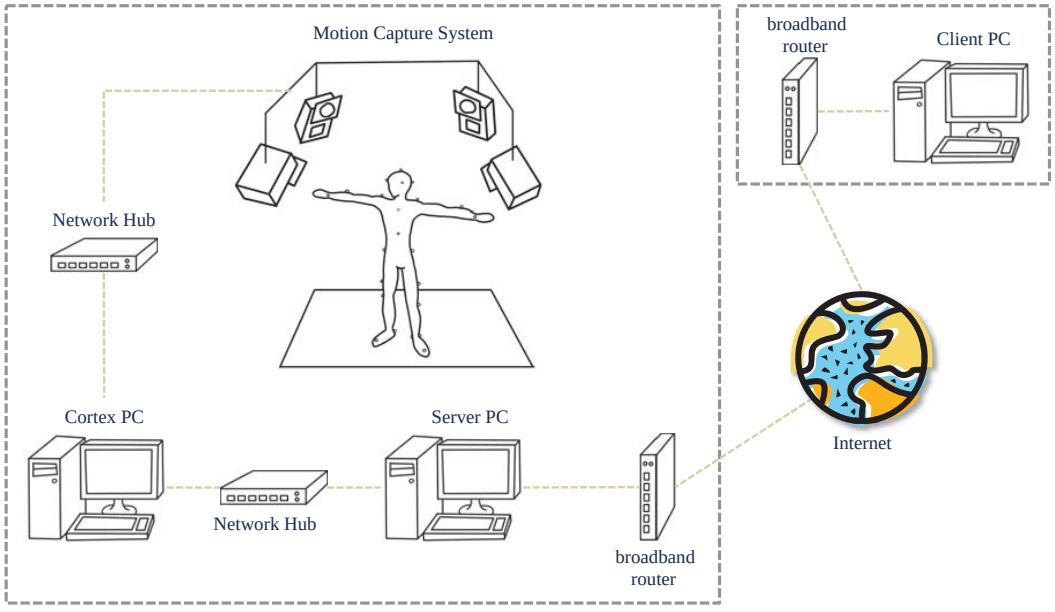
Solution of industrial infrastructure → Network transmission system

Innovative technology

Realtime MOCAP Projects

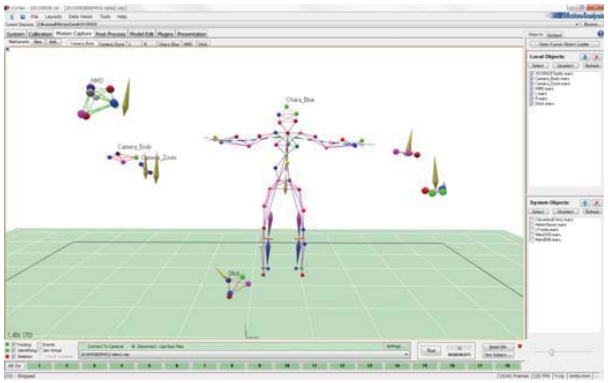


Outline of **RT2MOCAP System**



Realtime Motion Capture Technology

- Realtime processing of measurement data



Data Structure

sFrameOfData	
Variable type	Variable Name
int	iFrame
float	fDelay
int	nBodies
sBodyData	BodyData[MAX_N_BODIES]
int	nUnidentifiedMarkers
tMarkerData*	UnidentifiedMarkers
sAnalogData	AnalogData

Frame Data

Marker Set Information
-Marker Set Name
Segment Information
・Coordinate（X・Y・Z）
・Rotaion（Euler Angles X・Y・Z）
・Segment Length（mm）
Analog Data Information

sBodyData	
Variable type	Variable Name
char	szName[128]
int	nMarkers
tMarkerData*	Markers
float	fAvgMarkerResidual
int	nSegments
tSegmentData*	Segments
int	nDofs
tDofData*	Dofs
float	fAvgDofResidual
int	nIterations

tMarkerData	
Variable type	Variable Name
float	tMarkerData[3]

sAnalogData	
Variable type	Variable Name
int	nAnalogChannels
int	nAnalogSamples
short*	AnalogSamples
int	nForcePlates
int	nForceSamples
tForceData*	Forces

tMarkerData	
Variable type	Variable Name
float	tMarkerData[3]

tSegmentData	
Variable type	Variable Name
double	tSegmentData[7]

tDofData	
Variable type	Variable Name
double	tDofData

tForceData	
Variable type	Variable Name
float	tForceData[7]

Frame Data Structure

Realtime Network Transmission Technology

This system networking in real time transmits the measured motion data to client PC in the remote place. Client PC that receives the motion data displays the CG image by realtime rendering.

We aim to minimize the delay by data processing from the measurement of the motion capture to the display of the screen.

The development of low-latency network transmission protocol is indispensable.

We develop the realtime network transmission technology based on VRPN (Virtual Reality Peripheral Network).

VRPN[The Virtual Reality Peripheral Network]

This projects adopt VRPN by the realtime transmission technology, and are using improved VRPN.

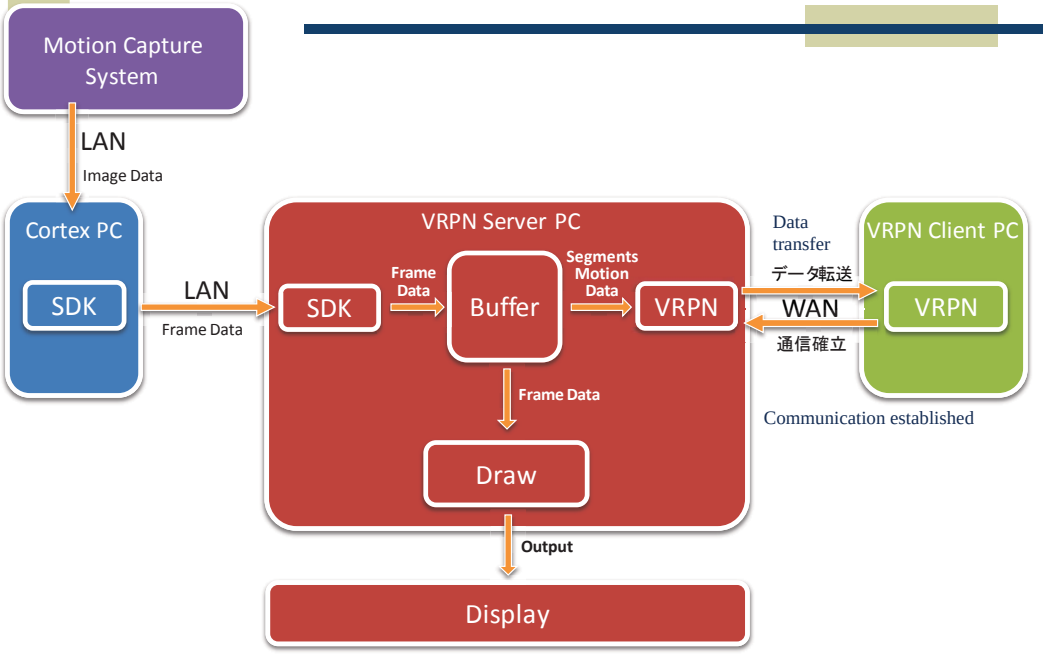
VRPN（Virtual Reality Peripheral Network）

- Network protocol of free license
- Connection by TCP/IP
- VRPN establishes the communication with TCP, and transmits data with UDP. Latency can be suppressed. It is able to effectively deal with resynchronization when a failure occurs in a network.

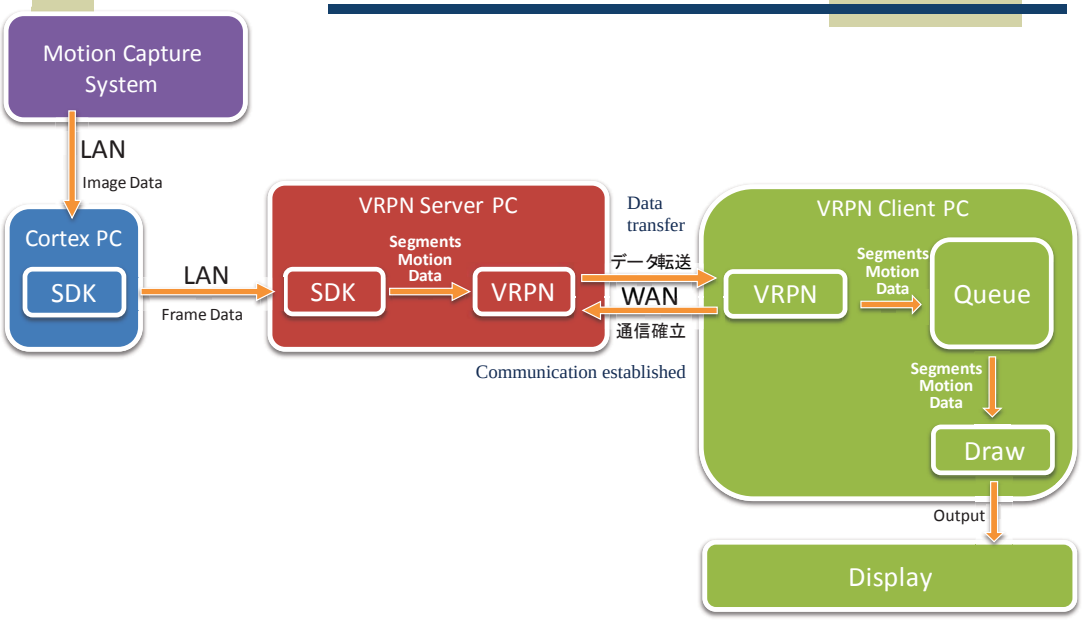
Improvement of VRPN

- NAT Connection
- Proxy to forward packets
- The frame rate in network can be controlled.
- Linear interpolation function of transmission data.

Outline of VRPN Server

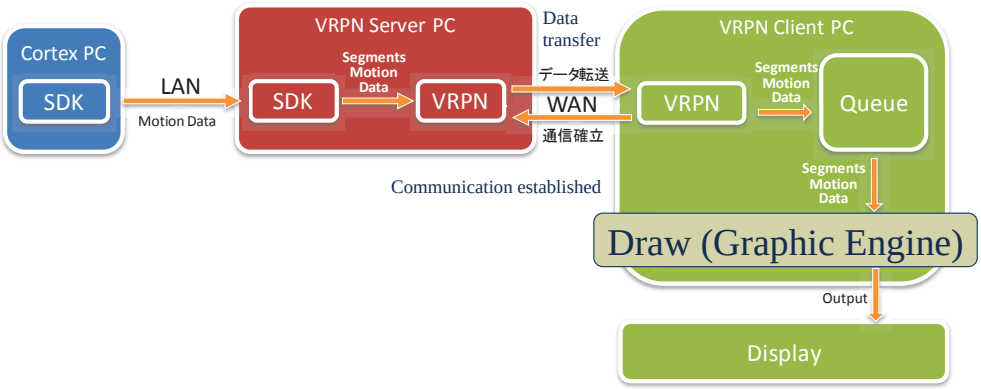


Outline of VRPN Client

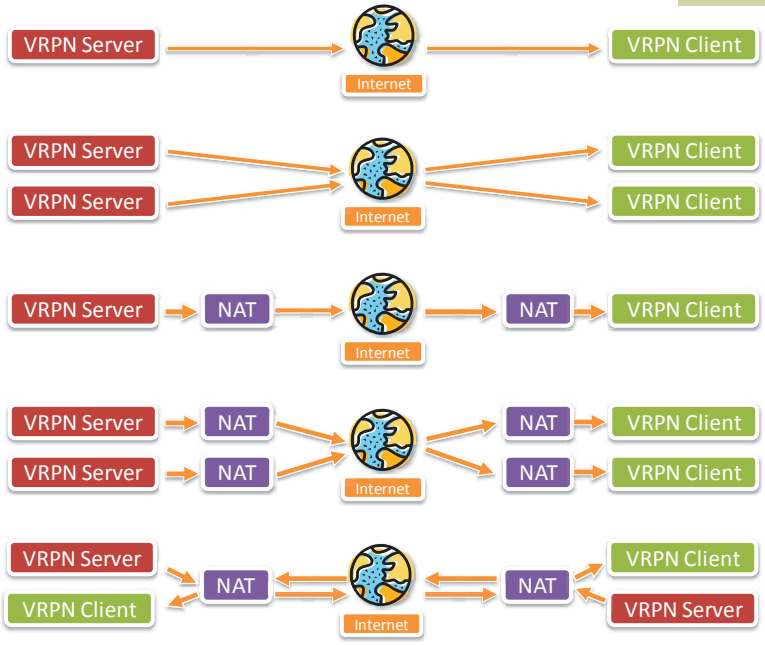


Realtime Rendering Technology

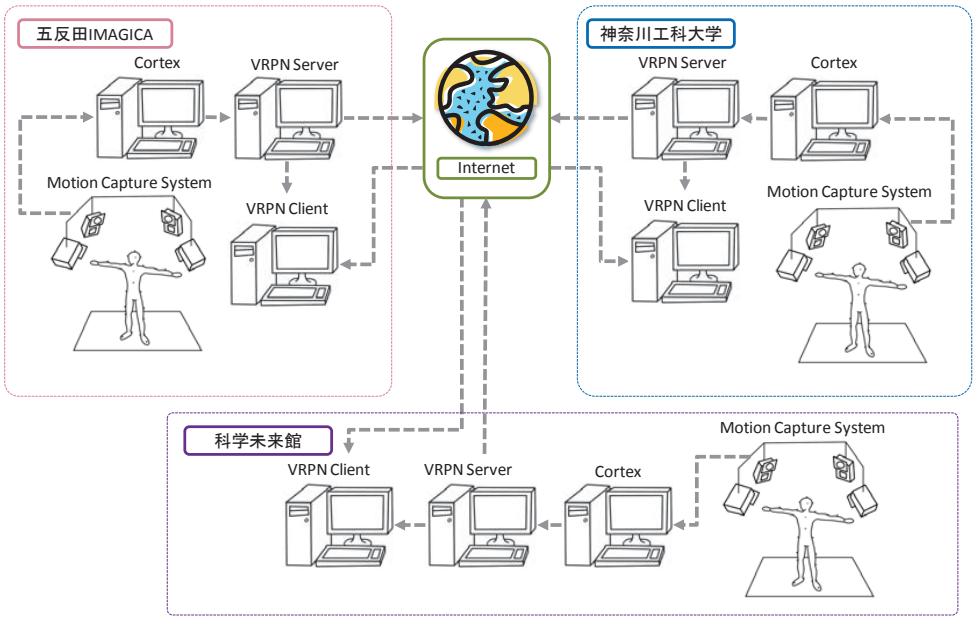
- This project adopts the Game Engine with the realtime rendering engine.



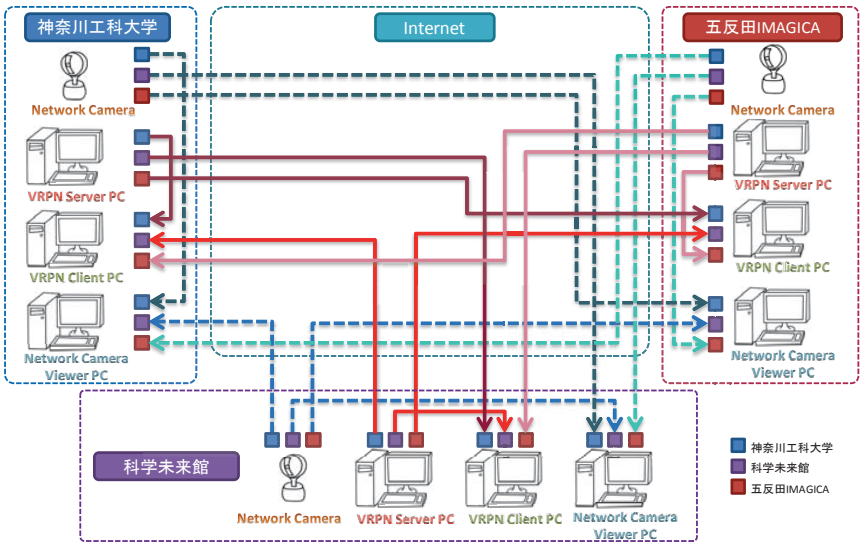
Connection Model of **Realtime Transmission**



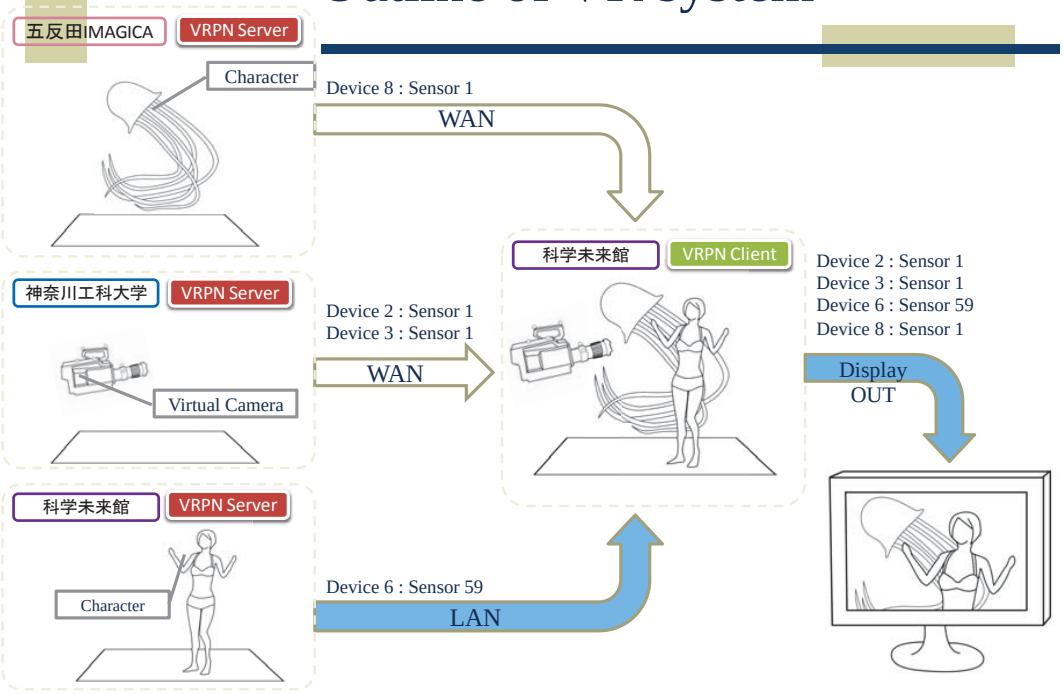
【RT2CharaAR】
Composition of system of two or more points



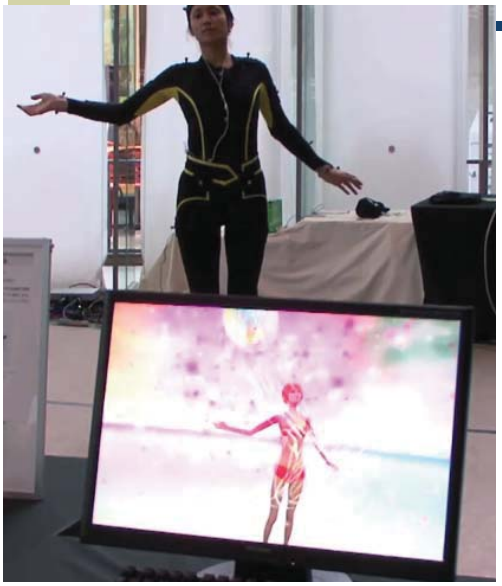
Flow of data between the point



Outline of VR system



VR system Image



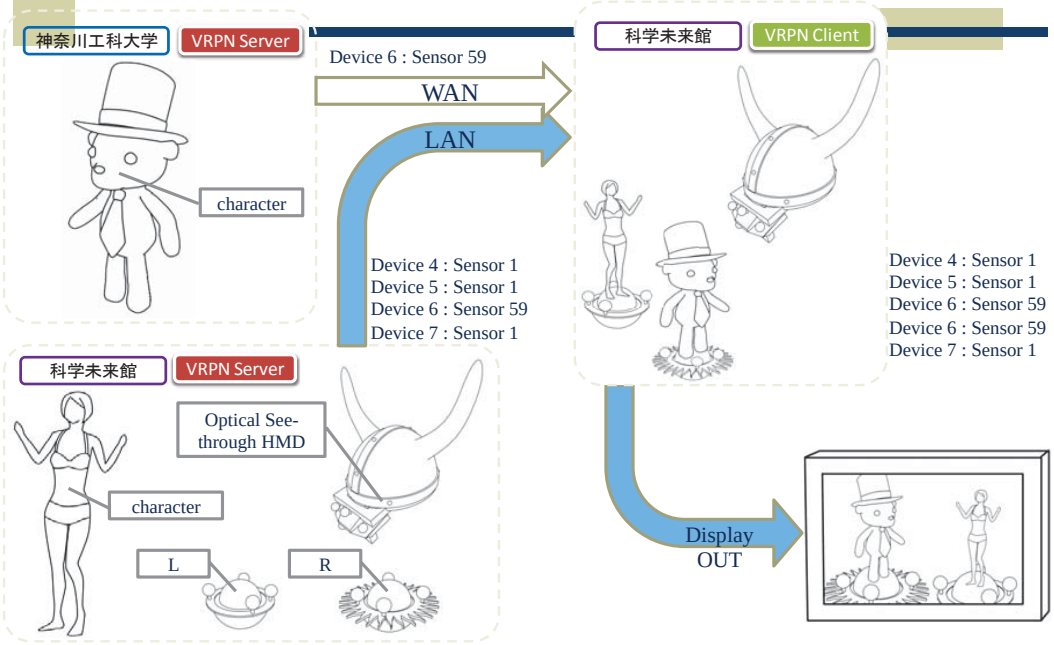
VR System Image

MOCAP(五反田)

MOCAP(神奈川工科大)

MOCAP(科学未来館)

Outline of AR system



AR system Image

Optical see-through HMD

AR System Image

Conclusion

- Construction of Development of Real-Time Network Content Creation Technology using Character Animation
  - This system has in real time achieved the measurement, transmission, and rendering.
  - Character Streaming System
  - Character AR/VR System
  - Possibility of application with broadcasting, event, digital signage, AR, and Pre-viz, Rehabilitation, Sports science.

Future Works of RT2MOCAP

- Time Synchronization in network transmission
  - Generator locking signal
  - Time code generator based on GPS
- Stability of MOCAP system
  - Noise cancellation without increasing delay in the system
- Auto configuration of camera direction and focus
  - Development of remote control system
- Realtime transmission of off-line motion data
  - Construction of asset server of motion data
- Communication channel for operators
  - To communicate among different places

Session7 :
Business Systems and Applications
(Chair : Kozo Okano)

Training Dataset to Induce the Personal Sensibility Model for a Music Composition System

Naoki Tsuchiya†, Takami Koori‡, Masayuki Numao*, and Noriko Otani‡

†Graduate School of Environmental and Information Studies, Tokyo City University, Japan
g1583109@tcu.ac.jp
‡Faculty of Informatics, Tokyo City University, Japan
otani@tcu.ac.jp
* The Institute of Scientific and Industrial Research, Osaka University, Japan
numao@sanken.osaka-u.ac.jp

Abstract - We aim to compose music adapting to personal sensibility. Our composition system learns a personal sensibility model using inductive logic programming and composes music based on the model. Therefore, the quality of the personal sensibility model affects the quality of the composed musical piece. In previous works, training dataset for inducing a personal sensibility model is generated using the individual's ratings for the various popular musical pieces. If he/she knows the musical piece well, the rating is influenced by his/her memories of the piece. Then the personal sensibility model generated based on the rating may include not only the structural features of music but also the memories of the piece. In addition, he/she often hesitates in evaluating. The rating that it took a long time to decide is likely not to represent an intuitive impression. In this paper, we reveal the difference between known music and unknown music as a target music used for rating. Besides, we also reveal the influence of music that was rated with hesitation in rating.

Keywords: Music Composition, Inductive Logic Programming, Personal Sensibility Model

1 INTRODUCTION

A large number of automatic music composition systems have been proposed. Some of those systems apply the interactive evolutionary computation (IEC)[1][2]. In order to reflect the user's preference or request to music, the system based on IEC uses the listener's evaluation as a fitness value. Whenever a musical piece is composed, the user has to evaluate a lot of candidates for the final result repeatedly. This task is excessive burden on the user. On the other hand, some systems target not a person but the people, and generate music using some existing music[3]. In this approach, examples that are pieces of existing music are combined and a mimic music is generated.

We have proposed an automatic composition system that composes music that arouses an individual's particular sensibility. Here, "sensibility" means the ability to cause some emotion by an external stimulus. The kind of sensibility that was aroused by a stimulus differs according to the individual. Our system induces a personal sensibility model, which consists of rules that represent characteristics of the individual's sensibility, using inductive logic programming (ILP). As music pieces are generated based on the model using some optimization algorithm, the quality of the personal sensibility model affects the quality of the composed musical pieces. Previeus works have demonstrated that it is indeed possible to use this technique to compose music that partially adapts to the individual's sensibility [4]–[7].

In previous works, training dataset for inducing a personal sensibility model is generated using the individual's ratings for the various popular musical pieces using a semantic differential method (SDM). If he/she knows the musical piece well and has a special fondness for it, the rating is influenced by his/her memories and emotions. The personal sensibility model generated based on those ratings may contain not only the structural features of music but also the memories of the piece. In addition, we have often observed situations in which listeners hesitated in deciding which ratings best matched to their impressions. If it took a long time to decide, then the ratings is likely not to represent an intuitive impression.

This paper targets two issues in generating training dataset for inducing a personal sensibility model. The first one is the difference between known music and unknown music in rating. The second is the influence of music that was rated with hesitation[1].

2 REPRESENTATION AND COMPOSITION PROCEDURE OF MUSICAL PISECES

In this study, the musical piece is a sequence of quarter notes with a 4/4 time signature, consisting of a frame structure, chord progression consisting of plural motif, melody, and bass part. A motif is the most basic component of a music, and its duration is of two bars. The frame structure has 10 components: *genre*, *key*, *tonic*, *tonality*, *time signature*, *tempo*, *melody instrument*, *melody instrument category*, *chord instrument*, and *chord instrument category*. The chord progression is a sequence of chords. A chord is a set of *root*, *type*, and *tension*. When the previous chord is played in succession, the chord is represented by "-" instead of the set.

The composition flow is illustrated in Figure 1. At first, training datasets are generated and personal sensibility models for the frame structure, motif, and chord progression are

[1]This work was carried out under the Cooperative Research Program (2014328) of "Network Joint Research Center for Materials and Devices" and JSPS KAKENHI Grant Number 26330318.

```
frame(tender,A) :-
   tempo(A,andante),
   chord_category(A,piano).
motif(tender,A) :-
   motif(A,bar(_,_,_,_),
   bar(_,(iv,add9),(i,major),(i,major))).
chords(tender, A) :-
   next_to(A,B,C,_),
   has_chord(A,B,D),type(D,minor),
   has_chord(A,C,E),root(E,v),
   next_to(A,C,F,_),
   has_chord(A,F,G),root(G,i).
```

Figure 2: Examples of personal sensibility model

induced using ILP. In the next step, a frame structure that adapts to the model for the frame structure is generated using an optimization algorithm. In the third step, a chord progression that adapts to the model for the motif and chord progression is generated using an optimization algorithm. After that, a bass part and a melody are generated, and they are combined with the frame structure and chord progression.

3 PERSONAL SENSIBILITY MODELS

A personal sensibility model consists of rules of musical pieces that affects a specific sensibility of a listener. This section explains the method for inducing a personal sensibility model.

3.1 Inductive Logic Programming

Personal sensibility models for the frame structure, motif, and chord progression are induced by ILP. ILP is one of the machine learning techniques that performs inductive reasoning on the first-order predicate logic. It has been used to solve various classification problems. ILP generates some rules based on inductive reasoning using a training dataset that consists of positive examples and negative examples. The generated rule that is called "hypotheses" covers positive examples, and do not cover negative examples. In other words, the rules represent the features of the target concept. The rule is described in the form of a clause of Prolog. The right part of ":-" means the assumption of the rule, the left side of ":-" means the conclusion.

Figure 2 shows the examples of personal sensibility models for the frame structure, motif, and chord progression. These clauses describe musical features that induce a feeling of tenderness in the listener. The first clause indicates that the listener has tender feelings upon listening to music whose tempo is andante and that is played by some kind of piano. The second clause is a feature of a motif wherein the first bar consists of four arbitrary chords and the second bar consists of an arbitrary chord, a IV add9 chord, and two beats of the I major chord. The third clause is a feature of chord progression with three successive chords: minor chord, V chord, and I chord.

One of the most important characteristics of ILP is the use of background knowledge. In this study, the frame structure, motif, and chord progression of each musical piece used for rating are described as background knowledge. An example of background knowledge in this study is shown in Figure 3. Here, `music` is a predicate for describing a frame structure and a chord progression of a musical piece. The first argument

```
music(1,
 song_frame(pops,c,c,major,four_four,
            allegro,piano,piano,
            piano,piano),
 [
  chord(i,major,null),
  chord(v,7,null),
  chord(v,major,null),
  chord(i,major,null)
 ]).
music_structure(1,
 [motif(bar((i,major),-,-,-),
  bar((v,7),-,-,-)),
  motif(bar((v,major),-,-,-),
  bar((i,major),-,-,-))]).
```

Figure 3: Example of background knowledge

of `music` is the serial number of a musical piece. The second argument `song_frame` means a frame structure. The `song_frame` in Figure 3 means that the genre is pops, the key is c, the tonic is c, the tonality is major, the time signature is four/four time, the tempo is allegro, the instrument of melody is piano, the instrument category of melody is piano, the instrument for chord is piano, and the instrument category for chord is piano. The third argument `chord` means a chord progression. The first `chord` in Figure 3 is a major 1st that has no tension note. `music_structure` is a predicate for describing motifs in a musical piece. The first argument of `music_structure` is the serial number of a musical piece. The second argument `motif` consists of two bars. The first `motif` in Figure 3 represents that the first bar contains a whole note of major 1st chord, and the second bar contains a whole note of 5th seventh chord.

3.2 Training Dataset

We target an individual's specific sensibility for inducing a personal sensibility model and composing a musical piece. Target sensibilities can be located in the direction "positive" or "negative" on the axis of sensibility evaluation. A target sensibility and its opposite sensibility are used together to generate training datasets.

The individual listens to various musical pieces using SDM, and his/her affective perceptions are noted. The individual rates a musical piece on a scale of 1-5 for bipolar affective adjective pairs, namely, favorable-unfavorable, bright-dark, happy-sad, tender-severe, and tranquil-noisy. The former adjective in each pair expresses a positive affective impression, while the latter expresses a negative one. The individual should provide a high rating for any piece that he/she finds positive. On the other hand, he/she may provide a low rating for any piece that he/she believes to be negative.

Two training datasets, t_1 and t_2, are generated using the individual's ratings. The individual's ratings for positive and negative examples in each training dataset are shown in Table 1. When learning a model for a positive impression, pieces with higher ratings are used as positive examples and the remaining pieces are used as negative examples. When learning a model for a negative impression, pieces with lower ratings are used as positive examples and the remaining pieces are used as negative examples.

Two personal sensibility models with different levels are

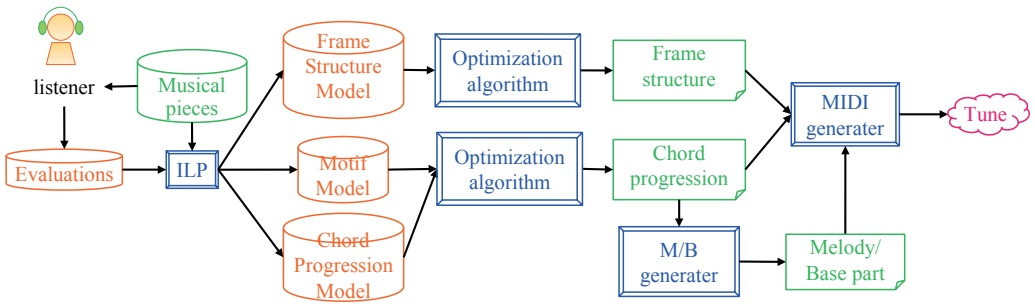


Figure 1: Composition flow

Table 1: Ratings for musical pieces in examples

Direction of target impression	Training dataset	Ratings	
		Positive examples	Negative examples
Positive	t_2	5	1-4
	t_1	4, 5	1-3
Negative	t_2	1	2-5
	t_1	1, 2	3-5

constructed using two kinds of training sets. They are reflected to a musical piece in different levels. In the process of composing a musical piece based on an optimization algorithm, a solution candidate Y is evaluated using the degree of adaptability to the personal sensibility model $m(Y)$ that is calculated in the equation (1).

$$m(Y) = \sum_{i=1}^{2}(2i-1)\left\{\sum_{j=1}^{r(X_i)} c(X_i,j)n(Y,X_i,j) - \sum_{j=1}^{r(X_i')} c(X_i',j)n(Y,X_i',j)\right\} . \quad (1)$$

Here, X_i is the personal sensibility model that was generated using training data t_i for the target adjective. X_i' is the personal sensibility model that was generated using training data t_i for the antonymous adjective. $c(X_i,j)$ is the number of positive examples covered by the j-th clause in the personal sensibility model X_i when the model was learned by ILP. $n(Y,X_i,j)$ is the number of parts in Y that satisfy the j-th clause in the personal sensibility model X_i. $r(X_i)$ is the number of rules contained in the personal sensibility model X_i. The solution candidate that contains the rules for the target adjective that cover a lot of positive examples has a higher value of $m(Y)$. On the other hand, the solution candidate that contains the rules for the antonymous adjective that cover a lot of positive examples has a lower value.

3.3 Ratings of Known Music and Unknown Music

In previous works, the training dataset for inducing a personal sensibility model is generated using the individual's ratings for the various popular musical pieces. If he/she knows the musical piece well and has a special fondness for it, the rating is influenced by his/her memories and emotions. It may be impossible to avoid the mere exposure effect: a psychological phenomenon by which people tend to prefer things merely because they are familiar with them. Therefore training dataset may contain musical pieces whose ratings were decided by factors other than the musical structure.

The features described in the background knowledge indicate only the musical structures. The features about the memories and emotions are not contained. It is difficult to generate rules using the background knowledge about musical structures and training dataset based on the musical structures and memories, because the difference between positive and negative examples is not clear in the space defined by the background knowledge. If we were to try to generate rules in such a situation, then less rules will be generated. Although they may cover more positive examples, they indicate less features of musical structures. It is expected that more rules that indicate features of musical structures will be generated by using the unknown music.

3.4 Evaluating with Hesitation

We have often observed the situations in which listeners hesitated in deciding which ratings best matched to their impressions. We analyzed the time taken for rating by 14 Japanese university students. They rated 50 musical pieces in terms of five impression adjective pairs using the evaluation system. The system records the times when the musical piece was started and when the rating was input. Students were instructed that they could rate in any order and change ratings as often as the liked. When a rating was changed, the system recorded the time again.

The histograms of the time taken for rating are similar in all students. Figure 4 shows the histogram of time taken for rating by a student. Ratings in the right side of the histogram might have been decided with hesitation.

Ratings that took a long time to decide are likely not to represent an intuitive impression. Therefore, training dataset that use such ratings may contain musical pieces whose ratings do not reflect the individual's impression correctly. Musical pieces rated with hesitation may be located on the boundary

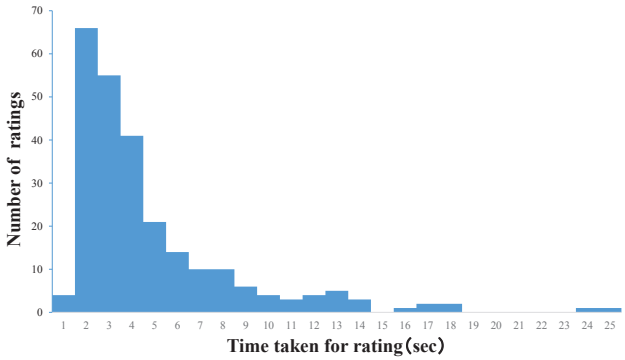


Figure 4: Example of the histogram of the time taken for rating

of the positive and negative examples in the hypothesis space, and it is difficult to generate rules that cover positive examples near the boundary. It is expected that rules that represent intuitive impressions will be generated by removing musical pieces rated with hesitation from the training dataset.

4 EXPERIMENTS

Experiments were conducted in which 14 Japanese university students participated. The participants were asked to listen to 50 known musical pieces and 48 unknown musical pieces. The known musical pieces are from popular movies, TV dramas, etc. The unknown musical pieces were newly composed for this study. They rated each piece in terms of the five impression adjective pairs on a scale of 1-5.

As mentioned in Section 3.4, the participants' ratings might have been decided with hesitation. Figure 4 indicates that the frequency of the ratings taking longer time tends to be lower. As the distribution is not normally distributed, we converted the histogram to normal distribution using Box-Cox transformation. We classify the musical pieces with ratings in the right 5% area of the converted distribution as the musical pieces rated with hesitation.

The training datasets t_1 and t_2 are generated as in the previous works. In addition, the training datasets t'_1 and t'_2 are also generated by removing musical pieces rated with hesitation from t_1 and t_2 respectively. In the following, the pair of t'_1 and t'_2 is called $Dataset1$, and the pair of t_1 and t_2 is called $Dataset2$.

Four cases were prepared to investigate the difference between $Dataset1$ and $Dataset2$ and the difference between known and unknown musical pieces. $Dataset1$ is used in the first and second cases, $Dataset2$ is used in the third and fourth cases. The unknown musical pieces are used in the first and third cases, the known musical pieces is used in the second and fourth cases.

The personal sensibility models for each participant and each case are learned by employing FOIL [8], a top-down ILP system that learns the function-free Horn clause definitions of a target predicate using background predicates.

Average number of the rules contained in each personal sensibility model for all participants is shown in Table 2. As mentioned in Section 3.3, more rules were generated for the motif and chord progression by using the unknown musical pieces. It makes possible to compose musical pieces that contain various motifs. However, there are a few differences between known and unknow musical pieces in the number of rules in the frame structure model. In addition, there are a few differences between $Dataset1$ and $Dataset2$ in the number of rules in all kinds of models. A paired t-test at a significance level $\alpha = 0.01$ was conducted to examine the difference not only between the unknown musical pieces and the known musical pieces but also between $Dataset1$ and $Dataset2$ in the number of the rules. In the results, the number of the rules for the motif and chord progression by using the unknown musical pieces were different.

We counted the number of fixed elements in each rule. The fixed element for the frame structure means the literal in the body of the rule. The fixed element for the motif means the argument except "_" in the predicate "`bar`." The fixed element for the chord progression means the literal whose predicate is "`has_chord`." Average number of the fixed elements in each rule contained in each personal sensibility model for all participants is shown in Table 3. More fixed elements in each rule were generated for the motif by using the known musical pieces. However, there are a few differences between known and unknow musical pieces in the number of rules in the frame structure model and chord progression model. In addition, there are a few differences between $Dataset1$ and $Dataset2$. A paired t-test at a significance level $\alpha = 0.01$ was conducted to examine the difference not only between the unknown musical pieces and the known musical pieces but also between $Dataset1$ and $Dataset2$ in the number of fixed elements in each rule. In the results, the number of the fixed elements in each rule for the motif by using the known musical pieces were different.

The Average gain value of each rule is shown in Table 4. Gain is the value used for evaluating hypotheses in FOIL. A hypothesis with larger gain value covers more positive examples and less negative examples. As mentioned in Section 3.3 and Section 3.4, the number of positive examples covered by a rule was decreased by using the unknown musical pieces. In addition, $Dataset1$ using the unknown musical pieces was decreased as compared to the $Dataset2$ using the unknown musical pieces in the all kinds of models. However, $Dataset1$ using the known musical pieces was increased as compared to $Dataset2$ using the known musical pieces in the motif model. In addition, the difference in the gain values between $Dataset1$ and $Dataset2$ are very small. A paired t-test at a significance level $\alpha = 0.01$ was conducted to examine the difference not only between the unknown musical pieces and the known musical pieces but also between $Dataset1$ and $Dataset2$ in the gain values. In the results, the gain values were different for only the motif by using the unknown musical pieces. Therefore, it is necessary to re-examine how to decide the musical pieces rated with hesitation.

5 CONCLUSION

In this study, we revealed the difference between known music and unknown music in rating and the influence of mu-

Table 2: Average number of rules in a personal sensibility model

Training dataset	Personal sensibility model	Known	Unknown
$Dataset1$	Frame structure	10.13	9.93
	Motif	2.22	6.31
	Chord progression	4.65	5.75
$Dataset2$	Frame structure	10.20	9.38
	Motif	1.80	6.67
	Chord progression	3.98	5.49

Table 3: Average number of fixed elements in a rule

Training dataset	Personal sensibility model	Known	Unknown
$Dataset1$	Frame structure	1.74	1.93
	Motif	1.70	1.18
	Chord progression	1.77	1.61
$Dataset2$	Frame structure	1.99	2.11
	Motif	1.72	1.27
	Chord progression	1.78	1.76

sic that was rated with hesitation in generating training dataset for inducing the personal sensibility model. The experimental results show that more rules for the motif and chord progression are generated by using unknown music. In addition, the number of positive examples covered by a rule is decreased by using the unknown musical pieces and removing the musical pieces rated with hesitation from the training dataset. In the future, we will consider another approach to decide the musical pieces rated with hesitation.

REFERENCES

[1] J.A. Biles, “GenJam: A Genetic Algorithm for Generating Jazz Solos”, Proc. of the International Computer Music Conference, pp.131-131 (1994).

[2] D.Ando, P.Dahlstedt, M.G.Nordaxhl, and H.Iba, “Computer Aided Composition by Means of Interactive Gp”, Proc. of the International Computer Music Conference, pp. 254-257 (2006).

[3] D.Cope, “Machine Models of Music”, MIT Press, (1992).

Table 4: Average gain value

Training dataset	Personal sensibility model	Known	Unknown
$Dataset1$	Frame structure	1.77	1.73
	Motif	3.56	2.07
	Chord progression	2.17	1.88
$Dataset2$	Frame structure	1.82	1.76
	Motif	3.54	2.09
	Chord progression	2.18	1.91

[4] R.Legaspi, Y.Hashimoto, K.Moriyama, S.Kurihara, and M.Numao, “Music Compositional Intelligence with an Affective Flavor”, Proc. of ACM International Conference on Intelligent User Interfaces, pp. 216-224 (2007).

[5] N.Otani, K.Tadokoro, S.Kurihara, and M.Numao, “Generation of Chord Progression Using Harmony Search Algorithm for a Constructive Adaptive User Interface”, Proc. of 12th Pacific Rim International Conference on Artificial Intelligence, LNAI 7458, pp.400-410 (2012).

[6] N.Otani, R.Kamimura, Y.Yamano, and M.Numao, “Generation of Rhythm for Melody in a Constructive Adaptive User Interface”, 4th International Workshop on Empathic Computing, (2013).

[7] N.Otani, S.Shirakawa, and M.Numao, “Symbiotic Evolution to Generate Chord Progression Consisting of Four Parts for a Music Composition System”, Proc. of 13th Pacific Rim International Conference on Artificial Intelligence, LNAI 8862, pp. 849-855 (2014).

[8] J.Quinlan, “Learning Logical Definitions from Relations”, Machine Learning 5, pp. 239-266 (1990).

Dissolve in Scents Using Pulse Ejection

Sayaka Matsumoto*, Shutaro Homma*, Eri Matsuura*, Shohei Horiguchi*, and Ken-ichi Okada**

* Graduate School of Science of Technology, Keio University, Japan
** Faculty of Science and Technology, Keio University, Japan
{matsumoto, honma, matsuura, shohei, okada}@mos.ics.keio.ac.jp

Abstract - A trial to raise a sense of reality by using scents with various types of media has lately attracted much attention. In addition, it is thought that we can raise a sense of reality more by not only adding scents, but also expressing the movement of scents with that of the picture. We aimed at the development of the presentation technique to express dissolve in scents with paying attention to changing the intensity of two types of scents. The results of experiments revealed that receivers may feel dissolve by presenting fade-in and fade-out in scents which are overlapped in three breathes. It is expected that the technique can raise realistic sensations when scents are presented in accordance with pictures by establishing the technique of dissolve in scents.

Keywords: Olfactory display, Pulse ejection, Dissolve in scents

1 INTRODUCTION

Information transmission and communication tends to be limited to visual information and audio information. However, the transmission of information via all five senses (sight, hearing, touch, smell and taste) has lately attracted much attention [1]. Olfactory information recognized by the olfactory organs differs from the information recognized via the other four senses [2]. The sense of smell powerfully affects humans since olfactory information is directly transmitted to the cerebral limbic system that governs emotions. In addition, olfactory information has high importance since it is thought that the presentation of olfactory information is effective as a means to enhance the sense of reality like three-dimensional vision and sound [3].

For transmitting scents together with other media, it is necessary to control the presentation of scent in accordance with the changes in images/sounds over time. In doing so, it is more effective to enhance the sense of reality. Therefore, we paid attention to both the change of types and the intensity of scents and developed presentation techniques in scents changing the intensity of two types of scents. Among these, this study presents a technique in scents that enables the receivers to feel "dissolve", which we defined as "the second scent becomes gradually strong at the same time as the first scent becomes gradually weak".

First, we constructed presentation techniques in scents that the receivers to feel "fade-out in scents" (the scent becomes gradually weak) and "fade-in in scents" (the scent becomes gradually strong) using pulse ejection. Then, we expressed dissolve in scents with combining fade-out in scents and fade-in in scents, and examined whether the receivers can feel dissolve in scents.

2 RELATED WORK

2.1 The Study of adding Scents

Trials on the transmission of olfactory information together with other media are currently being conducted. "Scents of Space" which Haque et al. developed is the art work of lights and scents, which can carry a scent with the wind from one wall of the room and send it to receivers [4]. A trial to present scents in accordance with movie at a movie theater carried out using a device called Aromageur [5] which is the scent generator can save the recipe of the scents [6]. These trials are aimed for the enhancing the sense of reality by adding scents in the room or with videos.

There are many scenes in videos and TV programs, such as the scene which many smelling objects appear at the same time, suddenly appear and gradually disappear. Therefore, it is necessary not only to present scents but also to control the presentation of scent in accordance with the changes in images/sounds over time. In doing so, it is more effective to enhance the sense of reality.

However, they paid attention only to adding scents with other media in related works and the study that are paid attention to changes of types and the intensity of scents are not performed so many. The conventional presentation method of scents which are used in related works continues emitting scent at high density for a long time that the receivers can feel enough scents. Too much scent emitted over a continuous period leaves in the air. If the scent presented before mixes with scent presented later, there is a possibility that the receivers cannot feel the change of the types of scents.

Moreover, too much scent causes human adaptation to the scent, and thus, the receivers may not feel the intensity of scents properly. From such problems, it was possible to add scents but it was difficult to change types and the intensity of scents.

2.2 Presentation Technique that Emit Scent for Very Short Periods of Time

In our previous research, we minimized the influence of the scents to spread in the air by emitting scents for just very short periods of time and reduced the fragrance which remained in the air. In this way, we can reduce human adaptation and change scents without scents being mixed.

We defined this presentation technique that emits scent for very short periods of time as "pulse ejection" [7], and we studied about the change of kind of scents and the intensity of the scents by using pulse ejection. About the change of the kind of scents, for example, we measured the interval of the ejection time that human can recognize two types of scents clearly without being mixed, and developed a presentation technique that the receivers can feel two types of scents in one breath [8]. Furthermore, applying this presentation technique, we developed presentation techniques that the receivers can feel the strength of the relation between two types of scents by presenting a weak scent earlier and a strong scent later [9]. Besides, when we paid attention to the change of kind of scents every fixed time, it was revealed that the receivers can feel the change of kind of scents every two breathes [10]. About the change of the intensity of scents, for example, we developed presentation techniques that the receivers can feel the scent is coming near or going away by changing the intensity of a scent every two breathes [11]. Like these, we enabled to present the change of scents by using pulse ejection.

3 DISSOLVE IN SCENTS

3.1 Presentation Technique of Dissolve in Scents

As it was previously mentioned at Section 2.2, we studied about the change of types and the intensity of scents independently to express the presentation of scents by using pulse ejection. In this study, we paid attention to both the change of types and the intensity of scents and devised presentation techniques in scents changing the intensity of two types of scents.

For transmitting scents together with video, it is thought that it is effective to use the presentation technique of scents that is suitable for scene conversions of the videos. Therefore, we focus on dissolve in scene conversions of the videos. In videos, dissolve is a scene conversion that "the next scene is gradually superimposed as the former scene fades out" [12]. We propose to express dissolve in scents based on dissolve in videos. This study presents a technique in scents that enables the receivers to feel "dissolve", which we defined as "the second scent becomes gradually strong at the same time as the first scent becomes gradually weak". To develop the presentation technique to express dissolve in scents, it is necessary to develop the presentation technique that the receivers can feel the scent becomes gradually strong and gradually weak. Therefore, we defined "fade-in in scents" as the presentation technique that the scent becomes gradually strong, and "fade-out in scents" as the presentation technique that the scent becomes gradually weak. Then, we present dissolve in scents with combining fade-out in scents and fade-in in scents.

3.2 Pulse Ejection

When we present dissolve in scents, if we use the conventional presentation method of scents that emit scent at high density for a long time that the receivers can feel enough scents, the scent presented before mixes with scent presented later and it is difficult to feel scents as we expected. Besides, the scent which was presented for a long time causes human adaptation to the scent. The receivers may not feel the intensity of scents properly. In this study, we propose the scent presentation technique to create dissolve by using pulse ejection. An olfactory display we developed uses the technique used in ink-jet printer, and can use pulse ejection.

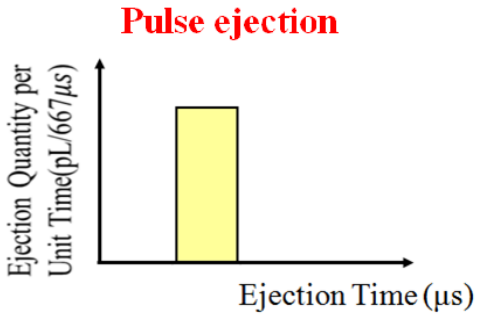


Figure 1: Image of pulse ejection

Pulse ejection is controlled scents as quantity by two parameters, ejection quantity of scent per unit time and ejection time (Figure 1). Besides, this device can change the ejection time at 667 μs intervals, so that it can present scent during only one breath. In presenting infinitesimal quantity of scent like this, pulse ejection can minimize the lingering of scents in the air.

4 EXPRESSION OF FADE-IN AND FADE-OUT IN SCENT

In this chapter, to determine how many times we should change the intensity of scents to express fade-in and fade-out at first. We examine how many times human can feel the change of the intensity of scents(Preliminary Experiment). After that, we developed the presentation techniques to express fade-in in scents and fade-out in scents.

4.1 Olfactory Display

We developed an olfactory display called "Fragrance Jet 2(FJ2)" as shown in Figure 2. This display uses the technique used in ink-jet printer in order to produce a jet which is broken into droplets from the small hole in the ink tank. This device can use pulse ejection for scent presentation so that the issue such as scent lingering and care to eject scent can be minimized. The display can set up one scent ejection head. This head can store three small tanks and one large tank, thus this display can contain 4 types of scents. There are 127 minute holes in the head connected to the small tank and 256 minute holes in the head connected to the large tank. Moreover, the display can emit scent from multiple holes at the same time. We denote the number of minute holes emitting at one time as "the number of simultaneous ejections". So, the ejection quantity is adaptable to 0-127 (small tank), 0-256 (large tank) if the ejection time is set. In this study, the pulse ejection time is

Figure 2: Fragrance Jet 2 (FJ2)

set to 100msec, so the ejection quantity is controlled by the change of the number of simultaneous ejections. We define the number of simultaneous ejections as the "intensity" of scents at this display.

4.2 Preliminary Experiment : the Number of Times of the Change

4.2.1. Experimental Method

At first, we measured the detection threshold of each participant. The detection threshold is the smallest density at which scent can be detected and where the user does not need to recognize the kind of a smell. The experiment to determine the detection threshold was conducted using the scent of banana stored in a large tank. The intensity was changed by 10 between 10 and 250. We use the triangle test [13] to judge the detection threshold in the measurement. In the triangle test, three stimuli are presented at random, where one of them is scented and the other two are odorless. The participant then answers when the scented odor was presented. Furthermore, we used the raising method (the first intensity was 10) to measured the detection threshold. The detection threshold was determined by the intensity which the participant answered correctly twice in a row. If the participant selected the wrong answer, the intensity was raised by 10.

After measuring the detection threshold, we examined how many times the participant could feel a change of the intensity of scent when the intensity was changed by 10 between the detection threshold of each participant and 250. We presented two scented ejections of different intensity to each participant. The interval between two scented ejections was 4 seconds. We presented first scented ejection in first breath and presented second scented ejection in second breath. We signaled the timing of breathing by sounds. When the scent is ejected, the countdown starts with the auditory cue. Scent emission then commences 0.5 sec after giving the cue "Go" according to previous study[14].

When we examine the number of times that the participant can feel the change of intensity of scent, we used two methods, rising method and dropping method. In rising method, we started the experiment from the intensity of detection threshold to 250. We prepared a reference value and a comparison value. The first reference value is the intensity of detection threshold and we started the experiment from the reference value. We presented two scented ejections to each participant in random order. One scented ejection was presented in the intensity of reference value and the other one was presented in the intensity of the comparison value which is larger than the reference value by 10. Then, we instructed the participant to answer which of the two was strong. If the participant answered correctly twice in a row, we judged that the participant can distinguish the intensity of the two, and recorded the comparison value. After that, we substituted the intensity of the comparison value for the next reference value and resumed the experiment. If the participant selected the wrong answer, we changed the comparison value to the value that is larger than the last comparison value by 10 and resumed the experiment. When the comparison value was reached 250(the maximum value), we finished the experiment. In dropping method, we started the experiment from 250 to the intensity of detection threshold. The same as the rising method, we prepared a reference value and a comparison value. The first reference value is 250 and we started the experiment from the reference value. We presented two scented ejections to each participant in random order. One scented ejection was presented in the intensity of reference value and the other one was presented in the intensity of the comparison value which is smaller than the reference value by 10. Then, we instructed the participant to answer which of the two was weak. If the participant answered correctly twice in a row, we judged that the participant can distinguish the intensity of the two and recorded the comparison value. After that, we substituted the intensity of the comparison value for the next reference value and resumed the experiment. If the participant selected the wrong answer, we changed the comparison value to the value that is smaller than the last comparison value by 10 and resumed the experiment. When the comparison value was reached the detection threshold of each participant (the minimum value), we finished the experiment.

In rising method, participants were 5 men and 3 women in their 20s. In descending method, participants were 4 men and 3 women in their 20s. Participants of the descending method were same as those of the rising method.

4.2.2. Results

Figures 3 and 4 show the results of the number of times that participants felt the changes of the intensity in case of the rising method and the dropping method. The horizontal axis shows the number of times that participants could feel the changes of the intensity. The vertical axis shows the intensity of the scent that participants could notice. We plotted the intensity for each participant, such as the first change of the intensity that he or she was able to feel, and the second change of the intensity that he or she was able to feel. It is revealed that all participants could notice the changes of the intensity more than 7 times both in rising method and in dropping method. Furthermore, it is thought that the results of Figures 3 and 4 can be approximated by a straight line. We calculate the average value of the intensity for each number of times and showed the approximation

straight line to these graphs. Tables 1 and 2 show the values of the intensity of scent that are calculated by substituting the number of times for the approximation and rounded off to the first decimal place. Using these values, we developed the presentation techniques to express fade-in in scents and fade-out in scents.

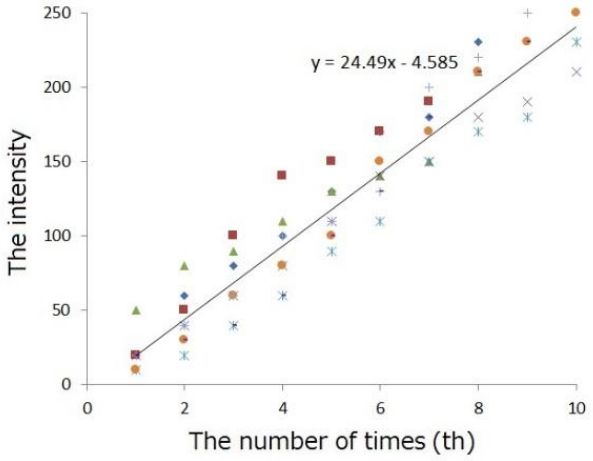


Figure 3: The intensity of scents that participants could notice (rising method)

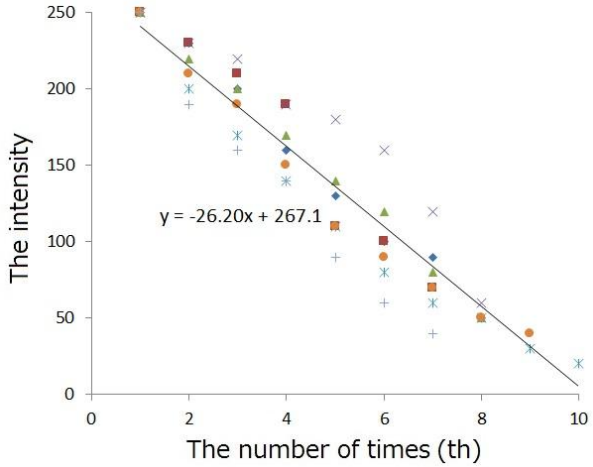


Figure 4: The intensity of scents that participants could notice (dropping method)

Table 1: calculated values (rising method)

The number of times	1	2	3	4	5	6	7	8	9	10
Intensity	20	45	70	95	120	145	170	195	220	245

Table 2: calculated values (dropping method)

The number of times	1	2	3	4	5	6	7	8	9	10
Intensity	247	220	193	166	139	112	85	58	31	4

4.3 Fade-in and Fade-out in Scents

Using the values of Tables 1 and 2, we examined whether the participant can feel fade-in in scents and fade-out in scents when we changed the intensity every 1 breath. When we presented one of the values of Table per breath to each participant and tested whether he/she can feel fade-in and fade-out. Participants were 5 people in their 20s. As a result, in fade-in, all participants answered that they felt "the scent becomes gradually strong". However, most of them were not able to feel scent when the intensity of the scent was less than 45, and were not able to feel the change of the intensity when the intensity of the scent was greater than 170. From this, participants could feel six changes of the intensity. In fade-out, all participants answered that they felt "the scent becomes gradually weak". However, most of them were not able to feel scent when the intensity of the scent was less than or equal to 31, and were not able to feel the change of the intensity when the intensity of the scent was greater than or equal to 193. From this, participants could feel six changes of the intensity.

Therefore, we decided the number of the changes of the intensity up to six times. We decided to present the intensity of scents in the range of 45 – 170 in fade-in. We decided to present the intensity of scents in the range of 35 – 170 in fade-out to make the maximum value of fade-out same to that of fade-in.

4.3.1. Experimental Method

We examined how short fade-in and fade-out in scents can be felt. As described above, we defined fade-in in scents as the presentation technique that the scent becomes gradually strong, and fade-out in scents as the presentation technique that the scent becomes gradually weak. In addition, we judged that participant can feel the change of the intensity smoothly when they can feel a lot of changes of the intensity. The experiment was conducted using the scent of banana stored in a large tank. In fade-in, we divided the intensity between minimum 45 and maximum 170 linearly into (a) 6 phases, (b) 5 phases, (c) 4 phases and (d) 3 phases. In fade-out, we divided the intensity between minimum 35 and maximum 170 linearly into (e) 6 phases, (f) 5 phases, (g) 4 phases, (h) 3 phases. As examples, Figure 5 and 6 show the images of presentation of scent about (a) 6 phases and (e) 6 phases. The horizontal axis shows the number of times and the vertical axis shows the intensity of the scent that we presented. We presented each scented ejection once per 1 breath, and numbers in figure 5 and 6 show the intensity of scents that we presented for each breath. The interval between each breath was 4 seconds. The same as preliminary experiment, we signaled the timing of breathing by sounds. After we presented 1 phase, we asked the participant how they felt about the changes of the intensity of scent. After taking a short break, we presented another 1 phase to the participant. We instructed them to answer by 2 ways, multiple choices and graphs, when they answered. In a way of answering by multiple choices, the participant was instructed to choose the most suitable item from seven evaluation items as follows; "The scent becomes strong

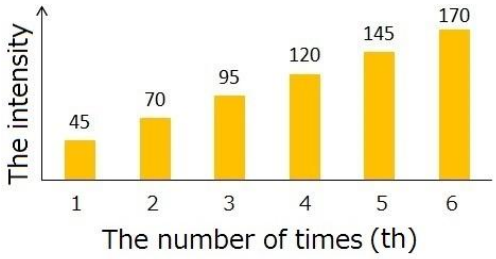


Figure 5: Image of (a) 6 phases (fade-in)

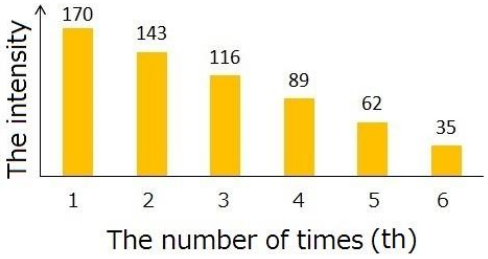


Figure 6: Image of (e) 6 phases (fade-out)

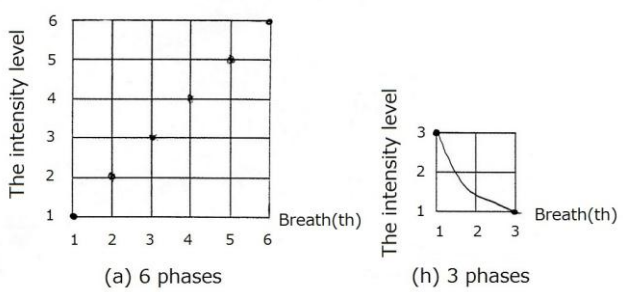


Figure 7: Example in a way of answering graphs

suddenly", "The scent becomes gradually strong", "The scent becomes weak suddenly", "The scent becomes gradually weak", "There was no change in feeling the scent", "There was no scents as a whole", "The scent irregularly changed from feeling strong or weak". In a way of answering by graphs, we used plotting paper like Figure 7. Figure 7 also shows the example of drawing and plotting. The horizontal axis shows the number of times and the vertical axis shows the intensity level of scent. Level 1 is the weakest, and Level maximum ((e) 6 phases: level 6, (h) 3 phases: level 3) is the strongest. In fade-in ((a) 6 phases - (d) 3 phases), we set the intensity in first breath to level 1. In fade-out ((e) 6 phases - (h) 3 phases), we set the intensity in first breath to level maximum. We instructed each participant to draw or plot the changes of the intensity of scents that he or she felt after second breath based on these levels. Participants were 10 men and 6 women in their 20s.

4.3.2. Results and Considerations

At first, we describe the result and the consideration of fade-in in scents. Next, we describe the result and the consideration of fade-out in scents.

- **Fade-in in Scents**

First, we describe the results about a way of answering by multiple choices. We judged the most suitable phases to create fade-in in scents with the number of participants that answered "The scent becomes gradually strong". Table 3 shows the number of the answers for four types of phases to create the impression of "gradually strong". It is revealed that the largest number of participants answered "gradually strong" in (a) 6 phases. The result of fade-in in multiple choices indicates that the participants could feel fade-in the best in (a) 6 phases.

Next, we describe the results about a way of answering by graphs. We judged that participants could feel the changes of the intensity smoothly when the number of the changes of the intensity participants could feel was the largest. We plotted the average values that are calculated from the values that each participant drew in each breath in each phase. For example, Figure 8 shows the plotting points of (a) 6 phases. The horizontal axis shows the number of times and the vertical axis shows the intensity level of scent. To examine whether the participants could feel the change of the intensity of scent between contiguous two breaths, such as the first breath and the second breath, the average values between contiguous two breaths were analyzed using t-test. If there was no significant difference between contiguous two breaths, the average values between n-th breath and (n+2)-th breath were analyzed. Significant differences were found in all contiguous two breaths at (a) 6 phases, (c) 4 phases and (d) 3 phases ($p < 0.05$). The result of these indicates that the participants could feel all changes of intensity of scent that we presented in those three phases. That is, the participants could feel six changes of the intensity in (a) 6 phases, four changes of the intensity in (c) 4 phases, three changes of the intensity in (d) 3 phases. On the other hand, in (b) 5 phases, it follows that participants could feel four changes of the intensity. The result of fade-in in graphs indicates that the participants could feel fade-in the most smoothly in (a) 6 phases. In light of results of two ways above, the participants were likely to feel fade-in in scents the best in (a) 6 phases.

Table 3: The number of participants that answered "The scent becomes gradually strong"

Fade-in	(a) 6 phases	(b) 5 phases	(c) 4 phases	(d) 3 phases
The number of participants	12	7	8	8

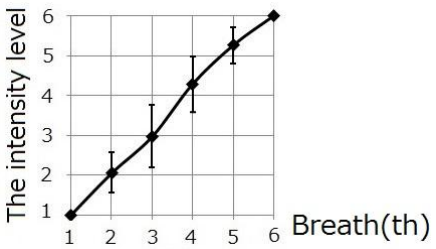


Figure 8: The result of the graphs in (a) 6 phases

- **Fade-out in Scents**

First, we describe the results about a way of answering by multiple choices. We judged the most suitable phases to create fade-out in scents with the number of participants that answered "The scent becomes gradually weak". Table 4 shows the number of the answers for 4 types of phases to create the impression of "gradually weak". It is revealed that the largest number of participants answered "gradually weak" in (e) 6 phases. The result of fade-out in multiple choices indicates that the participants could feel fade-out the best in (e) 6 phases.

Next, we describe the results about a way of answering by graphs. We judged that participants could feel the changes of the intensity smoothly when the number of the changes of the intensity participants could feel was the largest. We plotted the average values that are calculated from the values that each participant drew in each breath in each phase. For example, Figure 9 shows the plotting points of (e) 6 phases. The horizontal axis shows the number of breaths and the vertical axis shows the intensity level of scent. To examine whether the participants could feel the change of the intensity of scent between contiguous two breaths, such as the first breath and the second breath, the average values between contiguous two breaths were analyzed using t-test. If there was no significant difference between contiguous two breaths, the average values between n-th breath and (n+2)-th breath were analyzed. As a result of comparison at (e) 6 phases, significant differences were found in between the first breath and the second breath, the second breath and the fourth breath, the fourth breath and the fifth breath and the fifth breath and the sixth breath ($p < 0.05$). The result of this indicates that the participants could feel changes of intensity of scent at the first, second, fourth, fifth and sixth breath in (e) 6 phases. That is, the participants could feel five changes of the intensity in (e) 6 phases. On the other hand, in (f) 5 phases, it follows that participants could feel four changes of the intensity. In (g) 4 phases, it follows that participants could feel three changes of the intensity. In (h) 3 phases, it follows that participants could feel two changes of the intensity. The result of fade-out in graphs indicates that the participants could feel fade-out the most smoothly in (e) 6 phases. In light of results of two ways above, the participants were likely to feel fade-out in scents the best in (e) 6 phases.

From these results, we combined (e) 6 phases in fade-out and (a) 6 phases in fade-in, and examined whether the participants can feel dissolve in scents.

Table 4: The number of participants that answered "The scent becomes gradually weak"

Fade-out	(e) 6 phases	(f) 5 phases	(g) 4 phases	(h) 3 phases
The number of participants	12	6	6	5

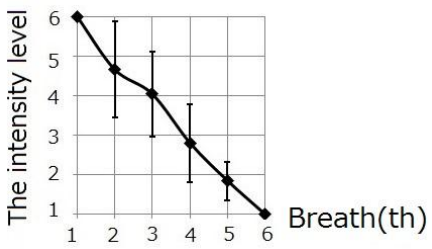


Figure 9: The result of the graphs in (e) 6 phases

5 EXPRESSION OF DISSOLVE IN SCENTS

We defined dissolve in scents as the presentation technique that "the second scent becomes gradually strong as the first scent becomes gradually weak" and "two types of scents were felt in a single breath". In addition, we judged that participant can feel the change of the intensity smoothly when they can feel a lot of changes of the intensity. That is, there are the parts that two types of scents are presented in one breath to express dissolve in scents. In this experiment, we examined how many overlapping parts the presentation technique have that participants can feel dissolve the best when we presented the overlapping part that two types of scents are presented once per one breath.

5.1 Experimental Method

The experiment was conducted using the scent of banana stored in a large tank to express fade-out in scents and the scent of mint stored in a small tank to express fade-in in scents. In fade-out, we changed the intensity of scents six times using (e) 6 phases in the experiment 4.3. In fade-in, as the size of the tank was different from the experiment 4.3, we converted the intensity of scent (stored in a large tank) of (a) 6 phases in the Experiment 4.3 into those of scent stored in a small tank.

In this experiment, we prepared five patterns, from "overlap 1" to "overlap 6". For example, "overlap 1" is a presentation that there is one overlapping part at which two types of scents are presented in one breath, and "overlap 6" is a presentation that there are six overlapping parts. Figure 10 shows the image of a presentation of "overlap 3". The "overlap 3" has three overlapping parts, so there is nine breaths overall in "overlap 3". In Figure 10, yellow shows the scent of banana, and green shows the scent of mint. Numbers in figure 10 shows the intensity of scents that we presented for each breath. When combining fade-out and fade-in, we used a presentation technique that the receivers can feel two types of scents in one breath [8]. Figure 11 shows a presentation of the overlapping part. The horizontal axis shows the breaths and the vertical axis shows the intensity of the scent. The interval between two types of scents was set to 0.7sec [8]. We presented the scent of banana before and the scent of mint after, in spite of the intensity of the scents. The interval between each breath was 4 seconds in the Experiment 4, for smelling two types of scents, the interval between each breath was set to 5 seconds in this experiment. The same as preliminary experiment, we signaled the timing of breathing by sounds. After we presented 1 overlap to the participant, we asked the participant how they felt. After taking a short break, we presented another one overlap to the participant. We

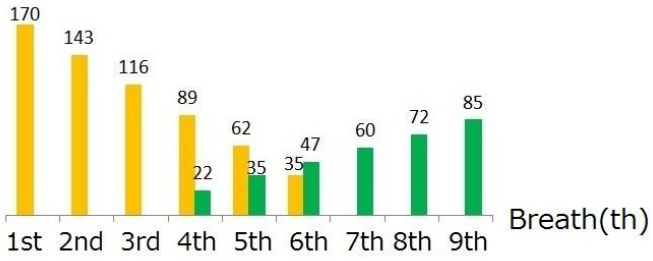


Figure10: Image of overlapping (“overlap 3”)

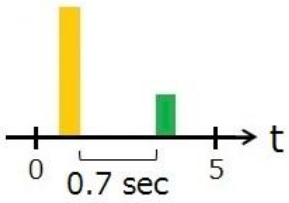


Figure 11: A presentation of the overlapping part

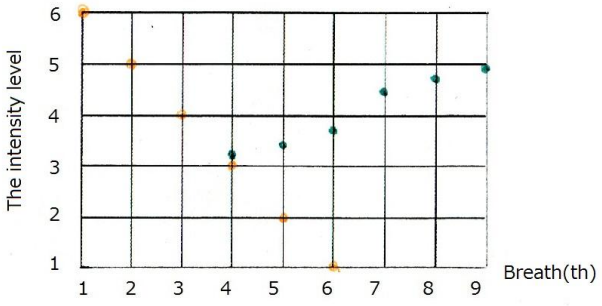


Figure 12: Example in a way of answering graphs (“overlap 3”)

instructed them to answer by two ways, sentences and graphs, when they answered. In a way of answering by sentences, the participant was instructed to answer about the “types” and the “intensity” of scents. In a way of answering by graphs, we used plotting paper like Figure 12. The horizontal axis shows the number of times and the vertical axis shows the intensity level of scent. Level 1 is the weakest, and Level 6 is the strongest. The numbers of breath were different for each “overlap”, so we changed the size of plotting paper for each “overlap”. Figure 12 shows the example of “overlap 3”. We set the intensity in first breath to level 1. We instructed each participant to draw or plot the changes of the intensity of scents that he or she felt after second breath. At the time, the participant was instructed to change the color of the pen when he/she felt some types of scents. Figure 12 also shows the example of plotting. Participants were 10 men and 6 women in their 20s.

5.2 Results and Consideration

First, we describe the results about a way of answering by sentences. In questions about the types of the scents, we asked participants “How many scents do you feel in one overlap?” Table 5 shows the number of the participants that answered “I could feel two types of scents in the overlap.” for six types of overlaps. Table 5 indicates that about 80% of participants could feel two types of scents. There were few participants that felt only one kind of the scent because they could not notice the changes of scents, and felt three types of scents because the scents were mixed. Also, in questions about the types of the scents, we asked participants “Can you feel two types of scents in one breath?” Table 6 shows the number of the participants that answered “I could feel two types of scents in one breath.” for six types of overlaps. Table 6 indicates that more than half of participants could feel two types of scents without mixed in one breath from “overlap 1” to “overlap 6”.

We judged that the participants who answered “I felt two types of scents in one breath” (the question about the types of scents) and “I felt the first scent becomes gradually weak and the second scent becomes gradually strong” (the question about the intensity of scents) could felt dissolve. Table 7 shows the number of the answers for six types of overlaps to create the impression of dissolve. It is revealed that the half number of participants could feel dissolve in scents in all overlaps. Even out of those, the number of participants that could feel dissolve in scents at “overlap 1” and “overlap 3” was the largest.

Next, we describe the results about a way of answering by graphs. We judged that participants could feel the changes of the intensity smoothly when the number of the changes of the intensity participants could feel was the largest. We plotted the average values that are calculated from the values that each participant drew in each breath in each overlap. For example, Figure 13 shows the plotting points of “overlap 3”. The horizontal axis shows the number of breaths and the vertical axis shows the intensity level of scent. The level which participants could not feel scents was set to the level 0. In “overlap 3”, the scent of banana was presented in between the first breath and the sixth breath and the scent of mint was presented in between the fourth breath and the ninth breath. That is, we presented two types of scents in between the fourth breath and the sixth breath. To examine whether the participants could feel the change of the intensity of scent between contiguous two breaths, such as the first breath and the second breath, the average values between contiguous two breaths for each scent were

Table 5: The number of participants that answered “I felt two types of scents in one overlap”

overlap	1	2	3	4	5	6
The number of participants	15	15	16	15	15	14

Table 6: The number of participants that answered “I felt two types of scents in one breath”

overlap	1	2	3	4	5	6
The number of participants	8	10	9	10	9	13

Table 7: The number of participants that could feel dissolve

overlap	1	2	3	4	5	6
The number of participants	8	7	8	7	7	6

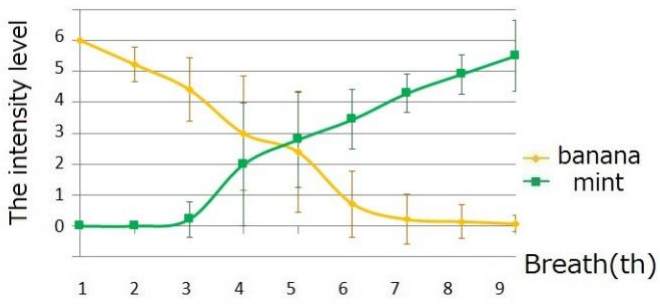


Figure 13: The result of the graphs in "overlap 3"

using t-test. If there was no significant difference between contiguous two breaths, the average values between n-th breath and (n+2)-th breath for each scent were analyzed.

In this paper, we describe the result about "overlap 3". As a result of comparison in the scent of banana at "overlap 3", significant differences were found in between the first breath and the second breath, the second breath and the third breath, the third breath and the fourth breath and the fourth breath and the sixth breath ($p < 0.05$). The result of this indicates that the participants could feel changes of intensity of scent at the first, second, third, fourth and sixth breath. That is, the participants could feel five changes of the intensity in the scent of banana at "overlap 3". As a result of comparison in the scent of mint at "overlap 3", significant differences were found in between the fourth breath and the sixth breath, the sixth breath and the seventh breath and the seventh breath and the eighth breath ($p < 0.05$). The result of this indicates that the participants could feel changes of intensity of scent at the fourth, sixth, seventh and eighth breath. That is, the participants could feel four changes of the intensity in the scent of mint at "overlap 3". The similar comparison was also conducted for other overlaps, and it was found that the number of changes of the intensity that participants could feel was the largest in "overlap 3". Therefore, the result of dissolve in graphs indicates that the participants could feel dissolve the most smoothly in "overlap 3".

In light of results of two ways above, the participants likely to feel dissolve in scents the best at "overlap 3" in those six overlaps. In this study, we used the scent of banana and the scent of mint. Based on this study, we will examined whether the participant can feel dissolve in scents in different types of scents.

6 CONCLUSION

Studies on transmitting scents together with various media to enhance the sense of reality are currently conducted. There are many scenes in videos and TV programs, such as the scene which many smelling objects appear at the same time, suddenly appear and gradually disappear. Therefore, it is necessary to control the presentation of scent in accordance with the changes in images/sounds over time. In doing so, it is more effective to enhance the sense of reality. However, the conventional presentation method of scents continues emitting scent at high density for a long time and gives problems of human adaptation to the scent and scents lingering. The receivers may not feel the intensity of scents properly. Therefore, it was difficult to present scents in accordance with videos.

To solve such problems, we studied about the change of kind of scents and the intensity of the scents by using pulse ejection which emits scent for very short periods of time. In this study, we paid attention to both the change of types and the intensity of scents and devised presentation techniques in scents changing the intensity of two types of scents. Among these, we especially studied a presentation technique in scents that enables the receivers to feel "dissolve" which we defined as "the second scent becomes gradually strong at the same time as the first scent becomes gradually weak". First, we developed presentation techniques in scents that the receivers to feel "fade-out" and "fade-in" using pulse ejection. As a result, the participants were likely to feel fade-out in scents and fade-in in scents the best in 6 phases. From these results, we combined 6 phases in fade-out and 6 phases in fade-in, and examined whether the participants can feel dissolve in scents. We aimed at the development of the presentation technique to express dissolve in scents with paying attention to the change of two scents and the intensity of the scents. The results of experiments revealed that participants could feel dissolve in scents the best by presenting fade-in and fade-out in scents which are overlapped in three breath. Based on this study, we will examined whether the participant can feel dissolve in scents in different types of scents. It is expected that the technique can raise realistic sensations more when scents are presented in accordance with pictures by establishing the technique of dissolve in scents.

REFERENCES

[1] Ministry of Posts and Telecommunications in Japan, Reports of the Association for Information and Communications Technology Using Five Senses, (2007) (in Japanese).

[2] Goodrich Hunsaker, N.J., Gilbert P.E. and Hopkins, R.O., The Role of Human Hippocampus in Odor-Place Associative Memory, Chemical Sences, vol.34, No.6, pp.513-521, (2009).

[3] Chika Oshima, Koichi Nakayama and Hiroshi Ando, Aroma That Enhances the Reality of Visual Images, Review of the National Institute of Information and Communications Technology, Vol.56, Nos.1/2, (2010) (in Japanese).

[4] Usman Haque, Scents of space: an interactive smell system, SIGGRAPH'04 ACM SIGGRAPH 2004 Sketches, 35, (2004).

[5] ITmedia LifeStyle, the Lessons of Aromatherapy with Internet by OCN, http://www.itmedia.co.jp/lifestyle/articles/0505/23/news096.html (in Japanese)

[6] INTERNET Watch, NTT Com creates the movie called "New World" with a fragrance, http://internet.watch.impress.co.jp/cda/news/2006/04/11/11594.html (in Japanese)

[7] Ami Kadowaki, Junta Sato, Yuichi Bannai and Ken-ichi Okada, Presentation Technique of Scent to

Avoid Olfactory Adaptation, Proc.of ICAT 2007, pp.97-104, (2007).
[8] Kaori Ohtsu, Junta Sato, Yuichi Bannai and Ken-ichi Okada, Measurement of Olfactory Characteristics for Two Types of Scent in a Single Breath, IFIP INTERACT 2009, pp.306-318, (2009).
[9] Daisuke Noguchi, Kaori Ohtsu, Yuichi Bannai and Ken-ichi Okada, Scent Presentation Expressing Two Smells of Different Intensity Simultaneously, 2009 Joint Virtual Reality Conference of EGVE-ICAT-EuroVR, pp.53-60, (2009).
[10] Sayumi Sugimoto, Kaori Ohtsu, Yuichi Bannai and Ken-ichi Okada, Scent Presentation Technique to Enable High Speed Switch of Scents, The Journal of the Virtual Reality Society of Japan, Vol.15, No.SBR-1, pp.17-22, (2010)(in Japanese).
[11] Junta Sato, Kaori Ohtsu, Yuichi Bannai and Ken-ichi Okada, Pulse Ejection Technique of Scent to Create Dynamic Perspective, In proc. Of ICAT 2008, pp.167-174, (2008).
[12] CUERBO English Dictionary, http://english.cheerup.jp/eedict/search?name=dis-solve
[13] Japan Association on Odor Environment, The guidebook of the Simplified measurement of Smell, (2005).
[14] Daisuke Noguchi, Sayumi Sugimoto, Yuichi Bannai and Ken-ichi Okada, Time Characteristics of Olfaction in a Single Breath, CHI'11 Proceedings of the 2011 annual conference on Human factors in computing systems, pp.83–92 (2011).

A Speculation on a Framework that Provides Highly Organized Services for Manufacturing

Takashi Sakakura*, Mitsuteru Shiba*, Tatsuji Munaka**

*Information Technology R&D Center, Mitsubishi Electric Corporation, Japan
Sakakura.Takashi@bx.MitsubishiElectric.co.jp
Shiba.Mitsuteru@ap.MitsubishiElectric.co.jp
** School of Information and Telecommunication, Tokai University, Japan
Munaka@tsc.u-tokai.ac.jp

Abstract - recently, a movement that seeks a novel level of manufacturing occurs in many countries such as Industrie 4.0 in Germany. The goal of such activities is to improve efficiency in every field of manufacturing by applying the ICT. One of the issues in this movement is vertical integration of services from plant floor to decision making via some BI tools for visualization. We propose a framework that can accommodate any kinds of services and provide a session control that enables a session with any combination of services on the framework, since the existing SOA software is not good enough in flexibility of adopting services and processing data with low latency. We also assume that the framework runs on data centers said "cloud", and issues regarding security are considered as well.

Keywords: manufacturing, services, efficiency, PLM, SCM

1 INTRODUCTION

All the people in the world are experiencing the unknown in every field of interest, everyday, everywhere. The total population has grown to over 7 billion from 1 billion at the beginning of nineteenth century [1]. And activities of 7 billion people may affect the climate, which has been stable for 10,000 years, seems changing. It's nothing in geological point of view, but does matter much for us such as that super storm appears often. The technologies shall realize the sustainability, especially foods, while keeping the quality of life in economically advanced countries and improving the quality of life for 1.8 billion of the juvenile on this planet. To achieve this, the nearest way is to eliminate waste and improve efficiency in every productive activity by utilizing the ICT.

On the other hand, innovations in the science and the engineering accelerate more and more while some of them are "destructive" as letting the existing technologies being obsolete. As for the ICT, the performance of supercomputers of NUMA architecture, the latency and the bandwidth of the communication systems including both wired and wireless, PC and handheld devices (e.g. smart phone or tablet) are drastically improved. As these as a background, said "cloud computing" was emerged. The cloud computing consists of data centers, which is basing on grid computing technology that is developed mainly for calculation or signal processing programs which require the CPU resource intensively. In service point of view, the user of the cloud computing buy the computer resources from the operator via the network. This business model brought conveniences to user such as that very small initial costs to begin providing services, flexible allocation of computer resources by on-demand basis, even the TCO might be cheaper than on-premise systems considering the costs of human resource who maintains the on-premise systems.

It seems as if back to a centric model of computer usage, the age of the mainframes, but brought back with huge computational power. Many organizations already provide services on the cloud. There are many success stories in cloud computing [2], and the novel level of services are available in combination of high performance mobile devices [3].

We can say that the cloud computing increases the benefit by eliminating the waste from every aspects of the business. If the existing cloud services augment to talk another service, may reduce waste, and improve efficiency by coordinating each other. The coordination makes more benefit for respective business as a result, and may contribute the sustainability. For example, in the food business, there are many business entities such as farmers, fishers, markets, food manufacturing, logistics, retailers, restaurants, and the consumers. If the all of these and more entities such as meteorological agency collaborated, eliminating huge amount of waste of the foods is achieved.

The services for manufacturing are not exception. Technical environments are ready to provide highly organized services for manufacturing, and that will enable to collaborate another service of business organizations. We believe the fusion of respective business services is a key to realize the sustainable society.

With the wish as described, we designed and implemented a framework that is applicable to control the plant floor , especially the discrete systems which require real time processing , and accommodate many services like the PLM, SCM, MES, ERP, BI for data visualization, the analytics for every level of manufacturing data, and the communication capability to talk the other service entity. This paper consists as follows. In chapter 2 we explain how the framework works and describe services on the framework in chapter 3. In chapter 4 discuss the interface that enable to collaborate another service entities. In chapter 5, describe evaluation of the framework including the collaboration, and conclude in chapter 6.

2 FRAMEWORK OF SERVICE

2.1 Overview

Figure 1 depicts an example of proposing systems. Around 30 of data centers are deployed in the world, and all the data centers synchronize the data in the storage via the dedicated communication lines among them (in case of the Google, the dedicated line is called “B4” [4]). Every services

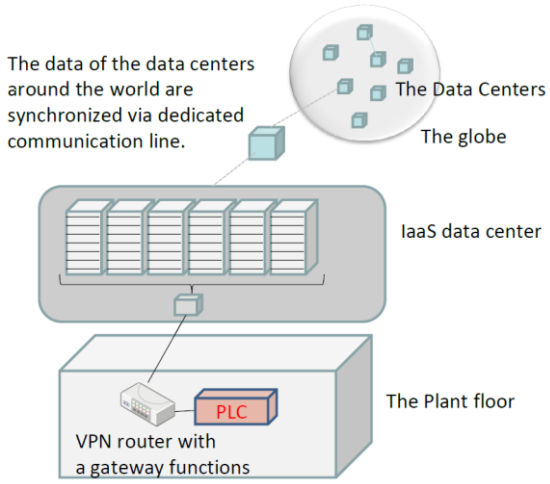


Figure 1: **Overview of the System**

on the framework share the specific data world-wide, so users can get good responses from the services by accessing the closest data center. The synchronization is not only used for storage data but also data in the volatile memories with a function we implemented.

Another characteristics should be observed in Figure 1 are that we apply the cloud service in form of the IaaS and connections between the plant floor and the data center are via the VPN and HTTPs while the interface for users is mainly HTTPs.

2.2 Framework

Basic functions of the framework are follows

- Storage data encrypting by functional encryption [5]
- Authentication and access control with the original public key of the functional description
- Session control that invokes a session which consists of plural of services
- Reflective distributed memory that reflects all of data in factories/plants

1) Data encryption

We adopted the functional encryption for encrypting storage since the functional encryption enables users to decode cipher data with one’s private key which is not a set of a public key, which is used originally for encrypting plain data.

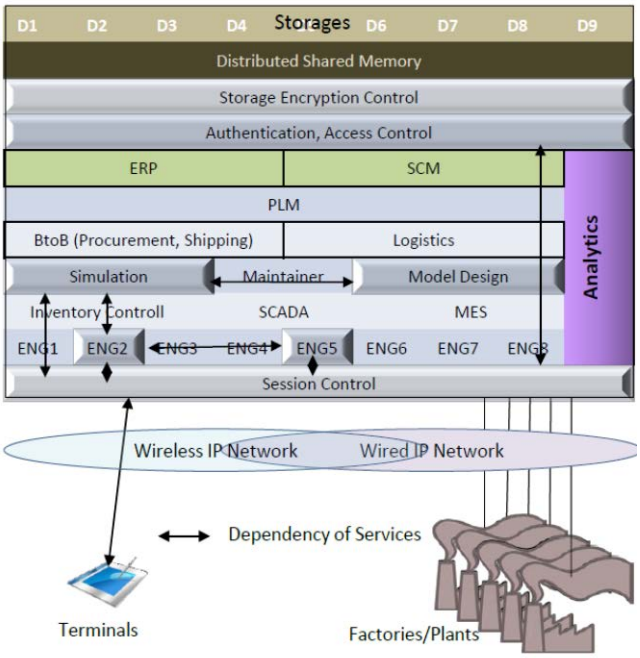


Figure 2: Framework and services

In another words, users can access the cipher data with a given private key if the access control admitted. This encrypting/decrypting scheme is necessary to protect the data even from administrators of the data center.

2) Authentication and access control

User authentication is processed via the TLS [6] based on a database which accommodates information of user password and privileges. By a demand of the session control, the access control generates a re-encryption key for cipher data that are required by a session invoked by the session control.

3) Session control

After the successful authentication, the session control invokes a set of services according to a dependency of services described in a form of XML. Figure 2 shows an example of session. Firstly, the session control gets the one re-encryption key for all the cipher data that some services in the session have to access from the access control. And setups inter process communications as described in the XML message.

ENG2 denotes an engineering tool for programming PLC, while ENG5 denotes an engineering tool for programming graphic operation terminals. Usage scenario in this example of session is below. The dependencies among the services are designated by arrows.

- There was a demand of minor specification change of a product
- An operator logs in to the system and selects the product.
- Opens model design editor of the product and makes a modification
- Checks if the modification is correct by using the simulator

- ENG2 modifies PLC program by information provided by the simulator
- ENG5 reflects the modification of PLC program to the graphic operation terminal

Other than the example above, sessions with various combination of services can be established such as alarming by disconnected form of session, decision making for business based on tons of data that are organized via the analytics, ERP, and BI tools.

4) Reflective distributed memory

This is a fundament of the framework. All the data memory of all the controllers in a plant is mapped to the memory of the framework. Updates both from framework and controller are reflected to the other in very low latency [6].

The memory mapped by the controllers is also mapped by the storage. Therefore the memory data are synchronized via the storage data synchronization mechanism provided by the data center operator results in synchronizing the memory data as well.

However, a portion of the latency is coming from the distance from the plant floor to the data center, and is not evitable. So the control operations that can accept the latency are able to execute on the data center.

3 SERVICES ON THE FRAMEWORK

As figure 2 illustrates, any kind of service, which satisfy some requirements can be deployed on the framework. The requirements are listed below.

- Should be a server program and communication interface are configurable
- Preferably with its source code to fit on the frame work more tightly
- No kluge code that is hard to detect influences to the framework
- Does not require particular GUI

Some services are under construction so far, however we think services in figure 2 are essential.

ERP is an essential service for decision making today with use of BI tools while SCM is to operate the business. The functional border of service is getting unclear because most of these kind software is augmenting their functions.

PLM is also an essential if a product needs many parts. Siemens's Teamcenter is becoming the de-facto standard especially for manufactures produce products that consist of many parts such as aircrafts or cars .

As a trend today, model based design is getting its place. The design method is usually coupled with simulation. The other services in figure 2 are useful for manufacturing and somehow deployed on the framework, especially the service of the engineering tools that gives field engineers freedom, anywhere, anytime.

Lastly, the analytics that provides data analysis in every layer of services. But no magic is in data analysis. What should be analyzed, what the data analyzed, how the data retrieved must be considered by experts. The analytics provides an engine for analysis

References should be collected at the end of your paper. They should be prepared according to a recognized style, e.g. the Harvard or sequential numeric system making sure that your accumulated list corresponds to the citations made in the main text and that all material mentioned is generally available to the readers. When referring to them in the text, type the corresponding reference number in square brackets as in this example [1].

4 COLLABORATION WITH ANOTHER SERVICE ENTITY

In Japan, a law that obligates all the medical care party to report their expense to a community on a cloud was established in 2011. The cloud community is functioning very well to reduce the cost for over killed medicate.

Imagine if the community (a service) collaborates with the framework tailored for a pharmacy company's automation system. The medicines are able to be tailored for patients, The mass of anti-virus medicine production can be determined depending on forecast of infection disease considering latent period and so on.

Collaborations that we wish to endorse as soon as possible are any parties in food business. A half of whole foods is abandoned in Japan despite the fact that Japan imports 60% of foods approximately [7]. Big portions of this waste is committed by retailers and restaurants. Almost of them have own system including POS. If those systems interface the framework of manufactures regarding foods, we can make a great contribution to starving people.

Almost of the companies provide some WEB service as well. That implies they have some staff or consultants to maintain the WEB service. So the WEB, more precisely an application protocol on the HTTP is the easiest and most acceptable candidate as interface with the framework.

We will investigate what they have in their mind as requirements of the interface and propose an application protocol to IETF. The protocol would be not peer to peer (i.e. multicast) like the Websocket [8], and contents of messages would be "what do you want to buy?", "what do you want to sell?", "what amount of a product manufacturer should make?", "what logistics do you want to use?".

We expect that let the framework achieve to collaborate to another services is not very hard in technical point of view, however easily anticipate that coordinating the low should be costly effort.

5 EVALUATION OF THE FRAMEWORK

Before describing about the evaluation, we have to determine what to measured and evaluated. Since what we developed was a framework that runs on various hardware or communication platform which is provided by IaaS providers, so quantitative approach of evaluation does not make sense very much in this case. How the framework scales out is a matter of IaaS providers as well.

Considering this situation the most interesting issue of the framework is how the framework is usable against the

several milliseconds of communication latency in updating the memory data both from PLC and Application program on the data center. We define some typical usages and examine how the latency affects the usages.

We assume 5 milliseconds as communication latency from a data center to a plant floor and vice versa based on the RTT to the Google's data center. The QoS is guaranteed by the DiffServ or a dedicated line.

5.1 Usage of BI tools for decision making

As accessing most of ordinary WEB pages, user are used to wait for slow response in browsing. Several milliseconds of the latency is negligible comparing the latencies of WEB response including getting an IP address by DNS inquiry, three way handshake of TCP, authentication with third party certification (i.e. TSL).

As long as using the framework for decision making we do not see any problem since the latency occurred by reasons above cover the latency communication of the framework. It is true for the other operations such as getting result of any kinds of analytics, creating or modifying model-base design, running and examining simulation, inventory control and that do not require strict real time property.

5.2 Cell phone alarming

The framework allows disconnected operation as well. Users of the framework can invoke a watchdog process. If the process detects something wrong in the system, the process pages the users' cell phone. The latency of paging cell phone is in order of seconds, so the communication latency from the PLCs is negligible also in this scenario.

However, after accepting the alarm is another story. He or she gets the SCADA screen alarmed and checks what is happening. The latency of cellar system is tens of milliseconds and the latency to the plant floor is several milliseconds, so he or she can make immediate temporary amendment to the system and goes to the plant site if needed.

5.3 Modifying control program in the fly

Modifying control program is costly operation even by using a dedicated standalone engineering tool with direct connection (e.g. USB) to a controller. A problem on this issue is several milliseconds of latency is inserted into every message exchange.

Some PLCs like of us do not disclose application protocol for control program modification. In case of our PLCs roughly 800 messages are exchanged to modify a few ten lines of ladder code. The latency is inserted into every message, and that results in exhausting 4 more seconds comparing exchanging the messages via direct connection. Control program modification is not done so often, so it is not a big problem either.

5.4 System monitoring

Time constraint in system monitoring should be tight comparing issues described so far. However, the most of SCADA products asks system status by inquiring via HTTP which may contain many overheads resulting in latency more than several milliseconds. Our engineering tool on a PC asks the system via the propriety protocol on polling basis. Polling is not a good idea for taking account of scalability. The framework can provide better system in real time monitoring point of view as well.

5.5 Control of plant floor from the cloud

Control of plant floor from the cloud could be the most challenging issue for the latency [8]. Basically, operations which allow several ten milliseconds of delay can be controlled from the cloud. In typical usages of PLCs, the operations that require strict time constrains are the functions for the safety. To satisfy certain level safety, some dedicated hard wired system is deployed often without interaction of PLC [9].

Since the most of PLC usage cases, the scan time is tens milliseconds, we can say almost of operations being executed on PLCs are also able to control from the data center. However, the controls that require more restrict real time operation are not applicable.

If a control program which doesn't need synchronized operation the latency is not an issue.

We think the control from the data center as just complementary to the control executed on the PLCs such as part of the MES, calibrating parameters for worker's ability. Those operations don't need strict latency requirement.

6 CONCLUSION

We proposed a highly integrated service framework for manufactures. The framework allows to establish a session which is consisted of multiple of services with the dependency information regarding the services. The framework also allows disconnected operations such as alarming operator.

We also introduced a concept that the collaboration of service entities in the every field which achieves great efficiencies to realize sustainable society.

In the near future, every organization which consisting our society continues pursuing efficiency by the services that are enabled by the ICT. The services on the framework will be entity that are the parts of service that operate in the society's cyber physical system, and we believe the framework contributes for the sustainability of the global society.

REFERENCES

[1] http://www.unfpa.org

[2] http://cloud.cio.gov/success-stories

[3] D. Kovachev, D. Renzel, R. Klamma, and Y. Cao, "Mobile Community Cloud Computing: Emerges and Evolves," in *Proceedings of the First* International Workshop on Mobile Cloud Computing (MCC). Kansas

[4] Jain, S., Kumar, A., Mandal, S., Ong, J., Poutievski, L., Singh, A., ... & Vahdat, A. (2013, August). B4: Experience with a globally-deployed software defined WAN. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM* (pp. 3-14). ACM.

[5] Lewko, A., Okamoto, T., Sahai, A., Takashima, K., & Waters, B. (2010). Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption. In *Advances in*

Cryptology–EUROCRYPT 2010 (pp. 62-91). Springer Berlin Heidelberg

[6] Sakakura, T. (2015). An Empirical study of applying a Reflective-Distributed Memory for Automation Systems, Journal of Information Processing, IPSJ (under reviewing)

[7] Parfitt, J., Barthel, M., & Macnaughton, S. (2010). Food waste within food supply chains: quantification and potential for change to 2050. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 3065-3081.

[8] Hickson, I. (2011). The websocket api. *W3C Working Draft WD-websockets-20110929, September*

[9] Brown, S. (2000). Overview of IEC 61508. Design of electrical/electronic/programmable electronic safety-related systems. *Computing & Control Engineering Journal*, *11*(1), 6-12.

With a Little Help from My Native Friends: A Method to Boost Non-native's Language Use in Collaborative Work

Tomoo Inoue[*], Hiromi Hanawa[**], and Xiaoyu Song[**]

[*]Faculty of Library, Information and Media Science, University of Tsukuba, Japan
[**] Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
inoue@slis.tsukuba.ac.jp, hanawa@slis.tsukuba.ac.jp, s1465331@u.tsukuba.ac.jp

Abstract - Due to the fluency difference between a native speaker and a non-native speaker, the non-native speaker often feel difficulty in joining conversation. This can be a loss from shared knowledge point of view. In this paper, we propose a simple method to boost language use of the non-native speaker. A little help of the native speaker can be a big help to the non-naitive speaker.

Keywords: Inter-cultural communication, Conversation support, Language Provide up to five keywords to be used for future on-line publication searches and indexing.

1 INTRODUCTION

The chances we communicate with a person of different mother tongue have been increasing more than ever [1]. Thus it is important to support such communication. When we have conversation in one language such as English, people's fluency of the language can be diverse and it often causes unbalanced speech opportunity [2]. For example, a Japanese whose mother tongue is not English often feels difficult in keeping up with the conversation between native English speakers and has little chance of speaking consequently. This is not very productive from shared knowledge point of view. If the Japanese is able to understand the conversation more in real time, and can get a chance of speaking even though his/her language skill is not very high, he/she can contribute to the conversation more. Conversation by native speakers is more shared with him/her, and all the participants' thinking is shared by all, which makes the conversation more valuable.

In this paper, we propose a simple method to help non-native speaker's conversation with a native speaker and explain its initial evaluation[1].

2 RELATED WORK

To ease the burden of non-native speaker's participation in audio conferencing, a method to insert short silent periods in the native speaker's speech thus giving a non-native speaker extra time for cognitive processing was proposed [3].

To help understand culture-dependent word in intercultural communication, a system to add extra explanation to such words obtained by voice recognition was developed [4].

However more simple methods and systems are desirable considering the need for robustness and availability.

[1] This work was financially supported by JSPS KAKENHI No. 26330218.

3 PROPOSED METHOD

We propose a simple method to help non-native speaker's conversation with a native speaker.

The method is to ask the native speaker input keywords of what he/she speaks in real time from the keyboard, and share those keywords with the non-native speaker. This very simple method is not yet proposed in this cross-language communication context as far as we know of. Also the usefulness of the method is not yet investigated.

The theoretical essence of the proposed method is asymmetry of load and benefit. For the native speaker, to input keyword while speaking is not considered to be difficult. Doing this can be extra burden but is supposed to be not very heavy. For the non-native speaker, the keyword is expected to be very helpful in understanding what the native speaker speaks. The benefit would be great. Clearly the direct benefit of this method is not reciprocal. The native speaker only helps the non-native speaker. However if the non-native speaker's language use increases as a consequence of increased understanding of the conversation, the conversation become more communicative and this would benefit all the participants of the conversation.

4 INITIAL EVALUATION 1

To investigate the proposed method, conversation with the proposed method and one without it was compared.

4.1 Participants

Five pairs of a native English speaker and a non-native English speaker participated. The native speakers were either born and raised in English speaking countries, or educated in English over 8 years. The non-native speakers lived in English speaking countries within 2 years, but have over 700 TOEIC score to guarantee essential communication capability in English.

4.2 Design

Conversations in 2 different conditions were compared. One was with the proposed method and the other was without the method, which was the control condition. Five pairs participated in each condition. Because we were in the initial evaluation stage of the proposed method, the combination of the participants for making pairs were mixed, and the participation order was not strictly balanced.

Figure 1: Experimentation setup.

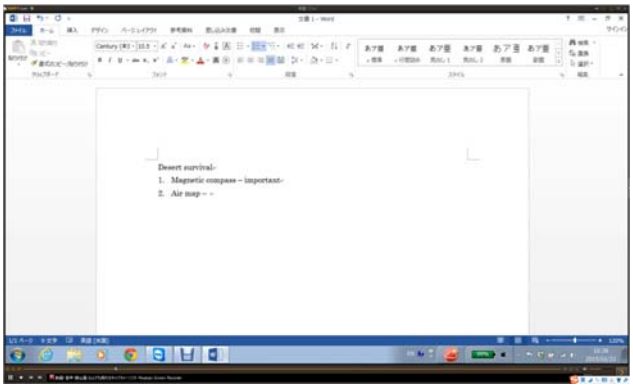


Figure 2: The shared screen in the proposed method.

However we thought this results could somehow foresee the results by more controlled one.

4.3 Setup

The experimentation was conducted in a quiet room. To avoid affection from other communication media than audio because of the basic investigation, the pair took seats back to back. In front of each participant was a laptop whose screen size was 15.6 inch and the resolution was WXGA (1366 x 768) on the desk (Figure 1).

The screen of the proposed-method condition is shown in Figure 2. A shared screen software Team Viewer version 10 was used for the non-native speaker sharing the screen with the native speaker, and Microsoft Word 2013 was used for the text input of the native speaker. This way the non-native speaker could see the keyword the native speaker put in in real time.

4.4 Procedure

The pairs got involved in solving the survival problem collaboratively [5]. The problem has been often used in group discussion. In the problem, the airplane or the rocket in which the participants get on lands on the desert, the North Pole, or the moon in emergency. The participants need to discuss and decide the priority order of the gadgets to be brought with from the ones found in the vehicle for survival in this situation.

Table 1: The average number of a non-native speaker's utterances

	Pair1	Pair2	Pair3	Pair4	Pair5	Ave.
Proposed	27	31	35	33	31	31
Control	36	8	16	20	23	21

Table 2: The average number of turn taking of a non-native speaker in a minute

	Pair1	Pair2	Pair3	Pair4	Pair5	Ave.
Proposed	4.2	4.8	3.9	3.7	3.1	3.9
Control	3.6	1.5	2.5	2.3	2.1	2.4

Table 3: The result of the questionnaire

	Non-native		Native	
	Proposed	Control	Proposed	Control
Q1	4.8	5.8	5.6	6
Q2	3.8	5.2	5.6	6
Q3	3.6	4.2	3.2	2.4
Q4	4.6	5.2	4.8	5.6
Q5	2.6	1.8	1.8	3.4

Because each pair worked twice, one in the proposed method condition and the other in the control condition, two different places were selected for each session. The combination of the condition and the place was balanced.

The given discussion time was 10 minutes each. To start conversation smoothly, 3 minutes were given for individual thinking of the solution before discussion.

4.5 Collected Data

The experiment was videotaped. The questionnaire was filled out after each 10 minutes session. The questionnaire items were as follows.

Q1: I felt easy to understand what my partner said.
Q2: I felt easy to say what I thought.
Q3: I felt frustrated.
Q4: I could communicate with my partner naturally.
Q5: I think I looked at my PC screen frequently.

4.6 Result

In this initial evaluation, a few descriptive values of the conversation were examined. The average number of a non-native speaker's utterances is shown in Table 1. The average number of turn taking of a non-native speaker in a minute is shown in Table 2.

The result of the questionnaire is shown in Table 3. The items were answered with the 7-point scale where 1 meant strongly disagree to 7 meant strongly agree.

4.7 Discussion

From Table 1 and Table 2, it is seen that the non-native speakers speak more actively with the proposed method than without it. From Table 3, no strong tendency was observed.

Table 4: The average number of a non-native speaker's utterances

	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	Ave.
Silence	46	28	43	41	48	26	39
Control	50	22	55	24	40	15	34

Table 5: The average number of turn taking of a non-native speaker in a minute

	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	Ave.
Silence	3.1	2.6	2.0	2.3	1.9	1.7	2.2
Control	2.7	6.2	1.9	4.1	3.5	1.5	3.3

Table 6: The result of the questionnaire

	Non-native		Native	
	Silence	Control	Silence	Control
Q1	5.7	5.8	5.3	5.8
Q2	3.7	4.3	3.7	5.5
Q3	4.3	3.0	4.5	3.2
Q4	4.3	5.5	3.3	5.5
Q5	0.5	0.5	1.2	1.0

5 INITIAL EVALUATION 2

Through the initial evaluation 1, the comparison between the conversation with the proposed method and the conversation without the proposed method, it was often observed that the native speaker did not speak when typing the keyword. This observation led to the idea that we should investigate possible cause of the active speaking of the non-native speaker in the proposed-method condition. Two causes were the keyword and the silence. It was possible that the keyword prompted the non-native speaker to speak. It was also possible that the silence of the native speaker when typing prompted the non-native speaker to speak. In the previous experiment these possible causes co-occurred most of the times. Thus to clarify the possible causes, conversation with the native speaker's occasional silences and one without them was compared.

5.1 Participants

Six pairs of a native Japanese speaker and a non-native Japanese speaker participated. All the participants were different from the previous experiment. The native speakers were born and raised in Japan. The non-native speakers were N1 holders of the Japanese-Language Proficiency Test to guarantee essential communication capability in Japanese.

5.2 Design

Conversations in 2 different conditions were compared. In one condition, the native speaker keeps silence occasionally in the same manner as in typing the keyword. The other was without such silence, which was the control condition. Six pairs participated in both conditions, and their participation order was balanced.

5.3 Setup

The setup was the same with the previous experiment, although the participant did not use the keyboard of the laptop.

5.4 Procedure

As in the previous experiment, the pairs got involved in solving the survival problem collaboratively. The combination of the condition and the place was balanced. The procedure was also the same with the previous experiment, but the experimenter asked the native speaker to stop speaking by showing a sign board only to the native speaker. This way the native speaker stopped speaking while the non-native speaker knew of the instruction. The silence time was about 20 second in a minute, which was roughly the same with the typing time in the proposed-method condition of the previous experiment.

5.5 Collected Data

Collected data was the same with the previous experiment.

5.6 Result

As in the previous experiment, the average number of a non-native speaker's utterances is shown in Table 4. The average number of turn taking of a non-native speaker in a minute is shown in Table 5.

The result of the questionnaire is shown in Table 6.

5.7 Discussion

From Table 4, it might be that the non-native speakers speak more in the silence condition, but its increase from the control condition is not more visible than in the previous experiment.

From Table 5, it is seen that the number of turn-taking of a non-native speaker is fewer in the silence condition, and the ratio of the silence condition to the control condition is close to the ratio of speaking periods of the native speakers in those conditions. When the native speaker stopped speaking, even though the instruction by the experimenter was not known, the non-native speaker felt somewhat unnatural, and tended not to speak frequently in the period.

From Table 6, strong tendency was not observed in most items, but Q2 "I felt easy to say what I thought" and Q4 "I could communicate with my partner naturally" were highly rated by the native speakers in the control condition. These answers seem to reflect restriction and unnaturalness of speaking in the silence condition.

The results indicate that simply making silence does not take over keyword input.

6 CONCLUDING REMARKS

An opportunity of conversation with a non-native speaker has been increasing. It is natural that when a person uses less fluent language, he/she becomes less eloquent. This is

not very desirable considering the productivity of conversation.

To solve this problem, a simple method was proposed. With a little help of typing the keyword from a native speaker, it is expected that the language use of a non-native speaker is boosted.

Although the evaluation of the method is still its initial stage and more rigorous experimentation is needed, the utility of the method is suggested.

REFERENCES

[1] R. Fujita, Yokoso Japan! : The Significance of Intercultural Communication Competence, The Journal of Communication Studies, Vol.30, pp.3-14 (2009).

[2] N. Yamashita, A. Echenique, T. Ishida, and A. Hautasaari, Lost in transmittance: how transmission lag enhances and deteriorates multilingual collaboration, Proc. CSCW 2013, pp. 923-934 (2013).

[3] N. Yamashita, A. Echenique, H. Kuzuoka, T. Ishida, and A. Hautasaari, A Method That Ease the Burden of Non-native Speakers' Participation in Conference Calls, IPSJ Journal, Vol.54, No.6, pp.1794-1806 (2013).

[4] K. Okamoto, T. Yoshino, Development of Face-to-face Intercultural Communication Support System Using Visualized Keyword of Conversation, Forum on Information Technology, Vol.8, No.3, pp.393-396 (2009).

[5] J.C, Lafferty, P.M. Eady, and J. Elmers, The desert survival problem. Experimental Learning Methods, Human Synergistics, Plymouth, MI, USA (1974).

Panel Discussion

On-vehicle Information Devices Based on User's Context

Ryozo Kiyohara

Kanagawa Institute of Technology

ITS: Research Topics

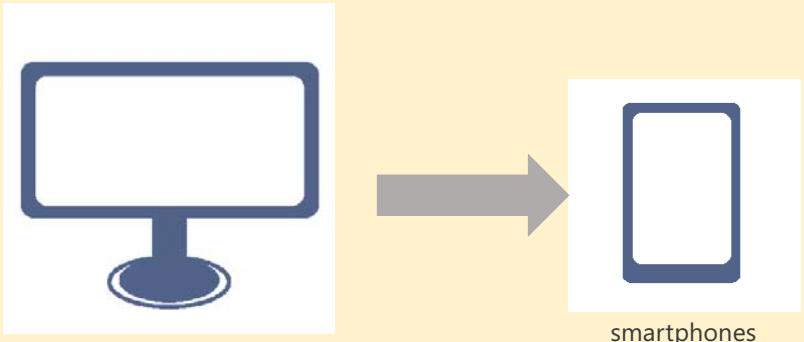
神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

1. Automatic Driving Vehicle
   - A) Sensing Technology(IoT)
   - B) Network Technology(GNSS)
   - C) Etc..
2. Comfortable Driving
   - A) M2M/IoT,
   - B) V2X
   - C) Etc.
3. Safety Driving
   - A) V2X
   - B) Platform(OS, Security, Software Updating)
   - C) UI for drivers distraction problem
   - D) Etc.

Background

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- Smartphones are replacing with...

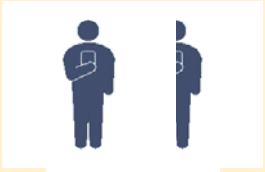


Background

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- The three reasons

Widely used



Many smartphones applications

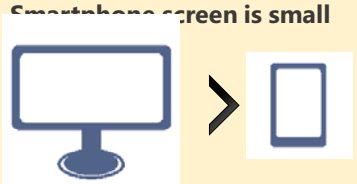
No additional costs

Background

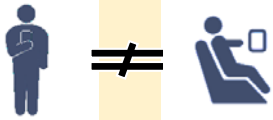
神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- Several problems of smartphone for on-vehicle information device

Smartphone screen is small



Difference between typical operation and while driving operation

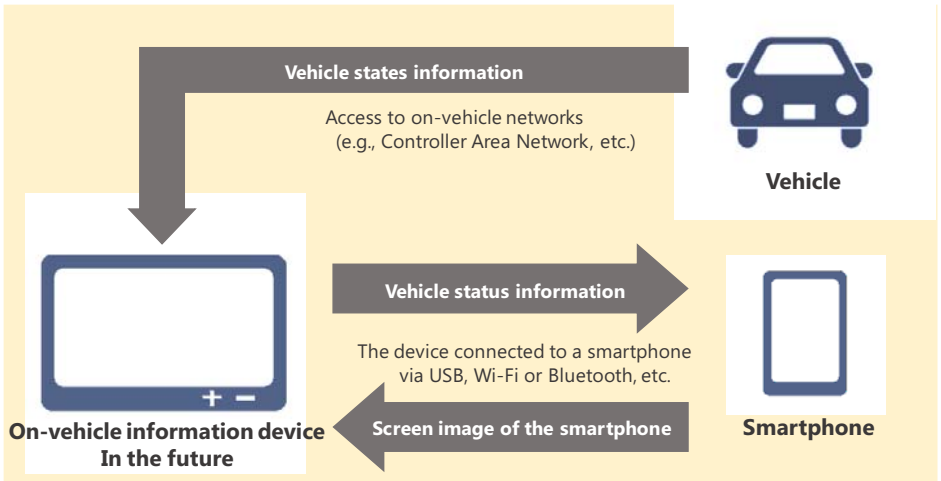


Applications are not inoperative while t driven

Background

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- The on-vehicle information device in the future



Background

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- An issue of smartphones
  for on-vehicle information devices

Drivers must operate the UI while they are stopped.
For example, while vehicle is stopping at a red light.

- **Drivers have a**
  limited time to operate the smartphone.

Purpose

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- Context-aware technologies be applied to
  the UI control system.

Context-aware technologies

A limited time to operate the smartphone
can be more effectively.

Purpose

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- The driver want to changed music.

Stopping at a red light

Launching audio application by the UI control system

Selecting music

Starting to move

Launching map application by the UI control system

Vehicle context

Driving

Stopping

Driving

The driver can make effective use of the limited time to operate the smartphone.

Related works

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

"ContextPhone: a prototyping platform for context-aware mobile applications"

Mika Raento, Antti Oulasvirta, Renaud Petit, and Hannu T T Toivonen

Pervasive Computing, IEEE, Vol.4, Issue2, pp.51-59 (2005)

Related works

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

**"The Sentient Car:
Context-Aware Automotive Telematics"**

Pablo Vidales and Frank Stajano

4th International Conference on Ubiquitous Computing, pp.47-48

Proposed method

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- The UI control system applied to context-aware technology will require the following processes.

1. To acquire the data to recognize the context.
- **Vehicle status information**
- **Data acquired utilizing the smartphone's sensors**
- **Data accessed via the Internet**

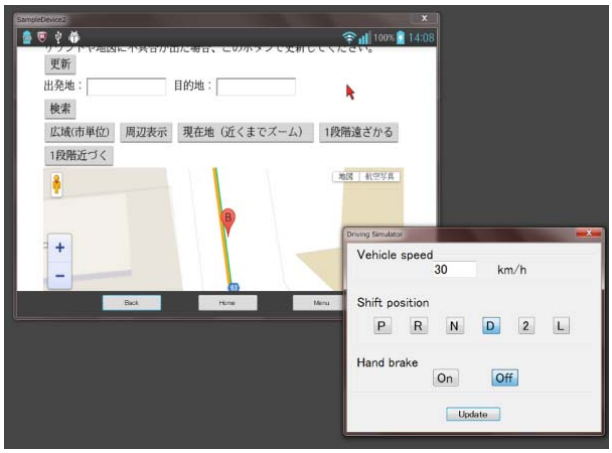
2. To recognize the context

3. To use of context

Proposed method

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- Implementing the prototype



Proposed method

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- Our expected effect on the number of times of the operation

	Conventional Device	Proposed Method Suggested function is wanted by driver	Proposed Method Suggested function is unwanted by driver
Operation	3	1	3

・**When the suggested function is the function that the driver wanted**
The proposed system reduces the number of steps from three to one.
We confirmed that **our proposed system is effective**.

・**If the suggested function is not the function desired by the driver**
The on-vehicle information device **must be operated in the conventional manner**.

Proposed method

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

- Our expected effect on the number of times of the sight line deviation

	Conventional Device	Proposed Method Suggested function is wanted by driver	Proposed Method Suggested function is unwanted by driver
Sight line deviation	3	1	3

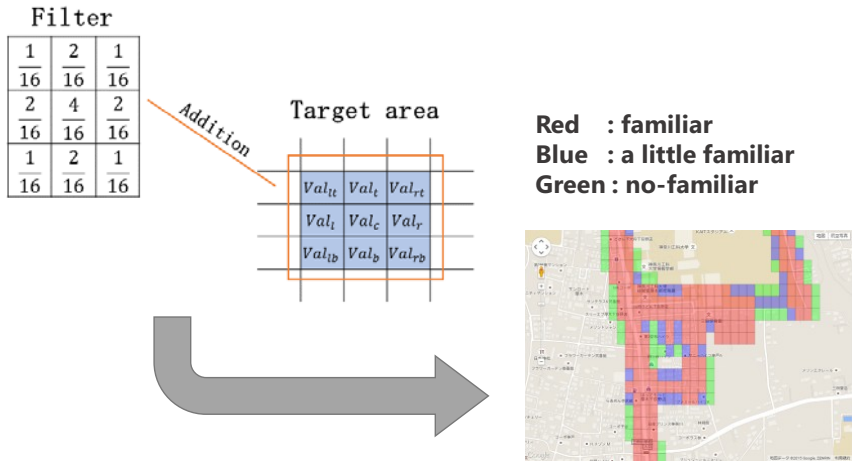
- **When the suggested function is the function that the driver wanted**
  The drivers to check the screen only once, and only a single selection is necessary.
  We confirmed that **our proposed system is effective**.

- **If the suggested function is not the function desired by the driver**
  The on-vehicle information device **must be operated in the conventional manner**.

Recognition of contexts

神奈川工科大学 KANAGAWA INSTITUTE OF TECHNOLOGY

Recognition of location context



Conclusions

神奈川工科大学
KANAGAWA INSTITUTE OF TECHNOLOGY

- In the future plan

The system is not effective when the suggested function is unwanted by driver.

We will continue a series of separate studies to recognize contexts.

Sept. 8, 2015

IWIN2015 Panel Session @ Amsterdam, the Netherlands

Promising Techniques toward Future Intelligent Transport Technologies

Multi-GNSS

(Global Navigation Satellite System)

Tomoya KITANI

(Shizuoka University, Japan)

Shizuoka_univ. Kitani_lab. BikeInformatics

Summary

Sept. 8, 2015

2

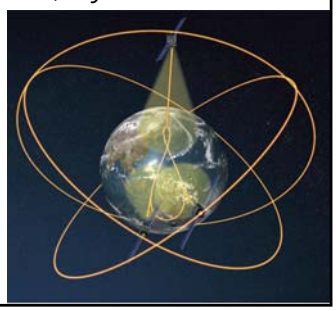
- How do you choose your research topic?
- To introduce a recommended research topic
  - **Multi-GNSS (global navigation satellite system)**
    - In fact, it is a “blue-ocean” research topic!!

Navigation Satellite Systems

Sept. 8, 2015

3

- **GNSS**: Global Navigation Satellite System
  - **GPS** (Global Positioning System) by U.S.
  - **GLONASS** by Russia
  - **BeiDou（北斗）**, a.k.a Compass, by China
  - **Galileo** by EU (planned)
- RNSS: Regional NSS
  - IRNSS by India (planned)
  - **QZSS** (quasi-zenith satellite system) by Japan



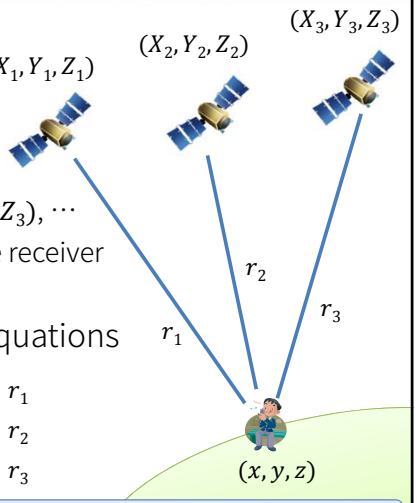
How to localize with GNSS

Sept. 8, 2015

8

- Location to be localized
  - Receiver's position (x, y, z)
- Given information
  - Location of GNSS satellites (X_1, Y_1, Z_1) , (X_2, Y_2, Z_2) , (X_3, Y_3, Z_3), ⋯
  - **Measured range** between the receiver and each satellite r_1, r_2, r_3, ⋯
- To solve the simultaneous equations

$$\begin{cases} \sqrt{(x-X_1)^2+(y-Y_1)^2+(z-Z_1)^2} = r_1 \\ \sqrt{(x-X_2)^2+(y-Y_2)^2+(z-Z_2)^2} = r_2 \\ \sqrt{(x-X_3)^2+(y-Y_3)^2+(z-Z_3)^2} = r_3 \end{cases}$$



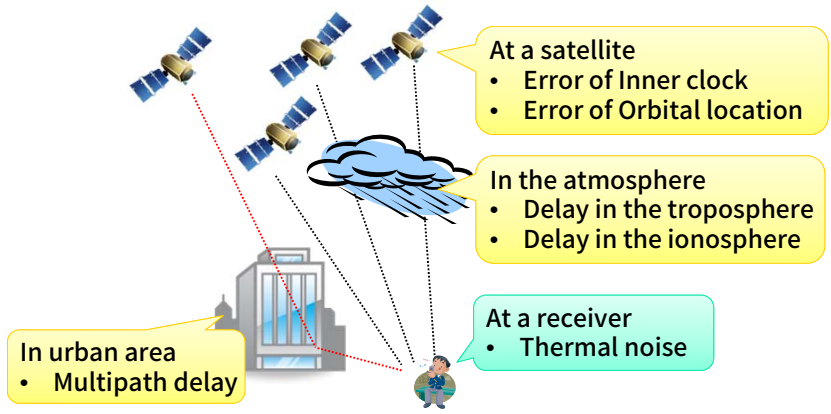
In practice, the clock error in a receiver is solved as the 4th unknown value.

Main factors of GNSS positioning error

Sept. 8, 2015

9

- Range to satellites includes many errors



- **Large number of satellites can make positioning error small**
  - With the least-square method, etc.

Current situation

Sept. 8, 2015

10

- Some receivers for consumers have been released that can capture two GNSS recently
  - GPS & QZSS, GPS + GLONASS, GPS + BeiDou
  - Positioning Chip
    - u-blox 8 series
  - Integrated Smartphone Chip
    - Broadcom

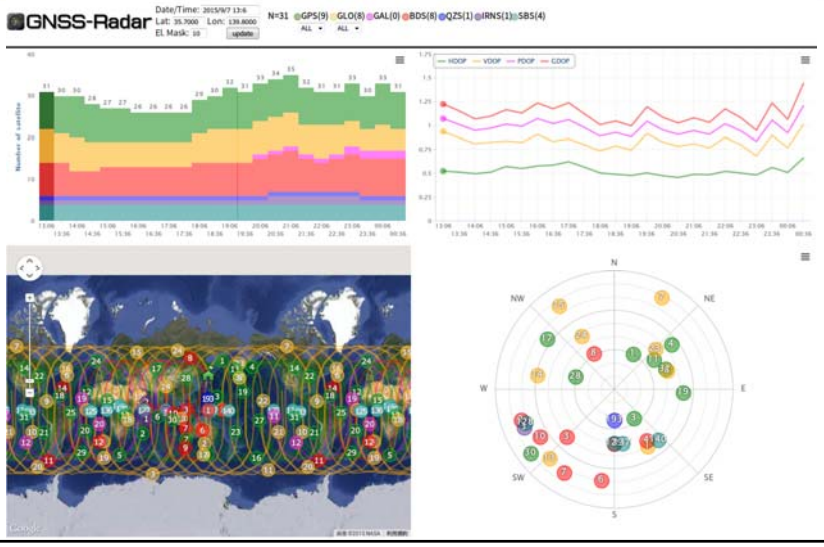
https://www.u-blox.com/

GNSS-Radar

http://www.taroz.net/GNSS-Radar.html#

Sept. 8, 2015

13



GPS vs. multi-GNSS

Sept. 8, 2015

14

- **GPS**
  - *# of visible satellites:* usually 4 to 8 (up to 12)
  - *Positioning accuracy:* more than 10 meters
- **Multi-GNSS**
  - *# of visible satellites:* usually more than 15 now
    - **It will be 30 over U.S. and 60 over Asia in 2035**
  - *Positioning accuracy:* a few meters now
    - It will be a few centimeters in the future
    - It could be a few centimeters even now for consumers **if novel good positioning algorithms are available**

DTN Based Information Processing Platform for Disaster Situation

Takaaki Umedu(Shiga University)
Takamasa Higuchi, Akira Uchiyama, Akihito Hiromori, Hirozumi Yamaguchi and Teruo Higashino (Osaka University)
Keeichi Yasumoto(NAIST)

Background

- Fast Grasping of Disaster Situation is Needed for Efficient Rescue, Relief and Evacuation
- There are Several Researches about Disaster Information Gathering
  - Safety Confirmation, Map of Available Paths, Distribution of Injured People, etc.
  - Each Study Focusing on Specific Purpose To Design Method
    - Each of Them is Efficient for Single Purpose, However Existence of Various Purposes and Requirements is not Well Considered.

Concept

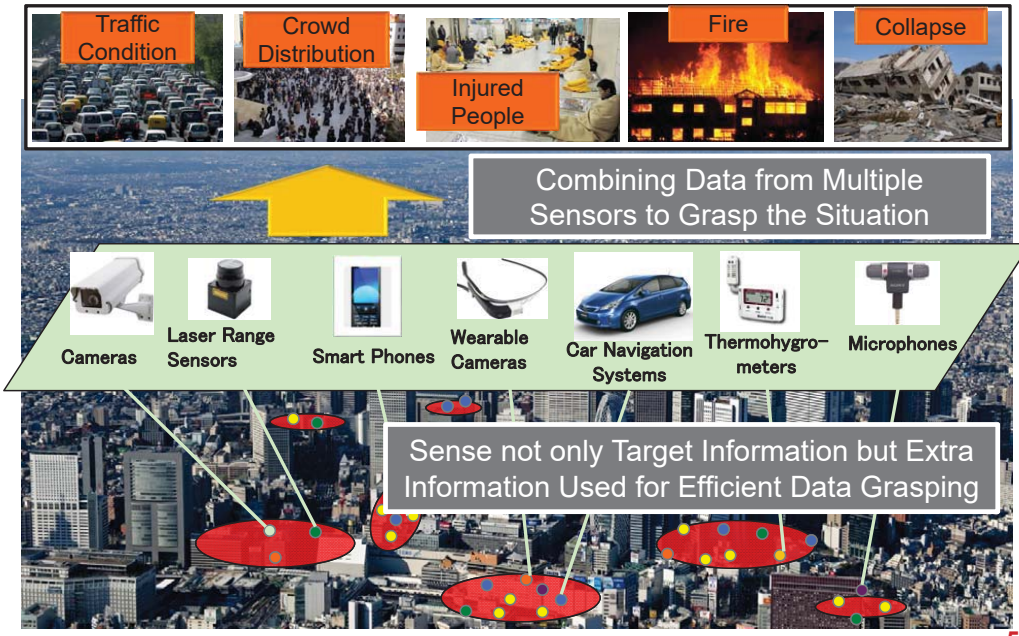
Information Collecting and Processing Platform for Various Requirements from Users

- Independent from Infrastructure: DTN Based
- Powerful Computation Function: Distribute Task Assignment Widely
  - Considering Task Assignment, Processing and Routing Together
  - Assuring Robustness by Speculative Execution
    - Task Assignments Often Missed Because of Nodes' Mobility
    - Computation Power might not be Enough
    - etc.

Existing Works

- There are Several Researches about Disaster Information Gathering
  - Safety Confirmation, Map of Available Paths, Distribution of Injured People, etc.
  - Each Study Focusing on Specific Purpose To Design Method
    - Each of Them is Efficient for Its Purpose, However Existence of Various Purposes and Requirements is not Well Considered.
- Serendipity [MobiHoc'12] : DTN Based Technology
  - Nodes Request Tasks to Encountering Nodes
    - Tasks: Heavy Data Processing Like Voice Recognition
  - Reduce Total Data Processing Time

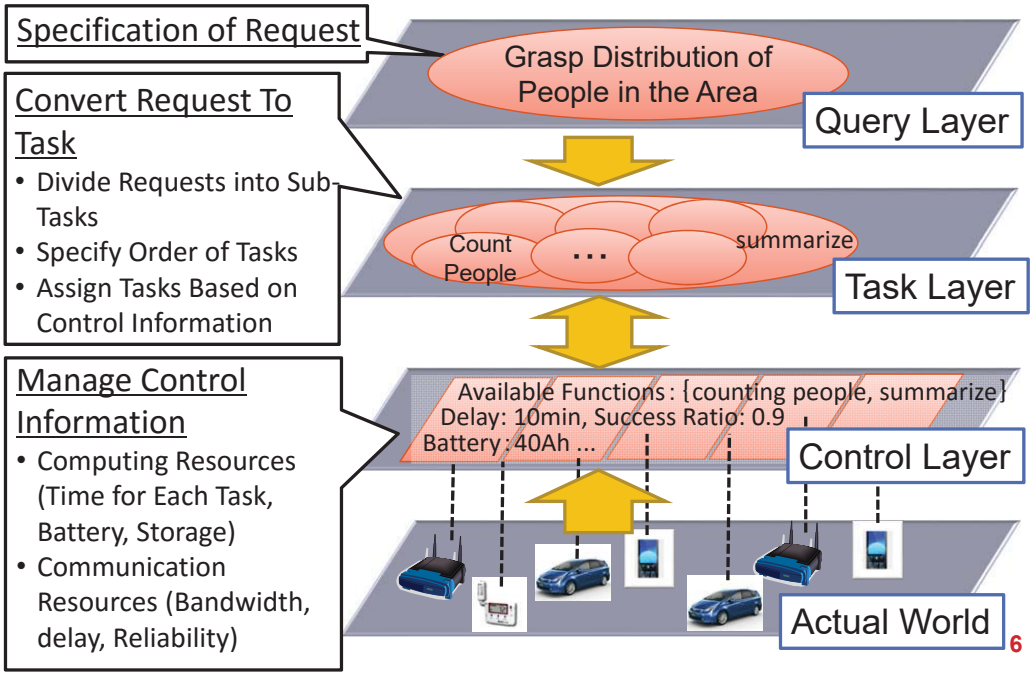
Grasping Urban Situation



IWIN2015 Panel Sesison

5

Overview of Our Platform



6

Elemental Techniques Needed to Realize our Platform

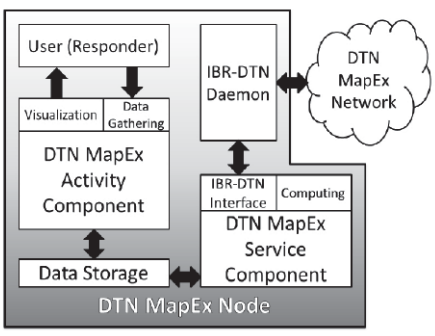
- Specification Language: The Design How Users Specify Requirements
  - Do not Require Detailed Knowledge about the Platform: Highly Abstracted Specification should be Accepted
- Task Decomposition and Assignment Mechanism
  - Considering Both Control Information and User Requirements, to Decompose Requests into Sub-Tasks and Make Assignment Them to Nodes and Routing
- Control Information Collecting Mechanism
  - Gather the Information Such as Available Computing Resources of Each Node, When and Which Node Pair Will Encounter
  - Includes Incorrect Information Based on Prediction

Case Studies Applying ITS
- **DTN MapEx**: Automated Map Generation App for Android
  → Discussed Problems through Implementation(Bottom-up Approach)
- Crowded People Count Estimation Using Photos Taken by Smart-phones
  → Heavy Load Computation Required on Each Node

DTN MapEx

E. M. Trono, Y. Arakawa, M. Tamai, K. Yasumoto: DTN MapEx: Disaster Area Mapping through Distributed Computing over a Delay Tolerant Network, ICMU 2015.

- Automated Map Generation App Based on DTN
  - Record and Share Travelled Trajectories
  - Additional Information can be Recorded, Placed onto Map Data and Shared
    - Text, Photo, Voice, Video, etc.
  - Information Transmission : PRoPHET/Epidemic + WiFi Direct



IWIN2015 Panel Sesison Structure of System

Map Data(OSM) Available (with Internet Access)

GPS Trajectory (w/o Internet Access)

Pilot Experiment

Tested: Data Transmission Time(How Much Data can be Exchanged Even Short Encountering Time)

Testing Scenario: Nexus 4 (Sender) →Nexus 5 (Receiver)

Application: DTN MapEx, IBR-DTN Daemon + Sharebox

Interface: Wi-Fi Direct

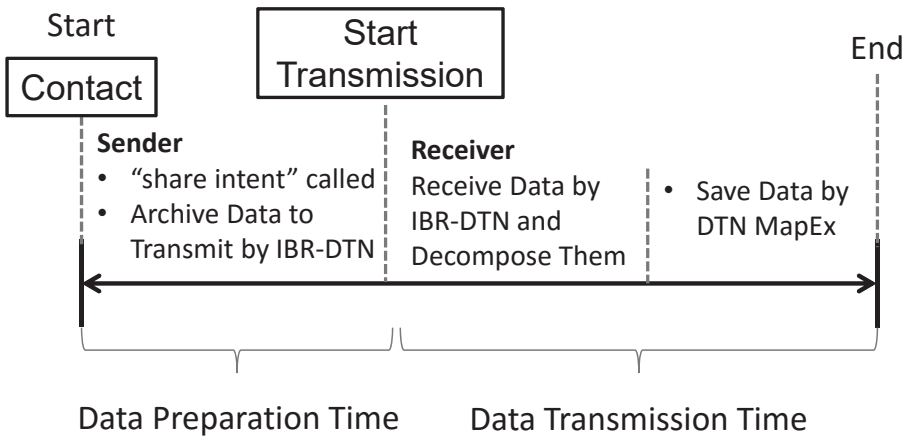
Transmitted Data:

Data Type	Dataset Details
Images (*.JPEG)	• 1, 2, 5, 10, and 20 images • Image batch size ranged from 757.8 kB to 21650.6 kB
Text (GPS traces in *.CSV format)	• 51.2, 153.6, 256, 512, 984 kB (corresponds to GPS Logs of 1, 3, 5, 10, 20 km)

IWIN2015 Panel Sesison

9

Tested Index: Data Transmission Time



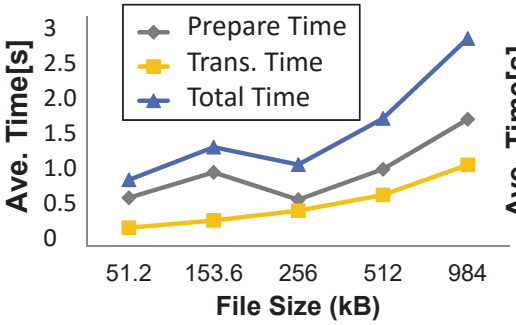
Test 10 Times to Take Average

IWIN2015 Panel Sesison

10

Results of Pilot Experiment: Data Transmission Time

- Preparation Time is Much Longer Than Transmission Time
  - Overhead of Processing "Intent" is too Large
  - Some Ideas to Hide the Overhead Needed
- For Large Data, Transmission Time may Exceed Contact Time
  - File Dividing Mechanism Needed



Corresponds to Text Data of GPS Trajectory of 1, 3, 5, 10, 20Km

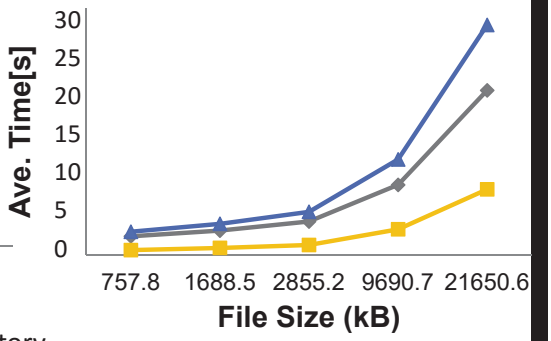
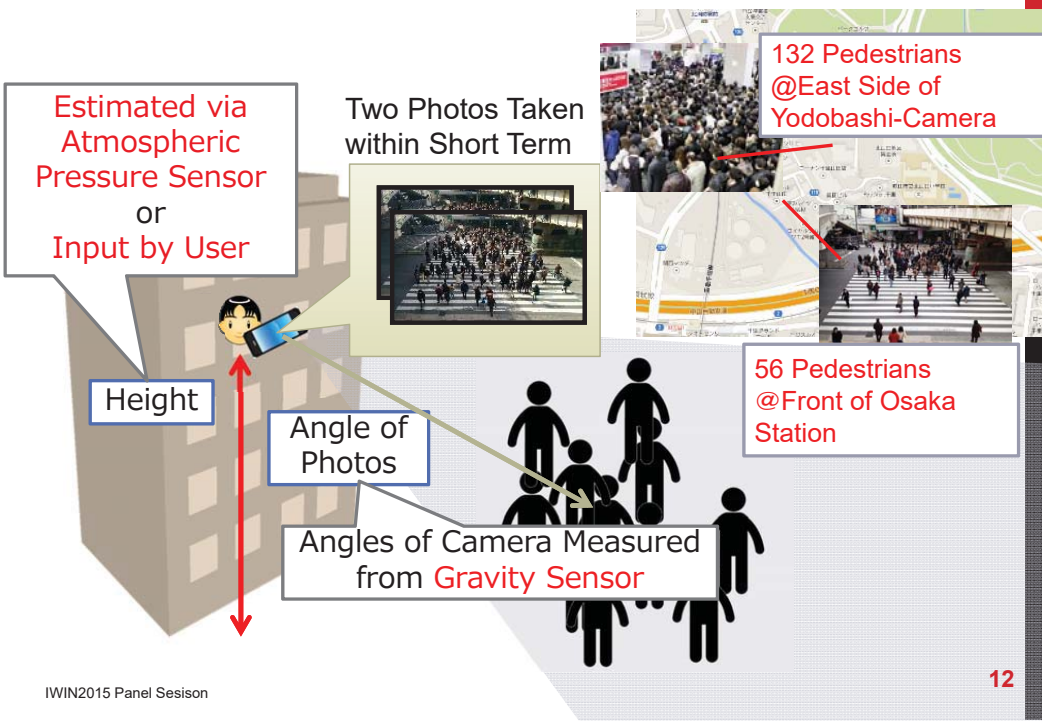


Image Data

Counting Crowded People



Key Idea for Counting Crowded People

$$\text{\# of People} = \sum_{y=0}^{y\max} g(y)\, f(y)$$

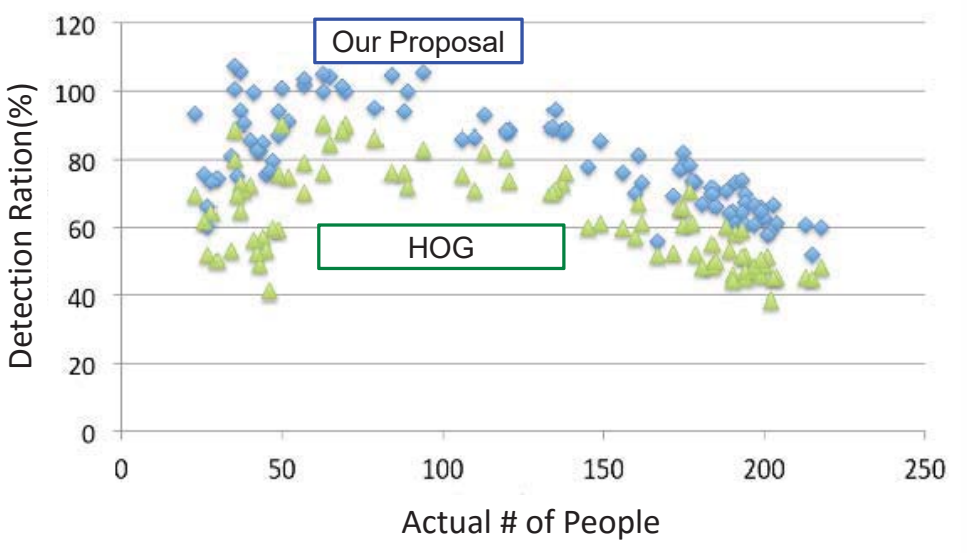
# of People / Pixel in the Photo for y

# of While Pixels in the Line for y

Evaluation Environment

- Take Photos Using Nexus5
  - # of Samples : 87 Pairs
  - Day : 2015/01/25(Sun.) around 14pm.
  - Place : Front of JR Osaka Station
    - Height:5.8m, Angle:25°
- Counting Process is Implemented Using OpenCV and Executed on a Laptop PC
- Evaluation Index: Detection Ratio = # of People Detected / Actual # of People
  - Actual Number is Counted by Hand

Comparison to Existing Method



IWIN2015 Panel Sesison

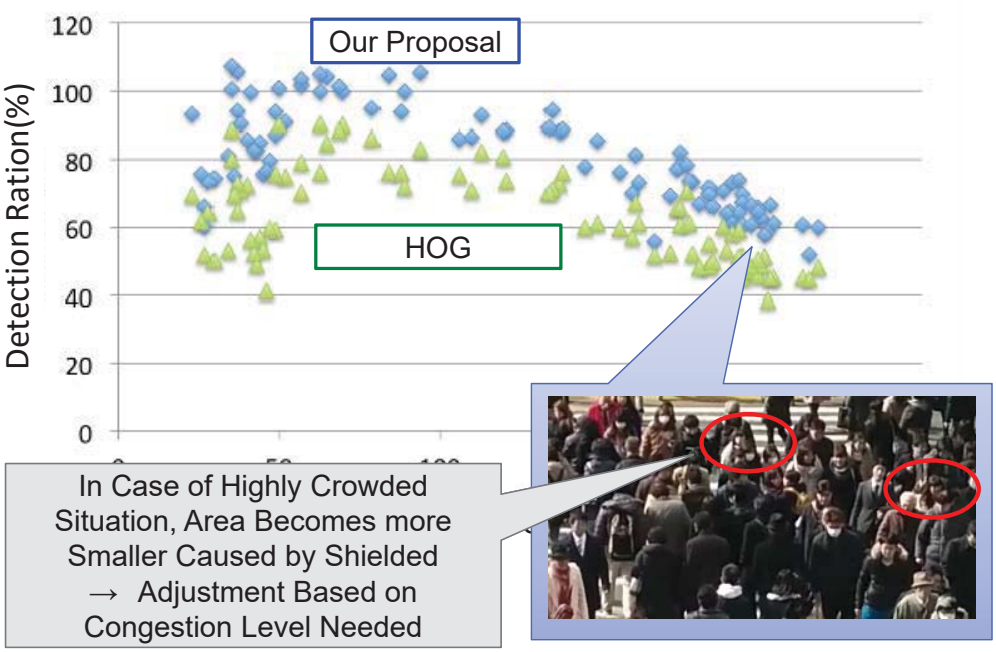
15

Comparison to Existing Method

IWIN2015 Panel Sesison

16

Comparison to Existing Method



17

Conclusion

- Summary
  - Designing Platform to Grasp Information
    - Discussing about Elemental Techniques through Case Studies
- Future Work
  - Detailed Design of Our Platform
  - Implement Each Elemental Techniques
    - Design Users Request Specification Language
    - Implement Converter from Request to Tasks
    - Implement Control Layer
    - Design Task Assignment Algorithm

IWIN2015 Panel Sesison

18