International Journal of Informatics Society

10/19 Vol. 11 No.2 ISSN 1883-4566

Informatics Society

Aims and Scope

The purpose of this journal is to provide an open forum to publish high quality research papers in the areas of informatics and related fields to promote the exchange of research ideas, experiences and results.

Informatics is the systematic study of Information and the application of research methods to study Information systems and services. It deals primarily with human aspects of information, such as its quality and value as a resource. Informatics also referred to as Information science, studies the structure, algorithms, behavior, and interactions of natural and artificial systems that store, process, access and communicate information. It also develops its own conceptual and theoretical foundations and utilizes foundations developed in other fields. The advent of computers, its ubiquity and ease to use has led to the study of informatics that has computational, cognitive and social aspects, including study of the social impact of information technologies.

The characteristic of informatics' context is amalgamation of technologies. For creating an informatics product, it is necessary to integrate many technologies, such as mathematics, linguistics, engineering and other emerging new fields.

Guest Editor's Message

Yuichi Bannai

Guest Editor of Thirty-second Issue of International Journal of Informatics Society

We are delighted to have the Thirty-second issue of the International Journal of Informatics Society (IJIS) published. This issue includes selected papers from the Twelfth International Workshop on Informatics (IWIN2018), which was held at Salzburg, Germany, Sept. 9-12, 2018. The workshop was the twelfth event for the Informatics Society, and was intended to bring together researchers and practitioners to share and exchange their experiences, discuss challenges and present original ideas in all aspects of informatics and computer networks. In the workshop 26 papers were presented in seven technical sessions. The workshop was successfully finished with precious experiences provided to the participants. It highlighted the latest research results in the area of informatics and its applications that include networking, mobile ubiquitous systems, data analytics, business systems, education systems, design methodology, intelligent systems, groupware and social systems.

Each paper submitted IWIN2018 was reviewed in terms of technical content, scientific rigor, novelty, originality and quality of presentation by at least two reviewers. Through those reviews 20 papers were selected for publication candidates of IJIS Journal, and they were further reviewed as a Journal paper. We have three categories of IJIS papers, Regular papers, Industrial papers, and Invited papers, each of which was reviewed from the different points of view. This volume includes six papers among those accepted papers, which have been improved through the workshop discussion and the reviewers' comments.

We publish the journal in print as well as in an electronic form over the Internet. We hope that the issue would be of interest to many researchers as well as engineers and practitioners over the world.

Yuichi Bannai is a professor of Information Media Department at Kanagawa Institute of Technology, Japan. After receiving ME from Waseda Univ., he joined Canon Inc. He also received MS from Michigan State Univ. in 1988 and Ph. D from Keio Univ. in 2007. His research interests include Virtual/Augmented Reality with five senses interaction and development of olfactory display. He received the best paper awards from IPSJ and ICAT07. He is a member of ACM, IEEE CS, IPSJ, VRSJ, the Japanese Society of AI, and the Japanese Association for the study of Taste and Smell.

Regular Paper

Investigation of the Influence of Scent on Self-Motion Feeling by Vection

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

[*] Graduate School of Engineering, Kanagawa Institute of Technology, Japan
[**] Department of Information Media, Kanagawa Institute of Technology, Japan
[***] Faculty of Design, Kyushu University, Japan
s1885007@cce.kanagawa-it.ac.jp, bannai@ic.kanagawa-it.ac.jp, seno@design.kyushu-u.ac.jp

Abstract - Vection is an illusion that gives the feeling of motion in the absence of bodily movement. This phenomenon may occur with the presentation of a screen displaying patterns with optical flows. In an immersive environment that uses a head mounted display (HMD), including virtual reality (VR) systems, vection is frequently induced. In recent years, many trials have been conducted in which scents are displayed with VR systems. We aimed to investigate the effects of scents on vection perception. In the current experiments, the subjects were seated in front of an olfactory display while wearing an HMD and were presented with moving images to induce vection, under several conditions: scent presentation, sound presentation, and presentation of no additional stimuli. We found that scent stimuli do not affect the perception of vection. However, there were many positive correlations between perceived vection strength and perceived scent strength, especially in the lavender condition. It seems that there is some relation between the lavender scent and vection.

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

1 INTRODUCTION

Vection is the visually induced illusion of self-motion [1], which may be felt when viewing a screen that is displaying patterns with optical flows. It is known that vection is induced not only by visual stimulation but also by auditory and somatosensory stimulation. Sakamot et al. [2] have reported that sound images of movement from front to back or from back to front induce linear self-motion perception and that the self-motion direction is influenced by the direction of the motion of auditory stimuli. This shows that auditory information also has a great influence on self-motion. Vection is also caused by somatosensory sensation. Murata et al. [3] performed experiments in which participants wore eye masks and were presented with white noise through a pair of earphones. A horse riding machine was used to produce bodily movement in the participant. Almost all of the participants in this condition felt the sense of forward motion on presentation of a constant stream of air to their front. As with the sense of vision and hearing, sense of smell can detect the object without touching it [4], so it is thought that olfactory stimulation may cause a sense of self-motion. However, research on olfactory vection has not been conducted.

It is known that auditory stimuli and cutaneous sensory stimuli promote vection perception by presenting them with visual stimuli. Riecke et al. [5] have reported that concurrently rotating auditory cues that match visual landmarks (e.g., a fountain sound) facilitated visually induced circular vection and presence. Seno et al. [6] have reported that consistent air flow to subjects' faces facilitated forward vection. The sense of smell is not as accurate as visual and auditory senses, but it can be a cue to estimate the distance to the object. Although it might be possible to promote vection perception, the effects of olfactory stimuli on visual vection have not been investigated at all.

In recent years, sense presentation technology in VR has been developed and using five senses in VR is much easier. Because of these developments in VR technology, senses other than visual and auditory senses can be and should be more used and investigated in science too. In fact, it was suggested in various articles that the importance of multimodal research is also increasing [7]. Many trials have been conducted in which scents are displayed in conjunction with movies or games using VR systems. For example, VAQSO, a device that adds a scent to the VR experience, is planned to be released for 2019 [8]. Because this device is compatible with any HMD, the contents using scent will increase. In an immersive environment, such as a VR game using an HMD, vection is frequently induced. However, few studies have investigated the relationship between vection and olfactory stimulation. Vection is closely related to VR sickness. Investigating the relationship between vection and olfactory stimulation, can enhance our understanding of vection and may become a clue to solving the VR sickness problem.

The aim of this research was to investigate the effects of scents on vection perception. We investigated whether olfactory stimuli affect vection perception and whether the effects are emphasizing or destructive. Furthermore, we researched whether the effect on vection perception changes depending on the type of scent. In this experiment, the subjects were seated in front of an olfactory display while wearing an HMD and were presented with two types of moving images, i.e., expansive and contractive optical flow, respectively under the conditions of with and without scents. During the moving image presentation, we measured the latency and duration of vection. After finishing the stimuli presentation, the subjects evaluated the strength of the vection that they experienced using subjective values. We examined the relationship between vection and olfactory stimuli using the results of this experiment.

2 METHODOLOGY

2.1 Olfactory Display

We used the Fragrance Jet 2 olfactory display (Fig. 1). This display uses the techniques of an ink-jet printer in order to produce a jet, which is broken into droplets by a small hole in the ink tank. Bubbles are formed in the ink by instantaneous heating, and ink is ejected by the pressure of bubbles. The display can set up one ejection head. This head can store three small tanks and one large tank; thus, this display can contain a maximum of four kinds of scents. There are 127 minute holes in the head that is connected to the small tank and 256 minute holes in the head that is connected to the large tank. Because the display can emit scent from multiple holes at the same time, the ejection quantity can be set at 0 to 127 (in the small tank) or 0 to 256 (in the large tank). We denote the average ejection quantity at one time from each hole as "the unit average ejection quantity (UAEQ)", and the number of minute holes that are emitting at one time as "the number of simultaneous ejections (NSE)". The unit average ejection quantity from the one minute hole in the small tank is 4.7 picoliters (pl) and that from the one minute hole in the large tank is 7.3 pl. The reproducibility of these values was confirmed without depending on the residual quantity of ink on examination. Because the emission occurs 150 times per unit time (100 milliseconds), the ejection quantity (EQ) in the case of the small tank that we used in this experiment was calculated as follows.

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

The ejection amount per unit time depends only on the NSE. In this experiment, we refer to NSE as the ejection level.

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

2.2 Scent Stimuli

Two kinds of scents were used in this experiment: lavender (oil of lavender) and banana (iso-amyl acetate). We used lavender because it is said to affect the movement of the body [9], and is thought to have some influence on vection. Banana scent has not been reported to have any special effects the movement of the body. Therefore we used it for comparison. The scents were diluted to 5% with ethanol and water, and the component ratios of each perfume are shown in Table 1. In order to determine the ejection level at which the subjects could sense the scent, we measured the detection threshold, which is the minimum detectable concentration of scent. This method is based on the two-point comparison method that was proposed by the Japan Association of Odor Environment. The initial ejection level was set to 5. If the subject correctly identified the scent twice at the initial level, we reduced the level by 2 according to the descent method and ended the measurement when the subject could not identify the scent. When the subject was not able to correctly identify the scent at the initial level, we adopted the rising method and increased the level by 3. The measurement was considered to be complete when the subject (between the age of 20 and 30 years, male) correctly identified the scent twice. Based on this result, the ejection level to be used in the following experiment was determined. Table 2 shows the measurement values of 6 participants.

Figure 1: Olfactory Display: Fragrance Jet 2

Table 1: Scent stimuli

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

Table 2: Result of olfactory detection threshold measurement

	Lavender	Banana
Avg.	2.67	3.67
SD.	1.97	2.07
Max.	5	7
Min.	1	1

2.3 Vection Stimuli

To induce vection, two kinds of moving images were used. The expansion stimulus was used to induce a feeling of forward movement which is frequently experienced in everyday life and VR games. The contraction stimulus was used to induce a feeling of backward movement. In the case of the expansion stimulus, when a dot reached the edge of the screen and disappeared, a continuous stimulus presentation could be obtained by rearranging the dot from the center of the plane. In the case of the contraction stimulus, a dot was repositioned from the edge of the plane when it reached the center of the plane and disappeared. In each stimulus, a total of 2400 dots are displayed on the screen. The size of the dot on the screen was changed to be physically constant according to the distance change simulation. Because the dots do not have a density gradient and do not give a static depth cue, the depth cues were only provided by motion. The speed of each dot was about 3.7 degrees per second at the viewing angle. Figure 2 shows a still image of the vection-inducing visual stimulus. Arrows indicate movement of points. The visual stimuli were presented using Oculus Rift DK2. The viewing angle was set at 110 degrees in the diagonal direction and 90 degrees in the horizontal direction and the motion stimuli that induced vection was displayed almost over the entire screen.

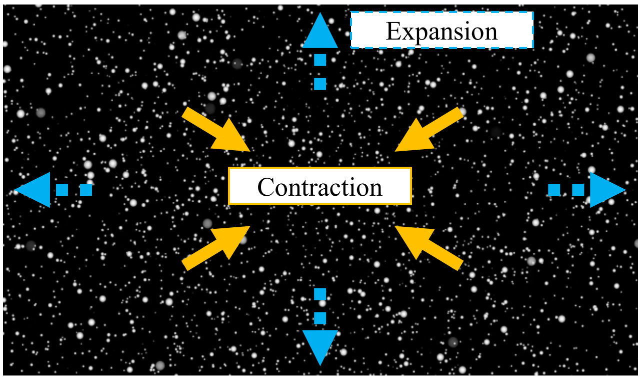


Figure 2: Still image of vection stimulus

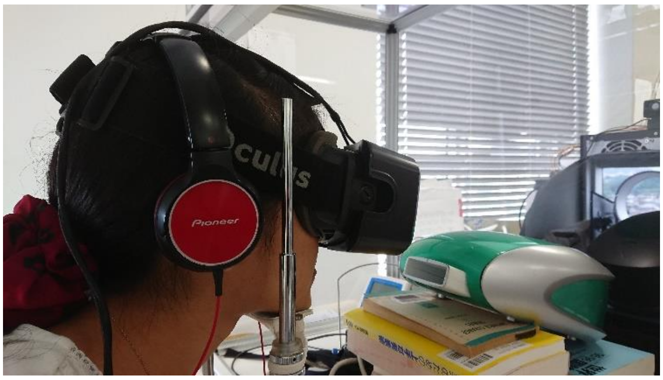


Figure 3: The experimental condition

2.4 Sound Stimuli

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

2.5 Experimental Conditions

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

2.6 Qualification of Subjects

Prior to performing the olfactory experiment, we conducted a test to confirm that the subjects had a general olfactory sense. In this test, we used odor panel analysis to qualify five standard odor solutions and tested whether each odor could be identified. The standard odor for panel selection was prepared based on the T & T Olfactory-meter, and it is used in national examinations to determine olfactory measurement operator [10]. Using odor solutions at the concentration determined by the Ministry of the Environment, the test was conducted according to the 5-2 method. Only subjects who passed the test were taken as subjects.

It is suggested that approximately 1 person in 20 people does not feel vection (vection blind). Therefore, subjects were screened before the experiment was conducted. We asked subjects to observe expansive and contractive moving images. Subjects were then asked to evaluate the vection they felt. In cases where subjects answered that they did not feel vection (intensity 0) in more than half of these tests, they were excluded from the experiment.

3 EXPERIMENT 1: PRESENTATION METHOD OF OLFACTORY STIMULI

It is necessary to examine the ejection method so that the subjects can continue to feel the scent without adaptation for 70 seconds including 30 seconds before the presentation of the vection stimuli. Therefore, a preliminary experiment was conducted with reference to a previous study [11].

3.1 Experimental Method

Two kinds of scents were used in this experiment: lavender and banana. Four times the average value obtained in the above detection threshold measurement experiment was set as a reference value 1, and 8 times the average value was set as a reference value 2.

Table 3 shows the respective ejection levels. As shown in Fig. 4, scent ejection with duration of 0.3 seconds and an interval of 1.3 seconds was repeated for 70 seconds.

After the stimuli presentation, subjects evaluated how they felt the continuity of scent ejection at the four stages in Table 4 and how they felt the intensity change for 70 seconds at five stages in Table 5. We determined a presentation condition that was suitable for this experiment; the selection of the condition was based on obtaining an average value of 2 or more, regarding the continuity, and at a level where no subjects selected 0 regarding the strength.

Measurements were made with a 1-minute break every 70 seconds and the presentation was repeated 4 times for each of the 2 scents. Taking the order effect into consideration, we determined the first and second scents randomly and adjusted them, so each order was presented approximately an equal number of times. Between the presentations of each scent, there was a five-minute break. Because of the nature of the experiment, it was assumed that the subjects could easily feel the scents. Therefore, prior to the first measurement, we confirmed verbally whether each subject could feel the ejection of scents twice. Next, we conducted the experiment with 5 subjects (between 20 and 25 years of age, male) using the reference value 1, and with 5 subjects

Table 3: Ejection levels

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

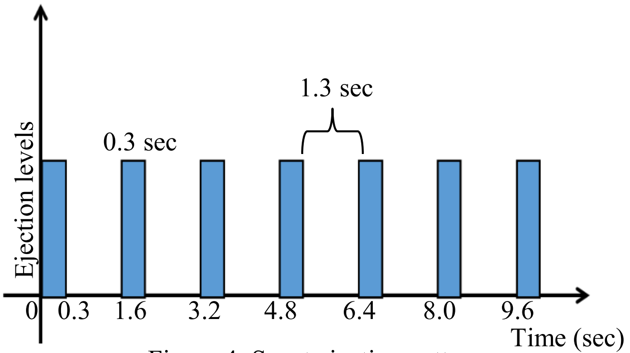


Figure 4: Scent ejection pattern

Table 4: Four-stage evaluation for continuity

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

Table 5: Five-stage evaluation for the strength

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

(between 20 and 30 years of age, male) using the reference value 2.

3.2 Experimental Results

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

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

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

Table 7: Six-level odor intensity indication

0	Odorless
1	A faint smell that you can hardly perceive (Equivalent to detection threshold)
2	Weak smell you can recognize (Equivalent to cognitive threshold)
3	Easily perceptible smell
4	Strong smell
5	Intense smell

Table 8: Scent intensity

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

When we interviewed the subjects who rated the continuity as 0, they made the following comments: "The ejection amount with the reference value 1 was too small to feel it", and "The ejection amount with reference value 2 was so large that my nose paralyzed". Thus, we asked three subjects (between 20 and 25 years of age, male) to rate the scent intensity with both reference values, using the six-level odor intensity indication method [12] (Table 7) consisting of odorless (0) to intense odor (5). This technique was devised in the field of odor control in Japan.

Table 8 shows the average and standard deviation of the evaluation values. Because the values are between 2 and 3, respectively, except for the banana scent reference value 1, we set the reference value 2 as the ejection level in this experiment so that the subject surely feels it in the experiment.

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

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

4.1 Experimental Method

The vection stimuli presentation time was set to 40 seconds. The scent and sound stimuli started 30 seconds

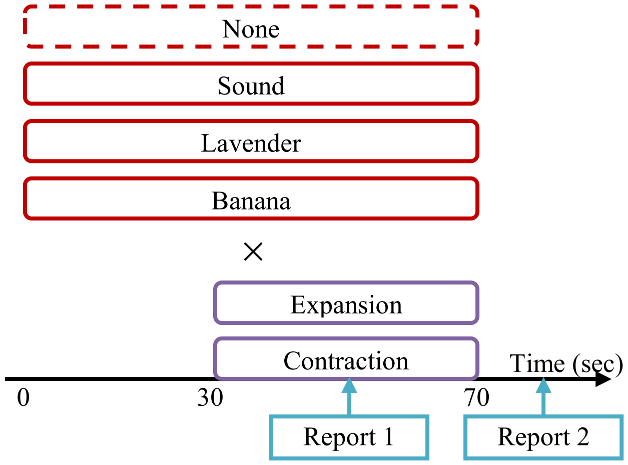


Figure 5: Flow of presenting stimuli

before the presentation of the vection stimuli and continued to be presented for 70 seconds. The flow of one trial is shown in Fig. 5. During the period in which the moving stimuli were presented, the subjects were asked to report the duration of time for which they were experiencing vection (Report1). This was reported by pressing the left button of the mouse. At this time, the time required for the first button press was recorded as the latency of vection. The total time during which the button was pressed in the 40 second period was recorded as the duration of vection. After the stimuli presentation was completed, the subjects were asked to report the strength of the vection by rating 0 when they did not feel vection, and 100 when they felt vection very strongly (Report 2). Acquisition of these three variables has been repeatedly used in previous vection experiments (e.g. Seno et al., 2013 [13]). Furthermore, in cases where an olfactory stimulus was presented, subjects were also asked to report the subjective intensity value of the scent by selecting from the six-level odor intensity indicator that is displayed in Table 8 after the stimuli presentation was completed (Report2). Under each condition, the flow of one trial, as shown in Fig. 5, was repeated 4 times, with a 1-minute break between each trial. The experiment consisted of 8 conditions, and each condition was carried out as one block with 4 consecutive trials.

The order of 8 block experiments was randomized for each subject. A 5-minute break was set between the blocks.

4.2 Experimental Results

The results under expansion stimuli conditions are shown in Fig. 6 and that under contraction stimuli conditions are shown in Fig. 7. The subjective vection strength in each condition was compared. It is known that when vection is strong, the latency is short, the vection duration is long, and the magnitude is large.

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

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

We next examined for a change of scent intensity. The result of Experiment 1 is used as a condition without HMD. In addition, the same operation as in Experiment 1 was conducted with three subjects (early 20s, male) presenting a gray scale image of Fig. 8 wearing HMD, and using the results of this as a condition without vection. The results of strength of scent under each visual stimuli condition are shown in Fig. 9. For each scent, a nonparametric test between visual stimuli was performed using the Kruskal-Wallis method. As a result, a significant difference was found in the main effect of visual stimulation in lavender condition (P <0.001), but a significant difference was not found in banana condition. Therefore, when multiple comparisons were performed using the Bonferroni method in lavender condition, significant differences were observed between without HMD and the expansion condition (P<0.05), between without HMD and the contraction condition (P<0.01), between without vection and contraction condition (P<0.05), respectively. Consequently, it was found that vection stimuli with HMD reduced the olfactory strength in lavender scent.

In two presenting scent conditions, we performed a correlation analysis by Pearson’s product moment correlation using SPSS by pooling the expansion and contraction conditions. Correlation analysis using SPSS can perform an uncorrelated test at the same time. It examines whether there is a relationship between the two variables. The strength of relevance is judged by the correlation coefficient. The results are shown in Table 9. In the table, R value represents the correlation coefficient of Pearson and P value represents

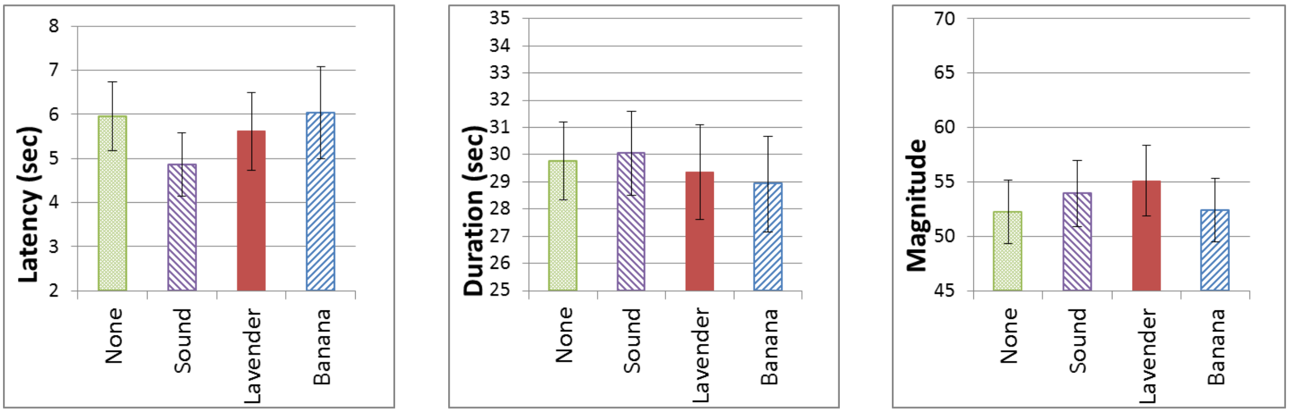


Figure 6: Expansion stimuli results

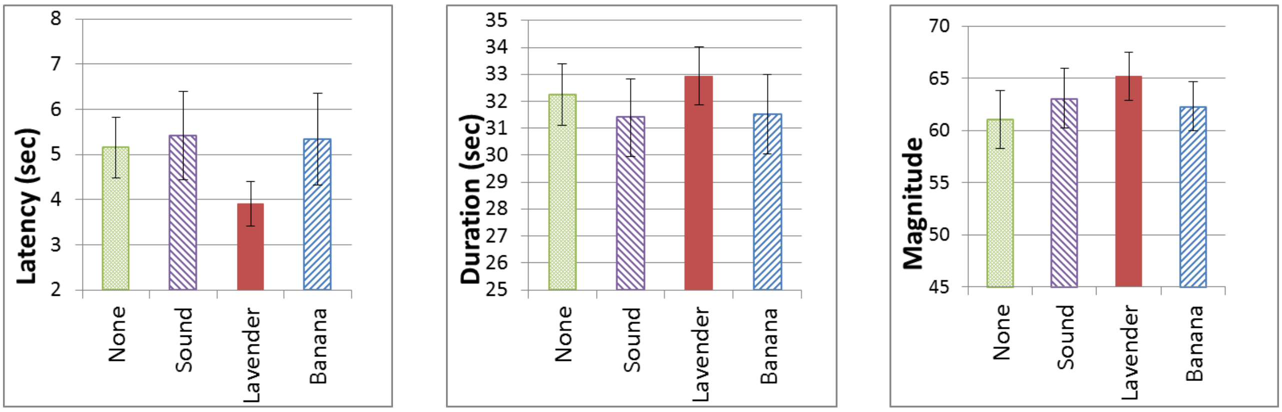


Figure 7: Contraction stimuli results

Figure 8: Gray scale image

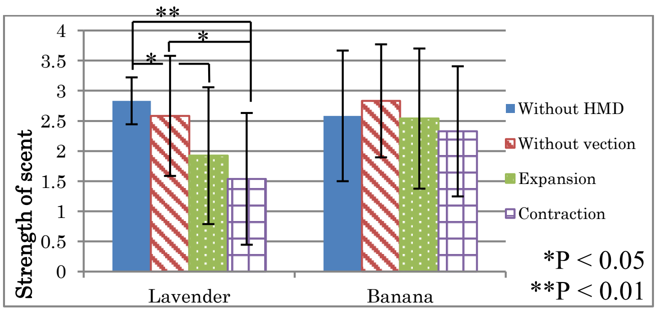


Figure 9: Strength of scent under each stimuli condition

Table 9: Correlation between vection perception and scent perception

	Scent	Value	Latency	Duration	Magnitude
Expansion & Contraction	Lavender	R	-0.326	0.346	0.308
		P	**0.001****	**0.000*****	**0.001****
	Banana	R	-0.256	0.212	0.313
		P	**0.009****	**0.031***	**0.001****
Expansion	Lavender	R	0.484	0.422	0.442
		P	**0.000*****	**0.002****	**0.001****
	Banana	R	-0.237	0.195	0.210
		P	0.091	0.167	0.137
Contraction	Lavender	R	-0.238	0.352	0.265
		P	0.089	**0.010****	0.057
	Banana	R	-0.287	0.272	0.540
		P	**0.039***	0.051	**0.000*****

the significance probability. In Table 9, *** is added to those that are significant at the 0.1% level, ** is added to those that are significant at the 1% level, and * is added to those that are significant at the 5% level. Furthermore, they are shown in bold. Significant differences are seen in all items by pooling the expansion and contraction conditions. Since the absolute values of the R values are about 0.2 to 0.3, it can be judged that the strength of the correlation is low. Then, we found that three vection indices were correlated with the perceived strength of lavender and banana scents. We also performed correlation analysis for expansion condition and contraction conditions individually and respectively. These results are also shown in Table 9. In the expansion condition, a correlation was found mainly in lavender condition but not seen in banana condition from significance probability. The R values in the item where a significant difference was observed are about 0.4, so it can be judged that the strength of the correlation is low. On the contrary, in the contraction condition, a correlation was mainly observed in banana condition rather than in lavender condition. When looking at the absolute values of R values in the item where a significant difference was observed, they are about 0.3 for the duration under lavender condition and about 0.2 for the latency under banana condition, so it can be judged that the strength of the correlation is low. In the magnitude under the banana condition, the highest correlation is found to be 0.5 R value. Thus, we could say that there was an interaction between the direction of vection and the types of scent.

5 CONSIDERATION

We examined whether olfactory stimuli affect the perception of vection by presenting subjects with visual stimuli while also presenting scents, using an olfactory display. No significant effect of scent presentation was observed in three variables, representing vection intensity. However, it was suggested that there were some effects of lavender scent on vection perception. The visual stimulus has a strong influence on the sense of smell in cases where the color of the liquid affects the scent identification [14] and in cases where the photos suitable for scent improve the intensity evaluation [15]. Furthermore, as mentioned earlier, the influence of visual stimuli on the sense of self-movement is large. The reason why there was no effect of scent on vection perception is possibly because olfactory stimuli were not sensed as consciously as vision and auditory stimuli. Olfactory information is not processed as much as visual and auditory information is, even if it does become consciously sensed. We believe that the visual stimuli were too strong for the scents to be consciously sensed. On the other hand, it was also found that vection stimuli reduced the intensity of lavender scent.

Moreover, we could find many positive correlations between perceived vection strength and perceived scent strength, especially in lavender scent. This is consistent with the result that discrimination function will be higher than normal when the HMD is worn and when vection stimuli are presented, as indicated by the research by Toju and Bannai [16]. However, this result is inconsistent with the result that the scent intensity is weakened under the condition that the vection stimuli presented. One of the reasons for this discrepancy is that the direction of attention is different in the case of evaluation only of scent (without HMD) and in evaluation of scent while evaluating a vection (HMD & vection available). In this experiment, it was not clear whether the scent is felt strongly when vection intensity is strong, or vection is felt strongly when scent intensity is strong. However, since there was little difference between the vection strength in the presence or absence of scent, we can assume that strong perception of scent will not always result in strong vection perception. If this assumption is correct, it suggests a causal relationship between strong perception of vection and perception of a strong scent. However, this is also inconsistent with the result that the scent intensity is weakened under the condition that the vection stimuli presented. With stronger olfactory or weaker vection stimuli, a more detailed investigation is necessary for this part.

In any case, from the above, it was suggested that the effects on vection perception differ depending on the type of scent. It seems that there is some relation between lavender scent and self-motion feeling induced by the vection. Lavender essential oil has a sedative effect and is believed to have an inhibitory effect on the central nervous system and the autonomic nervous system [17]. In addition, it has been reported that the fall rate of elderly people who smelled lavender scent for one year decreased [9]. These studies indicate that the scent of lavender has some effect on the human body. It is necessary that more detailed investigation using different scents be undertaken.

Sound stimuli are not considered to affect vection perception and though no significant difference was observed, a change was observed for each vection stimulus.

6 CONCLUSION AND FUTURE WORK

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

(1) Using a scent ejection time of 0.3 seconds and an ejection interval of 1.3 seconds, it is possible to present a subject with the scent at a level of about 8 times the detection threshold for 70 seconds without adaptation.

(2) There was no significant difference between the perceived vection in the presence or absence of scent presentation in the three variables representing vection intensity. However, in the contraction stimuli, there was a tendency towards a difference.

(3) Many positive correlations were observed between perceived vection strength and perceived scent strength, especially in lavender scent. Even though it was not clear whether vection affected smell perception or smell perception modified vection strength, these two perceptions were positively correlated.

(4) From the findings that were presented in (2) and (3), it we can conclude that there is a relationship between interaction with vection and the type of scent. Lavender scent may promote vection perception.

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

REFERENCES

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

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

[3] K. Murata, T. Seno, Y. Ozawa, S. Ichihara, "Self-Motion Perception Induced by Cutaneous Sensation Caused by Constant Wind.", *Psychology*, **Vol.5**, **No.15**, pp. 1777-1782 (2014).

[4] M. Tonoike, "Aroma and five senses", FRAGRANCE JOURNAL LTD., pp.52 (2016) (in Japanese).

[5] B. E. Riecke, A. Väljamäe, J. Schulte-Pelkum, "Moving sounds enhance the visually-induced self-motion illusion (circular vection) in virtual reality", *ACM Transactions on Applied Perception*, **Vol.6**, **No.2**, pp. 7:1-7:27 (2009).

[6] T. Seno, M. Ogawa, H. Ito, S. Sunaga, "Consistent air flow to the face facilitates vection." *Perception*, **Vol.40**, **No.10**, pp. 1237-1240 (2011).

[7] Y. Yanagida, "Multi-modal Interfaces for Virtual Reality", *Japan Society for Fuzzy Theory and Intelligent Informatics*, **Vol.19**, **No.4**, (2007) (in Japanese).

[8] VAQSO, https://vaqso.com/, (2018).

[9] Y. Sakamoto, S. Ebihara, T. Ebihara, N. Tomita, K.. Toba, S. Freeman, H. Arai, M. Khzuki, "Fall prevention using olfactory stimulation with lavender odor in elderly nursing home residents: A Randomized Controlled Trial", *Journal of the American Geriatrics Society*, **Vol.60**, **Issue 6**, pp. 1005-11 (2012).

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

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

[12] S. Saito, J. Inouchi, S. Ayabe, F. Yoshii, S. Nakano, "Outline of olfactory sense: Basis of the odor evaluation", Japan Association on Odor Environment, (2014) (in Japanese).

[13] T. Seno, K. Abe, S. Kiyokawa, "Wearing heavy iron clogs can inhibit vection", *Multisensory Research*, **Vol.26**, **No.6**, pp. 569-580 (2013).

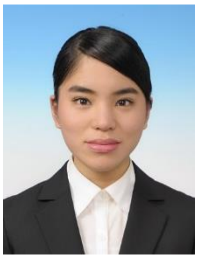
[14] N. Sakai, "Interaction between colors of foods / beverages and flavor perception", *Color Science Association of Japan*, **Vol.34**, **No.4**, pp. 343-347 (2010) (in Japanese).

[15] N. Sakai, S. Imada, S. Saito, T. Kobayakawa, Y. Deguchi, "The Effect of Visual Images on Perception of Odors", *Chemical Senses*, **Vol.30**, **Issue suppl_1**, pp. 244-245 (2005).

[16] M. Toju, Y. Bannai, "Effects of Vection on the Perception of Smell Based on Pairwise Comparisons of Two Pulses of Scents", *Virtual Reality Society of Japan Research Report*, **Vol.21**, **No.SBR-1**, pp. 7-12 (2016) (in Japanese).

[17] P. H. Koulivand, M. K. Ghadiri, A. Gorji, "Lavender and the nervous system", *Evidence-Based Complementary and Alternative Medicine*, **2013**, **681304**, pp. 1-10, (2013).

(Received October 20, 2018)
(Revised April 16, 2019)



Aoi Aruga She graduated from ICT Specialist Major, Department of Information Media, Faculty of Information Technology, Kanagawa Institute of Technology. She entered the Graduate School of Information Engineering, Kanagawa Institute of Technology. She studies about sensory information presentation and interaction between olfactory sense and other senses.

Yuichi Bannai He is a professor of Department of Information Media at Kanagawa Institute of Technology, Japan. He received a BE and a ME from Waseda Univ. in 1978 and 1980, respectively, and joined Canon Inc. in 1980. He also received MS from Michigan State University in 1988, and Ph D from Keio University in 2007. His research interests include human five senses interaction, artificial intelligence and virtual/augmented reality. He is a member of ACM, IEEE CS, IPSJ, Japanese Society for AI, and Virtual Reality Society of Japan (VRSJ).

Takeharu Seno Takeharu Seno, is an Associate Professor in the Faculty of Design at Kyushu University, in Fukuoka, Japan. His research topic has been “Vection” for more than 15 years. He became interested in vection while working on his PhD under the supervision of Professor Takao Sato at the University of Tokyo and later on as a post-doctoral fellow working with Professor Hiroyuki Ito at Kyushu University. Also, he studied and worked in the University of Wollongong under the supervision of Professor Stephen Palmisano, in Wollongong, Australia.

Regular Paper

A Web Course Based on SAT Counseling Method Reduces Anxiety by Continuous Use

Takeshi Kamita[*], Tatsuya Ito[*], Atsuko Matsumoto[**], Tsunetsugu Munakata[***], and Tomoo Inoue[****]

[*]Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan
{s1730527, s1721654}@s.tsukuba.ac.jp
[**]Graduate School of Comprehensive Human Science, University of Tsukuba, Japan
s1130368@u.tsukuba.ac.jp
[***]SDS Corporation, Japan
munakata21@yahoo.co.jp
[****]Faculty of Library, Information and Media Science, University of Tsukuba, Japan
inoue@slis.tsukuba.ac.jp

Abstract – To keep good mental health of employees is one of the prioritized issues in corporate management recently. It increases the number of potential clients, and the number of existing medical doctors and counselors is not enough. Hence the growing need for self-care. Previously, a self-care course for mental health using virtual reality (VR) based on the SAT method of counseling and therapy techniques has been developed. In this study, we propose a web-based self-guided mental healthcare course (WEB course) that can be practiced by a smartphone to improve usability for continuous use. It enables a user to repeat the course easier. The user is expected to acquire a self-care skill to cope with daily stresses so that he/she can gradually stabilize emotional arousal by oneself and improve stress tolerance eventually.

In order to examine the stress reduction effect of the WEB course, we conducted an experiment to compare between one-time use of the WEB course, continuous use during 14 days, and a breath relaxation method as a baseline. We confirmed anxiety was reduced in the continuous use of the WEB course, and discussed the merits and demerits of the course.

Keywords: Self-guided mental healthcare, Healthcare course, SAT counseling method.

1 INTRODUCTION

Research of online courses on mental healthcare has become active, as the importance of keeping good mental health has been widely recognized. The stress check system to keep employees' good mental health in companies has even been legislated recently in Japan. This resulted in sudden increase of the number of potential clients for medical counseling most of whom are not in sick, whereas the number of industrial physicians specialized in psychology who usually carry out counseling does not increase. Thus, it is in high demand to provide the means to take care of their mental health by themselves.

A self-guided mental healthcare course [1][2] based on the SAT (Structured Association Technique) method [3], which adopted virtual reality (VR) was proposed as one of such means. The course enables the user to carry out the self-guided therapy process by wearing a VR Head Mounted Display (HMD). It obtained good stress relief evaluation.

However, wearing a dedicated HMD is likely to put a burden on employees. Assuming the actual use in a company, they are required to move to a common space of the office where the HMD is installed when they carry out the course. Further, the number of users to use it simultaneously is limited. Removing these obstacles is desirable.

Further, from the viewpoint of the effect of therapy, such usability shall be also taken into consideration in terms of easiness of repeated course practice. In a SAT counserling session, a client prctices the imagery work to stimulate the intuitive association and inspiration by images, with guidance of the therapist. The client is asked of watching the images used in the session as homework when he/she feels stress until the next session. Eventually, the stress tolerance of the client is increased through such session and homework [4][5].

The self-guided course with VR (VR course) showed the possibility for a user to practice a therapy process for oneselft and get effect to reduce dailiy stresses.

In this research, we have developed a new web-based self-guided mental healthcare course (WEB course) that can be practiced by a popular device, a smartphone, using the same structure with the VR course, to improve the usability. It enables a user to repeat the course easier and acquire a self-care skill to cope with daily stresses so that the user can gradually stabilize the emotion for oneself and improve stress tolerance eventually. In order to examine the stress reduction effect of the WEB course, we conducted an evaluation experiment to compare the effect by one-time use of the WEB course and continuous use during 14 days with a breath relaxation method.

ISSN1883-4566

2 RELATED WORK

In recent years, researches have been conducted to apply psychotherapy to digital contents and use them with mobile devices as a complementary tool for treatment and counseling, or a training tool. Researches on one of major psychotherapy, Cognitive Behavioral Therapy (CBT) have especially progressed[6]-[8], and there are many commercially available mobile applications [9]. In the CBT session, the counselor modifies the negative cognitive distortion of the client through dialogue with the client to encourage positive thinking and behavioral change. After the session, clients are given a homework called the diary or column method in which they write their daily thoughts, and the counselor analyzes them in the next session for use in therapy. The CBT application is mainly digital content of this homework part and does not cover the entire CBT process. Therefore, it is hard for users to realize effect of stress reduction or stress problem solving in one-time use. While it may be an effective auxiliary tool for professional support, in case of using as a self-guided tool, users are required to be fully aware of the program and to maintain a high level of motivation to continue using it.

The Cognitive Bias Modification (CBM) approach has attracted attention as a counseling technique, mainly in Europe and the United States, and is being used extensively in research and psychology [10]. Cognitive bias refers to the assumption that people with high levels of anxiety or depression are more likely to negatively interpret vague information that can be interpreted positively or negatively. CBM-supported smartphone applications include Mood Mint, a training tool to reduce anxiety and depression [11]. In the Mood Mint, a screen displays a smile and three negative faces, all four of which are scored by immediately tapping the smile. Repeated implementation may increase the speed of response to positive images and reduce the focus on events with negative cognition. However, one-time use of Mood Mint is not intended to reduce stress or solve problems, so the user will continue to train repeatedly without knowing the stress-reducing effect. As with CBT applications, users themselves are required to maintain motivation. Mood Mint uses a system to provide point incentives for the token economy [12] as a method to encourage its use.

Mindfulness stress reduction using meditation (MBSR) and mindfulness cognitive therapy (MBCT) are also increasingly used in research and psychological clinics in Europe and the United States [13][14], and are widespread in Japan [15]. In the United States, changes in brain function were measured after 8 weeks of meditation, confirming the effectiveness of meditation [16]. MBSR refers to the "state of focus here," which is conducted in groups and individuals in combination with sedative meditation, walking meditation, and breathing techniques. Research and development on digital content of meditation has been advanced [17], and "Headspace" [18] is available in the smartphone application. In this application, courses are provided for each purpose, such as anxiety and depression, and the user performs 10 to 30 10-minute sessions per course in accordance with voice guidance. However, because some of the functions of therapy are implemented, and the main objective is to guide meditation, a single practice is not implemented with a sense of the effectiveness of stress reduction or problem solving. Thus, as with CBT and CBM applications, users themselves need to maintain high motivation. Some studies have found that meditation poses a risk of increasing discomfort and pain [19], and some aspects of the study aim require careful handling as a self-care tool.

This study aims to realize a self-care tool that assumes the use of a large number of employees with varying degrees of motivation to self-care. This program is to provide self-care measure when experiencing stress, to realize the effectiveness of stress reduction and problem-solving, and to realize a tool that can be used continuously to solve stress problems on a daily basis.

3 SELF-GUIDED WEB-BASED MENTAL HEALTHCARE COURSE

3.1 SAT Method

SAT counseling method is a structured and interview form counseling method proposed by Munakata. The SAT method has a wide effective range such as a mental disorder (such as Depression, bipolar disorder, obsessive compulsive disorder, personality disorder, schizophrenia, etc.) and various stress diseases. Unlike other conventional counseling methods focusing on the psychological aspects, the SAT method puts an importance on physicality, and approaches mental problems from the bodily symptoms. Therefore, instead of working thought by linguistic stimulus, use visual stimulus from the presented image. It is possible to grasp unconscious true feelings and an essential desire in a short time because it can functionalize an association and a flash and intuition well.

3.2 SAT Imagery Therapy

When a person who wants to have counselling recalls a stress scene, it is perceived as physical discomfort (such as Stomach shrinks, nervous, sweating hands, chest tightening). The SAT Imagery Therapy using light image is a technique to change the discomfort to a good feeling and reduce the stress by watching the light image selected and perceived as a pleasant stimulus [4].

The SAT Imagery Therapy using smile face image is a

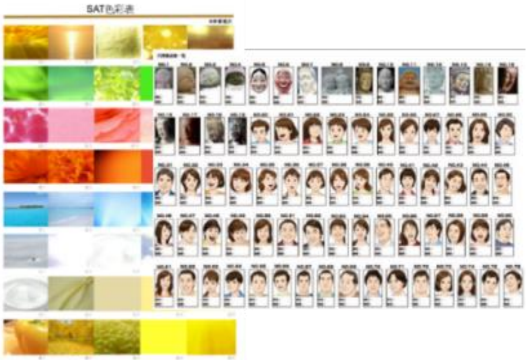


Figure 1: The list of images in printed form used in SAT

technique for transforming the image for self to a good one by replacing the primitive land-scape (for example scenery that many yell at around childhood) in the interpersonal relationship of the consultant with the image of smile face symbolizing pleasure. In psychology research, it is generally known that influence on self-esteem of a person is influenced by how the child care attitude of a child career is positive or negative. By allowing the person to select an image of smile face with a sense of good feeling and recalling the image of a scene that makes a sense of security and providing a feeling of security, person perceives a sense that is safely protected, enhances self-esteem and encourages stress reduction.

3.3 Self-Guided Course based on SAT Method

SAT counseling method does not need a client to tell his/her traumatic episodes or secrets and uses visual stimulation by images of light and positive face representation instead of nuanced linguistic expressions. Also, it is well structured which can be practiced in relatively short time in 5 to 10 minutes. But, in the conventional SAT Imagery Therapy, the expert evokes client's association through hearing, counseling or presenting thumbnails on paper media without images (Fig. 1). In some cases, the image is not sufficiently evoked by merely looking at the image on the paper medium, and the counselor has a supplementary voice call or encourages eyes to close to arouse the image while seeing the reactions such as the words and expression. Therefore, counselors play an important role in their progress. In this research, we created a course as a technique that can make self-progression even without counselor guidance support by converting SAT method to digital content and using a smartphone.

3.3.1 Course Composition

In the SAT method, first, a client is requested to answer psychological scales, and then gets to imagery therapy, and again answers the scales to check the effect of the therapy. During a session, a counselor conducts psychoeducation to deepen client's understanding of therapy if necessary (Table 1). According to this procedure, the composition of the self - guided course was designed as follows; (1) knowing their own mental condition (Assessment Part), (2) stress reduction (Solution Part), (3) knowledge and training to improve mental resistance (Learning Part). In the learning part, based on the analysis of the data obtained in the assessment part and the solution part, learning contents suitable for individual stress characteristics are provided. In this study, the assessment part and the solution part are developed prior to the learning part and the stress relief effect of image therapy in the solution part are investigated.

3.3.2 Assessment Part

In the assessment part, we conducted a mental characteristic check test (Table 2) using the SAT four psychological scale with the aim of measuring the mental condition and characteristics of the user and clarifying the changes before and after the use of the system.

Table 1: Self-guided course and SAT method content comparison

The course category	Content of self-guided course	Corresponded contents of the SAT method
Assessment Part	To know mental conditions and characteristics.	The SAT psychological scale used for health coaching by a SAT therapist
Solution Part	To carry out a therapy process to reduce stress.	The SAT imagery work using light image and smile face image
Learning Part	To learn the methodology of the SAT method and understand the course.	Psycho education conducted by a SAT therapist.

Table 2: Mental characteristic check test

Scale	Content	Total score range (SAT criterion)
State-trait anxiety inventory (STAI) (Spilberger 1970, Japanese version - Mizuguchi et al, 1982) [20]	The tendency to become anxious, not state anxiety that varies over time, but a vague degree of anxiety that reflects an individual's past experience.	20-80(20-31 lower/32-34mid/35-41higher/42-80 much higher)
Self-rating depression (SDS) (Zung 1965, Japanese version – Fukuda et al 1973) [21][22]	The depressive symptoms in mood, appetite, and sleep.	20-80(20-35 non/36-48 lower/49-68 higher, 69-80 painful)
Self-repression behavioral trait (Munakata, 1996) [23]	The behavioral characteristics that suppress one's own feelings and thoughts.	0-20(0-6 lower/7-10 average/11-14 slightly higher/15-20higher)
Difficulty in recognizing emotions (Munakata, 2001) [24]	The tendency to avoid feeling of one's own feelings, either subjectively or involuntarily. Higher scores tend to accumulate stress and become chronic with physical symptoms even if they are not aware of them.	0-20(0-5lower/6-8higher/9-20 much higher)

3.3.3 Solution Part

The solution part presents the set questions in order and is proceeded by the process that the user answers to reduce the stress. In the first half (Table 3), first, the user is asked to remember one of stress scenes in accordance with the question and make aware of how much stress it is. Next, by comparing the stress to color and form and making imagination as if the stress image compared by color and form are approaching the user oneself, the user is encouraged to perceive the physical discomfort. Furthermore, by specifying body part and type of the perceived physical discomfort, user is prompted to focus consciousness on the discomfort in the body. And then, by expressing the stress level caused by the discomfort as a numerical value (%), the user recognizes more clearly that the discomfort is occurred while feeling the stress.

In the second half, the course steps based on the process of SAT Imagery therapy using light image and smile face image (Table 4) are proceeded to decrease the stress level%, in short, relieve the discomfort and reduce the stress.

Table 3: First half of solution part

Order	Question
1	Please remind me again what you are concerned about now
2	What is it like? Please choose (Choose from 34 sources of stress such as your future, family health etc.)
3	How much is that degree? Please choose (Choose from 3 stages "not so" to "very much")
4	Does that stress comparable to color? (Choose from red, brown, black, gray, purple, navy blue, light blue)
5	If you compare the stress to the shape? (Square, rugosum, muddy, fluffy, pointed, flat, selected from spheres)
6	Close your eyes, thinking about where this thing comes and imagining this image, where do you feel strangeness in your body?
7	How is that strangeness? (Choose from throbbing, cold, heavy, dull, sore, tight, numb, stretch)
8	What is the stress level of current discomfort? (answer from 0%~100%)

Table 4: Second half of solution part

Order	Question
1	The part that feels that discomfort is healed by which color light is being protected?
2	Please choose a comfortable face that came into your eyes. Do you have anyone who smiled easily?
3	Looking at that face, what percentage of stress is the same as before? (Answer from 0%~100%)
4	What kind of character are you going to be when you see these people?
5	If such a personality, in the situation of stress, how can you handle it? It's okay with what you came up with intuition.
6	What do you think is the result if you do that?
7	Who is the most interested of those who have chosen?
8	What message will you give me?
9	How will you feel?
10	How did you feel about the stress that first came up when you were watching all the faces of these people?
11	How has the degree of stress changed? (Choose from 3 stages "not so" to "very much")

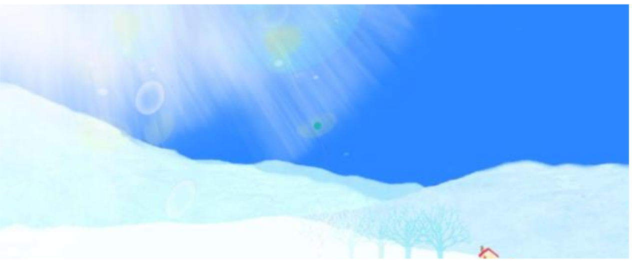


Figure 2 VR screen to view a light image

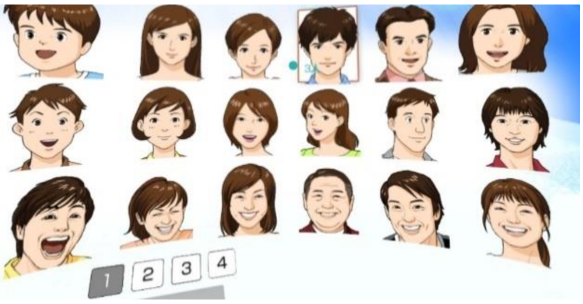


Figure 3 VR screen to select smile fece images

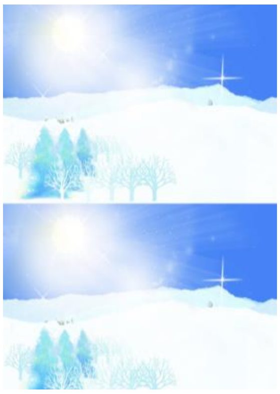


Figure 4 Smartphone screen to view a light image

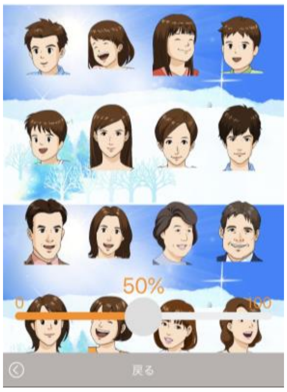


Figure 5 Smartphone screen to select smile face images

3.4 Implementation of WEB course

The WEB course using smartphones was developed in accordance with the composition of the digitized SAT method. It is constructed as a web site that can be realized with multiple platforms so as to flexibly respond to users' usage situations. Therefore, it can be accessed by using PC, smartphone, etc. In this research, we will describe contents assuming use on smartphones.

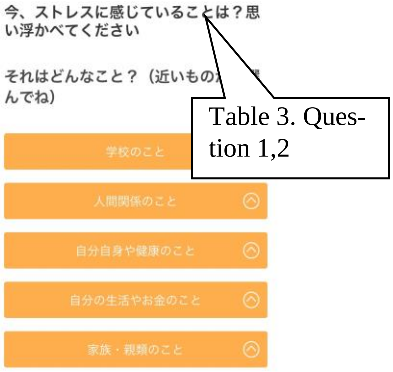
In the VR course, the immersive feeling image fits the SAT imagery therapy and can be as one of factors to bring effect on the stress reduction. However, such effect is not expected with the small flat display of the smartphone. In particular, it is assumed that the surface expressivity of the screen in the scenery of viewing light image is very different between the 360-degree screen of the VR (Fig. 2) and the small flat screen of the smartphone (Fig. 4), which will affect the stress reduction effect. On the other hand, from the viewpoint of operability, swiping and tapping operation on smartphone is easier and more familiar than moving the head to manipulate the cursor on the VR screen. In the setting of the scenery of selecting a facial image, VR requires moving the head from many facial images displayed on the front of the eye (Fig. 3), but users can easily select with the fingertips on the smartphone screen (Fig. 5). In the SAT therapy process, it is desirable to intuitively perform selection operation in a short time rather than carefully selecting using long time. By using a smartphone, relatively complicated operations such as button selection, cancellation, page advancement and return can be performed more intuitively and quickly.

When a user logs in using the ID and password on the login page, a start page is displayed. The user selects ether the button of the Assessment part or the Solution part.

The mental characteristic check test is displayed in the assessment part.

When the user selects the button of the solution part, the page shown in Fig. 6. In this screen, the user is requested to recall the stress scene (Table 3, Question 1), and select from the list of prepared stress sources (such as things of own future, family health etc.) the one closer to the problem of the stress scene (Table 3, Question 3). Then choose from 3 options for the degree of stress. In the scene where stress is compared to color and shape (Table 3, Question 4), make a selection from the image list. Returning to the chat screen, while recognizing the color and shape of the selected color, perceiving physical discomfort and specifying the part and type (Table 3, Questions 6, 7) (Fig. 7). Finally, answer by entering% of stress received by physical sense of discomfort.

Subsequently, questions are presented according to the latter half of the solution part (Table 4). The user is asked to select light images (Table 4, Question 1) from the light image list (golden, green, peach, orange, blue, white, cream, yellow color, provided based on the light image of SAT method), and then select smile faces from the smile face list (Table 4, Question 2). These selected images are displayed (Fig. 8). After that, select a representative from the selected smile face images and deepen the feeling of being protected by imagining speaking. Finally, it asked the user to answer how stress level against stress source confirmed in the first half has changed, and it ends.



今、ストレスに感じていることは？思い浮かべてください

それはどんなこと？（近いものを選んでね）

Figure 6: The Stress source list screen

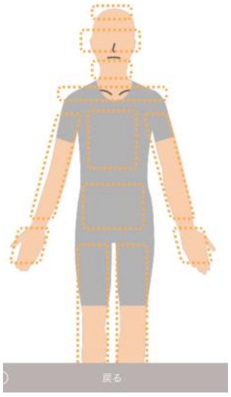


Figure 7: Identifying physical discomfort

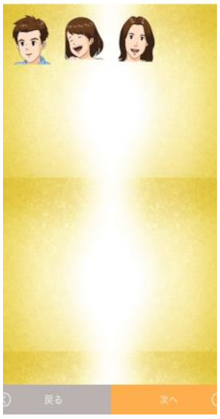
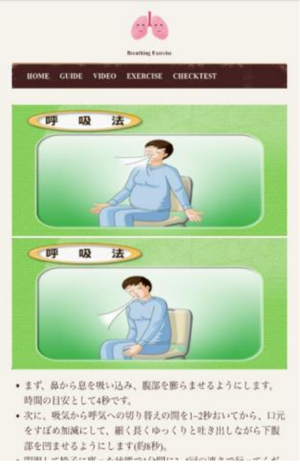
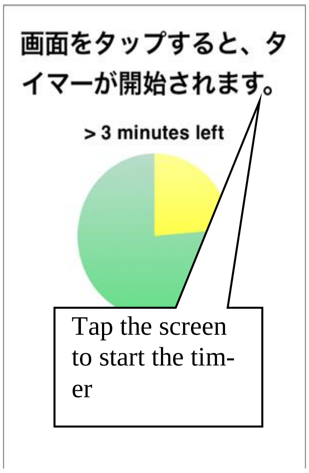


Figure 8: Selecting light image and face images



- まず、鼻から息を吸い込み、腹部を膨らませるようにします。時間の目安として4秒です。
- 次に、吸気から呼気への切り替えの間を1-2秒おいてから、口元をすぼめ加減にして、細く長くゆっくりと吐き出しながら下腹部を凹ませるようにします(約8秒)。

Figure 9: Breathing exercise practice guide screen



画面をタップすると、タイマーが開始されます。

Figure 10: 5 minutes timer screen

4 EXPERIMENT

In this research, the evaluation experiment was carried out with the approval of the ethics review committee in Faculty of Library Information and Media Science, University of Tsukuba (Notification No. 29-109)

In order to examine the stress reduction effect of one-time use and continuous use of WEB course and the difference of the effect, an evaluation experiment to have the subject continue to use this course for 14 days and to compare it with the breath relaxation method.

4.1 Breath Relaxation Method

We prepared a course to carry out the breath relaxation method as a control group.

The breath relaxation method is a training method aiming at improving the function of the mind and the body by breathing. It is introduced in the data of the Ministry of Education, Culture, Sports, Science and Technology in Japan [25] etc. as a relaxation method to consciously control breathing. It has been also reported that it is a technique that can be instructed safely and effectively in clinical practice [26]. Even a busy worker can do with a little time hanging on a chair and does not need special physical strength.

This method intends to increase the mobility of the diaphragm, and the respiratory movement that emphasizes the process of expiration is performed at a speed of 3-4 times per minute with the eyes closed and sitting on the chair [27] [28]. First, take about 4 seconds, breath in through your nose and inflate your abdomen. Next, after 1-2 seconds between switching from inhalation to exhalation, pull out the lower abdomen while drawing slowly and slowly for about 8 seconds. In this experiment, according to this method, the subjects in the control group were asked to perform this method.

We created a website (BREATH course) to guide the experiment subject's breathing practice. This has a screen (Fig. 9) for presenting the implementation method and a timer screen (Fig. 10) for displaying the execution time.

Table 5: Stress characteristic check scores

Scale	Course	Day 1-Before Average ± SD	Day 1-After Average ± SD	Day 14 Average ± SD	N	Chi-Square	df	Asymp. Sig. *2
STAI	WEB	42.95±9.23	41.85±11.03	36.88±9.52	15	6.621	2	0.037*
	BREATH	37.68±6.57	33.84±6.91	33.67±8.51	18	11.514	2	0.03*
	p-Value *1	0.080	-	-	-	-	-	-
SDS	WEB	32.48±13.04	33.05±11.26	29.50±7.67	15	2.621	2	0.270
	BREATH	31.16±7.37	30.37±4.97	27.83±6.44	18	2.303	2	0.316
	p-Value *1	0.964	-	-	-	-	-	-
Self-repression behavioral trait	WEB	10.00±2.67	9.85±3.30	8.75±2.08	15	1.750	2	0.417
	BREATH	9.79±3.33	10.21±3.88	9.22±3.74	18	3.085	2	0.214
	p-Value *1	0.743	-	-	-	-	-	-
Difficulty in recognizing emotions	WEB	9.24±2.74	9.40±3.32	8.00±2.63	15	2.327	2	0.312
	BREATH	8.63±4.19	9.26±5.30	8.94±4.28	18	0.818	2	0.664
	p-Value *1	0.548	-	-		-	-	-

*1 Man-Whitney's U test, *:p<0.05 *2 Friedman Test *: p<0.05

Table 6. Post hoc analysisa in Stress characteristic check scores changes

Course	Measured point in time	Z	Bonferroni adjusted p-Value
WEB	Day 1-Before to Day 1 After	-1.549^{b}	0.363
	Day 1-Before to Day 14	-2.592^{b}	0.030*
	Day 1-After to Day 14	-2.624^{b}	0.027*
BREATH	Day 1-Before to Day 1 After	-2.927^{b}	0.009*
	Day 1-Before to Day 14	-1.987^{b}	0.141
	Day 1-After to Day 14	-0.332^{b}	1.000

a.Wilcoxon Signed Ranks Test *: p<0.05 b. Based on positive ranks.

4.2 Procedure

33 college students and 7 office workers were selected as participants and randomly assigned to two groups, the WEB course group (N=21) and the BREATH course group (N=19). On Day1, the subjects were given the guidance on either the SAT method course or the breath relaxation method course. Before actually using the course, they took the mental characteristic check test. After using the course, they took the same test again.

After Day 2, participants were asked to take the assigned course once a day, which was also reminded by an email. They were again asked to take the same check test after Day 14.

4.3 Measurement

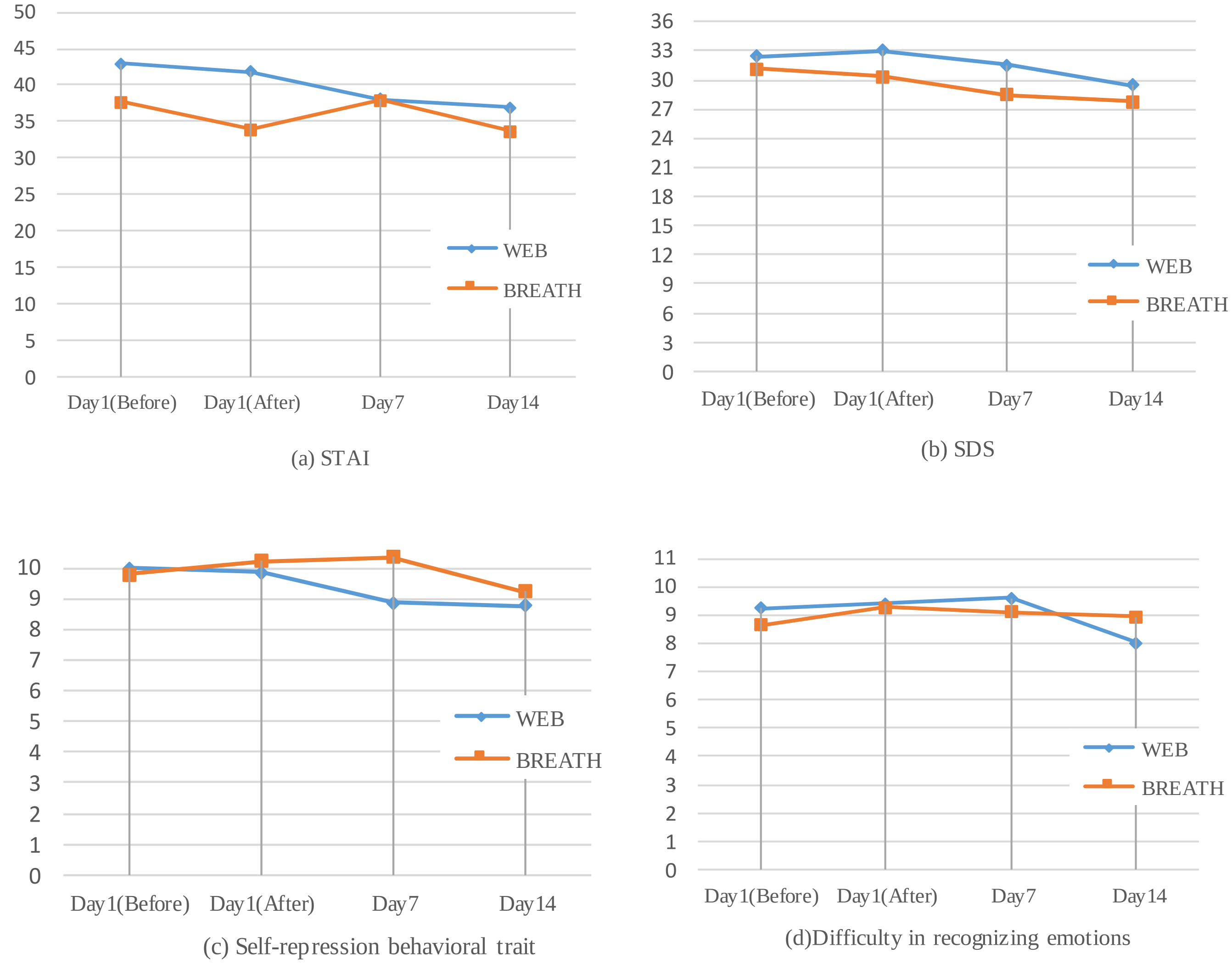
Stress was evaluated using the stress characteristic check test (Table 2). Four psychological measures (State-Trait Anxiety Inventory, Self-rating Depression Scale, Self-repression behavioral trait, Difficulty in recognizing emotions) used in the usual SAT method were used.

With regard to the obtained data, the difference of stress before using the courses between two groups was tested by Man-Whittney's U test at 5% significance level. In addition, the difference of stress before using each course was tested by Friedman Test at 5% significance level and post hoc analysis by Wilcoxon 's signed rank test with a Bonferroni adjustment applied at 5% significance level. IBM SPSS Statics Ver. 25 was used for the statistical analysis of this study.

5 RESULT

First, we performed Man-Whitney's U test to determine whether there were differences in stress status before the course between the two groups in the WEB course group and the BREATH course group (Table 5). No significant difference was found in any of the scale scores.

Then, Friedman test was performed on the four psychological scale scores of both groups to analyze whether changes



(a) STAI

(b) SDS

(c) Self-repression behavioral trait

(d)Difficulty in recognizing emotions

Figure 11: The changes in 4 psychological measures

in the scores affected the frequency of use (Table 5). It was found that there was a significant difference in the STAI score change of each group (WEB course: χ2(2) = 6.621, p = 0.037, BREATH course: χ2(2) = 11.514, p = 0.03). Subsequently, Wilcoxon's signed-rank test with Bonferroni adjustment for the difference in the STAI scores between two measured points confirmed the significant difference in the reduction of Day 14 score to Day1-Before score of the WEB course (Z=-2.592, p=0.030) and in the reduction of Day 14 score to Day1-After score of the WEB course (Z=-2.624, p=0.027). Also, a significant difference was found in the reduction of Day 1-After score to Day 1-Before score of the BREATH course (Z=-2.927, p=0.009) (Table 6). The changes are graphically presented in Fig. 11.

6 DISCUSSION

First, it was confirmed that there was no significant difference between the BREATH course group and the WEB course group in the four scale scores before the course was implemented, indicating that the groups were not with different stress characteristics.

The STAI score of both groups show a gradual decrease for 14 days (Fig. 11(a)). The STAI score of the WEB course group is within “much higher” level when the score before the course on Day 1 is compared with the SAT evaluation criteria (Table 2), suggesting that the group is much anxious and sensitive to stress. The STAI score of the BREATH course group is in “high” level and is also anxious. Comparing with the difference between before and after the course on Day 1, the STAI score of both groups decreased, but a significant difference was observed only in the BREATH course. On the other hand, after continuous use during 14 days, the STAI score in the WEB course group decreased to “high” level indicating a significant reduction in anxiety.

All SDS scores before the course on Day 1 of both groups fell within “non” level, indicating no depression. However, in both groups, the SDS score was further reduced by Day 14. The SAT method interprets anxiety as a barometer of stress, with persistent high STAI scores increasing stress, depression, and elevated SDS scores [3]. Both groups have high STAI scores but do not yet increase SDS scores, indicating that stress may not accumulate.

Comparing with STAI and SDS scores which indicate the presence or absence of stress and are relatively variable, Self-repression behavioral trait score and Difficulty in recognizing emotions score are harder to be changed than STAI and SDS as representations of the personality traits that produce stress. In the clinical setting of the SAT method, improvement of these indicators is one of the objectives of reducing stress, solving problems, and increasing stress tolerance. For Self-repression behavioral trait score before the course on Day 1, both groups are in "average" level, and the tendency to suppress their opinions and feelings is moderate and not at the level of special attention. On the other hand, both scores on Difficulty in recognizing emotions tend to be quite high and not to feel their own feelings. High levels of stress are said to be more likely to be manifested by physical illness, not mental illness. In both groups, Self-repression behavioral trait score and Difficulty in recognizing emotions score decreased at 14 days on average, but no significant difference was found, and no improvement was seen as the criteria changed.

The breath relaxation method itself has traditionally been shown to be effective in calming the autonomic nerves, relaxing the nerves, and reducing anxiety and depression [26][29]. However, the use of the BREATH course developed in this study showed a reduction in anxiety with one-time use, but no significant difference was found in the effect after continued use. Because breath function itself cannot be complemented with digital contents, the BREATH course remains in a guidance of the breath relaxation method. The first use in this study was enthusiastically tackled because the participants gathered together and started using it simultaneously. However, since the second and subsequent use, the significant effect could not be confirmed because the motivation could not be maintained by themselves. In addition, the breath relaxation method does not include cognitive changes, awareness of their own issues, and problem-solving processes, such as those used in the WEB course. Indeed, in the VR course, significant differences have been observed not only in anxiety and depression, but also in the stress personality traits, such as Self-repression behavioral trait score and Difficulty in recognizing emotions score in our latest study. Because the breath relaxation method does not include a process that promotes cognitive alteration, it is not expected to produce such changes.

Although the continuous use of WEB course demonstrated its effect in anxiety reduction, the one-time use did not. It is one of the differences from VR course, and it might come from the smaller screen size of a smart phone. However, just because of its screen size, WEB course could be more acceptable in daily life. Thus one usage scenario could be using VR course to gain the sense and feeling of its effect, followed by continuous use of WEB course. It is possible that combinational use of this type brings higher effect.

Because repetitive use is needed in WEB course, to maintain the motivation of use is one of the challenges. A good aspect to this goal is that the WEB has better usability than the VR, which makes it easier for continual use. More important factor to be motivated is that the user realizes the stress reduction in every use and expects to get better if used continually. Showing the changes in the psychological or physiological scales such as heart rate before and after the use of the course may help, even when the effect cannot be experienced by the user. For the other users who are not motivated to use the course from the beginning, it may need a mechanism to actively encourage their use. Prompting by a chat bot rather than simply waiting for their launch is an option. We will deepen our research from a multifaceted perspective regarding motivation in the future.

7 CONCLUSION

We developed a self-guided mental healthcare course based on the SAT counseling method aiming at long-term and continuously available self-guided mental healthcare tool, using a simplicity of smartphone and intuitive operation of web-based self mental healthcare course.

In this research, we conducted an experiment to have the subjects continue to use the course over 14 days and examined the stress reduction effect of one-time use and continuous use.

As a result, although stress-relieving effect in one-time use was not seen, there was a possibility of anxiety and stress reduction effect by continued use for 14 days. It is suggested that it may be effective in improving mental health by continuously using this system. On the other hand, no significant improvement in the stress characteristics to be cause stress has been confirmed, which will be a future research subject.

REFERENCES

[1] T. Kamita, et al.: "Realization of self-guided mental healthcare through the digital content based on the counseling technique SAT method", IPSJ Transactions on Digital Content, Vol.6, No.2, pp.32-41(2018).

[2] T. Ito, et al. : "A Self-guided Mental Healthcare Digital Content for Smartphone VR Based on the Counseling Technique SAT Method", IPSJ SIG Technical Report, Vol.2018-DCC-18, No.37, pp.1-8 (2018).

[3] T. Munakata : "SAT therapy", KANEKOSHOBO, Japan (2006).

[4] T. Munakata: "The applicability of the simple edition of SAT method in promoting universal health", Journal of Health Counseling, Vol.17, pp.1-12 (2011).

[5] T. Munakata : "Does SAT Re-scripting Expression Imagery Enable Us to Overcome Un-endurable Hardships toward True Life Career ?", Journal of Health Counseling, Vol.15, pp.1-12 (2009).

[6] P. J. Batterham, et al. : "FitMindKit: Randomised controlled trial of an automatically tailored online program for mood, anxiety, substance use and suicidality", Internet Interventions, Vol.12, pp.91-99, (2018).

[7] K. H. Ly, et al. : "Experiences of a guided smartphone-based behavioral activation therapy for depression: A qualitative study", Internet Interventions, Vol.2, Issue1, pp.60-68, (2015).

[8] D. Bakker, N. Rickard : "Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts

changes in mental health: MoodPrism", Journal of Affective Disorders, Vol.227, pp.432-442, (2018).

[9] J. Torous, et al.: "Cognitive Behavioral Mobile Applications: Clinical Studies, Marketplace Overview, and Research Agenda", Cognitive and Behavioral Practice, Vol.24, Issue 2, pp.215-225, May (2017).

[10] Y. hakamada, H. Tagaya: "Cognitive Biases in Anxiety and Depression：Emergence of Cognitive Bias Modification Approach", Japanese journal of biological psychiatry, Vol.22, issue 4, pp.277-295(2011).

[11] "Mood Mint", http://www.biasmodification.com/, (view 2018-11-30).

[12] F.B. Dickerson, et al. : "The token economy for schizophrenia: review of the literature and recommendations for future research", Schizophrenia Research, Vol.75, pp.405-416, (2005).

[13] Jon Kabat-Z.: "An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: Theoretical considerations and preliminary results", General Hospital Psychiatry, Vol.4, Issue 1, pp.33-47(1982).

[14] Rinske A. G., et al.: "Standardised Mindfulness-Based Interventions in Healthcare: An Overview of Systematic Reviews and Meta-Analyses of RCTs", PLOS ONE, DOI: 10.1371/journal.pone.0124344(2015).

[15] Y. Kimura: "Literature review of mindfulness training", Doshisha Women's College of Liberal Arts annual reports of studies,Vol. 67, pp.79-82(2016).

[16] R. A.Gotink, et al.: "8-week Mindfulness Based Stress Reduction induces brain changes similar to traditional long-term meditation practice – A systematic review", Brain and Cognition, Vol.108, pp.32-41, DOI: 10.1016/j.bandc.2016.07.001(2016).

[17] I.H. Bennike, et al.: "Online-based Mindfulness Training Reduces Behavioral Markers of Mind Wandering", Journal of Cognitive Enhancement, Vol.1, issue 2, pp.172-181(2017).

[18] "Headspace", https://www.headspace.com/, (view 2018-11-30).

[19] J.R. Lindahl, et al. : "The varieties of contemplative experience: A mixed-methods study of meditation-related challenges in Western Buddhists", PLOSONE, e 12(5): e0176239 (2017).

[20] C. D. Spielberger : "STAI manual", Palo Alto, Calif, Consulting Psychologist Press (1970).

[21] W. W. K. Zung : "A self-rating depression scale", Archives of general psychiatry, Vol.12(1), pp.63-70 (1965).

[22] K. Fukuda,S. Kobayashi: "SDS Manual", Sankyobo, Kyoto (1973).

[23] T. Munakata.: "Health and disease from the view point of behavioral science", Medical Friend Co. Ltd, Tokyo, pp.25-29, pp.128-129(1996).

[24] T. Munakata: "The science of mind recollection, communication, and conversation", Health Counseling, Vol. 3(6), pp.94-102(2001).

[25] "Safety measure materials for overseas educational institutions", Ministry of Education, Culture, Sports, Science and Technology Home page, http://www.mext.go.jp/a_menu/shotou/clarinet/002/003/010/004.htm, (view 2017-10-14).

[26] K. Kosakabashi : "How to incorporate the relaxation method into clinical nursing - Research, education and practice of relaxation method as nursing intervention –", KMJ The Kitakanto Medical Journal, Vol.65, Issue 1, pp.1-10 (2015).

[27] S. Arai : "Examination of effective posture for diaphragmatic respiration", Physical education science, Vol.41, pp.813-817(1991).

[28] "Useful for nursing care【3 volumes】" , https://www.igakueizou.co.jp/product/product_detail.php?product_code=NK, (view 2018-01-08).

[29] M. Okuno: "Stress reduction effect by Zen breathing method – evaluation by physiological index (spit liquid amylase and blood pressure) and psychological index (POMS and impressions)" , PhD Thesis. Kyoto University, Japan(2016).

(Received December 24, 2018)
(Revised May 8, 2019)

Takeshi Kamita is a doctoral student in the Graduate School of Libraty, Information and Media Studies, University of Tsukuba, and the representative director and co-founder of MILOQS Inc. He received his M.E. from Keio University in 1994. His current research focuses on research and development of the self-guided mental health care system.

Tatsuya Ito received his M.E. degree in Information and Media Studies from the University of Tsukuba in 2019. He is currently a system engineer at System Integrator Inc. His research interests include human interface and interaction.

Atsuko Matsumoto received her M.A. in Clinical Psychology from California School of Professional Psychology, Alliant International University in 2008 and her MHS from the University of Tsukuba in 2011. She is currently a director and co-founder of MILOQS Inc. Her current research focuses on stress management of woman mangers.

Tsunetsugu Munakata is President of SDS Corporation and Honorary professor of University of Tsukuba. He received his Ph.D. in health science from Tokyo University. He became Professor of University of Tsukuba in 1996 and Chair, Department of Human Care Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba in 2012. He served as Society and Culture Laboratory Manager of National Institute of Mental Health in 1986, Visiting Professor of Harvard Medical School in 1989 and Advisor of World Health Organization, Division of Substance Abuse in 1990, and is currently serving as President of the Academy for Health Counseling. His research interests include development and evaluation of educational method and lifestyle modification method using SAT. He was registered in MARQUIS Who's Who in the World from 2010 to 2013. He is a recipient of awards including 2000 Outstanding Intellectuals of the 21st Century Award, England 2010, Top 100 Health Professionals Award, England 2010, and Great Minds of the 21st Century Award, USA 2011.

Tomoo Inoue is Professor of the Faculty of Library, Information and Media Science of University of Tsukuba, Japan. His research interests include HCI, CSCW, and Educational Technology. He received his Ph.D. in Engineering from Keio University in 1998. He is a recipient of awards including Outstanding Paper Award, Activity Contribution Award and SIG Research Award from Information Processing Society of Japan (IPSJ). He has served a number of academic committees, including IPSJ Journal Group Editor-in-Chief, IPSJ Transactions on Digital Content Editor-in-Chief, IEICE Technical Committee on Human Communication Science Vice Chair, IPSJ SIG Groupware and Network Services Steering Board, ACM CSCW Papers Associate Chair, and IEEE TC CSCWD. He has co-authored and edited "Idea Generation Methods and Collaboration Technologies (Kyoritsu Shuppan)(in Japanese)," and "Communication and Collaboration Support Systems (IOS Press)" among others.

Industrial Paper

Proposal of Tamper-Proof IoT System Using Blockchain

Tetsuo Furuichi [*], Tomochika Ozaki [**] , and Hiroshi Mineno [***]

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

Abstract - In recent years, IoT which connect things to everything has become more widespread. Many practical systems are actually in operation, and are of great use in our lives. On the other hand, many information security incidents are reported, and the demands for countermeasures against them have been further increased. While the measures against malicious third parties have been mainstream until now, there has been an increasing demand for data falsification detection and blocking by operators and parties. Therefore, we focused on blockchains used in virtual currency as data tampering prevention technology. By applying the blockchain to the IoT system, we built an IoT system with a tamper-proof function for sensor data. Specifically, an IoT Gateway, which had been directly implemented in hardware, was realized with a smart contract, and devices which enable to use the blockchain efficiently were implemented in the configuration module. The blockchain with the tamper-proof function has a penalty of data propagation time, so we evaluated the data propagation performance in the implemented system and examined practical application examples of this system.

Keywords: blockchain, smart contract, IoT, tamper-proof

1 INTRODUCTION

IoT (Internet of Things) realized by connecting things to things and people are put to practical use, and create new value and bring great economic opportunities [1]. In this paper, we propose an IoT system with a powerful tamper-proof function by using a smart contract of a blockchain specialized for IoT. The evaluation results show that the developed prototype system is effective as an IoT system for logging system of IoT data. In this chapter, the outline of the proposed system and the situation of the logging IoT system are explained. In Chapter 2, we introduce conventional research that applies IoT, blockchains, and blockchain to IoT. In Chapter 3, we describe the issues, configurations, functions and usage examples to be solved in the IoT system using the smart contract of a blockchain. In Chapter 4, we explain the evaluation environment of the system implemented as a prototype and shows the results of the evaluation. Finally, in Chapter 5, we describe the conclusions of this survey, analysis, development, and evaluation.

1.1 Outline of the Proposed System

The IoT technology has already been used in factories and industrial products, and the merits of use for agriculture have also been reported.

When IoT devices collect more data for various purposes, those data will become more important. Then, since information security attacks targeting those data are beginning to occur, demands for information security solution are increasing. There are three major elements of information security, confidentiality, integrity, and availability. The general IT (Information Technology) system which is already put into practice has a function for protecting these software elements, but in IoT devices these countermeasures are delayed. Many damages are caused by an attack by a third party whose identity is unknown. One of the technology to prevent data tampering is blockchains that support virtual currency. The blockchain treats transaction information as a decentralized managed ledger and has a mechanism that cannot change the transaction using encryption technology in a fixed mining cycle. Therefore, anyone can view the transaction information but no one can tamper with it.

We expected that the demand to prevent IoT data tampering will increase in the future, and examined and developed an IoT system using blockchains. Since this system directly handles IoT data acquired from sensors as blockchain transaction information, it can take advantage of the blockchain tampering prevention feature as it is. And we implemented the IoT gateway using a smart contract, which is a function to process blockchain transaction information.

Latency time is one of the important performances in IoT systems. The IoT system with low latency time has a wide range of use. However, low latency systems may increase costs and power consumption. Since blockchains have mining cycles, there is a penalty in latency time. Therefore, in this research, we aimed to make it possible to use in various usages by shortening the latency time as much as possible.

1.2 Outline of the IoT System

As devices for personal use, a fitness tracker or wearable device acquires information on activity and exercise from our body and transmits the information as data. As devices for home use, there are smart home devices that control air conditioning and monitor electricity usage, and the like. There

ISSN1883-4566 © 2019 - Informatics Society and the authors. All rights reserved.

are also smart security devices that are useful for home security. In the retail environment, proper stock managements and self-checkout functions are realized with IoT. In addition, in offices, security and energy managements have been implemented to improve the efficiency of the operation of buildings and raised the productivity of employees. Advance sale analysis, usage-based design, and condition-based maintenance are effective for vehicles. By using IoT devices for industry, safety and productivity can be improved. Utilization in cities is used for resource managements, environmental monitoring, smart meters and adaptive traffic managements, and it is desired in cities to put autonomous vehicles into practical use by using smart IoT devices [2].

And, in the IoT system, the latency time is an important requirement for measuring its performance. However, in order to shorten the latency time, a lot of resources are required. Therefore, it is necessary to set the latency time according to the application.

The information obtained by the IoT devices has often been affecting the interests of operators and users. Typical examples are log information of public transportations, cars, equipment, and healthcare products, and the like.

(1) Traffic probe data

Traffic probe data obtained from taxis and buses is one specific example of log information. These data are used as a congestion degree and snow removal information each time by statistically analyzing them collectively after a lapse of time. When this information concerns someone's interests, the importance of the data increases and accuracy is required. When the data amount is large, statistically it is possible to eliminate erroneous information including noise and alteration. However, since the number of sampled probe data acquired in rustic areas is small, filtering is difficult. For that reason, prevention of tampering is an important requirement for that information.

In the case of traffic probe data, it can be used for real-time navigation and warning systems for drivers when information can be acquired or processed with low latency time. Even when the probe data cannot be processed with a low latency time, it is possible to analyze road conditions by season and time zone by batch processing.

(2) Log of vehicle and machine

Working log of a vehicle, driver's driving log, machine working log, and an operator's operation log can be used for various purposes. The authenticity of the data is also important for this information.

In the case of log information of cars, drivers and machines, if information can be acquired and processed with low latency, it will be possible to construct a warning system and accident avoidance system for drivers. In addition, if the latency time is in minutes, it can be used for machine failure prediction and doze prediction. Even with even larger latency times, if any accident happens in future automated driving, it can be used to analyze the cause of the accident and clarify the location of responsibility.

(3) Measuring equipment and inspecting apparatus

As for the information on the measuring equipment and the inspecting apparatus, not only the measurement information and the inspection information of the object but also the accompanying information such as the identification information of the measurer and the measuring time are important. For example, in order to determine the shipment of products, their measurement information may be used. In many cases, the product shipping yield depends on that measurement information. The determination of the shipment by a specific measurer or measured value may affect the final product shipment number and profit. Prevention of tampering with that information is also important. If a third party tamper with that information, it will cause confusion in product shipment. In addition, producers who stick to the number of deliveries may alter the measurement data and give priority to securing their profits.

In the case of information on measuring instruments and testing equipment, information on low latency time, especially in shipping determination, is important, because it affects throughput. In addition, when investigating the influence of yield and production number, low latency time is unnecessary, because it only refers to past measurement information as batch processing.

(4) Sensor information on healthcare

Sensor information on healthcare may change the amount of insurance subscribed to by that party. Tampering with the sensor owner may possibly change the insurance premium or the amount of insurance.

In the sensor information of healthcare, if sensor information can be handled with low latency time, it can be used for abnormality detection and notification of vital data. Also, in the case of latency time in minutes, it may be used for abnormal announcement of vital data. In the case where the latency time is long, it is possible to diagnose the health condition of the parties by batch processing.

In the current IoT devices, the operator or the user often has administrative authority. In the case where these operator and user have to take some responsibility and compensation, if they have an administrator authority for tampering data, the authenticity of acquired data may be suspected.

For these measures, human measures, technical measures and physical measures can be considered. Human measures include moral education of information security and operation management education. Technical measures include encryption with multiple keys, obfuscation, signatures, timestamps and log management at multiple sites. Physical measures include entry and exit managements and locking managements are typical. These countermeasures are time-consuming and require a lot of large-scale measures, so it is a difficulty to incur development and operation costs.

2 PREVIOUS RESEARCH AND TASK

In this chapter, we will introduce IoT technology, blockchain technology, and previous research applying blockchains to IoT.

2.1 IoT

Different types of IoT modules are made according to their purposes and conditions. The sensor type IoT module communicates the sensor data with the cloud server via the Internet using some communication method. In the actuator type IoT module, control from the cloud server reaches the actuator via the Internet or proprietary communication.

As an index of the performance of the IoT module, there are a latency time and a data transmission band. These indicators are selected depending on the purpose, with a balance between cost and performance.

Figure 1 (a) shows the simplest connection between the IoT module and the server. In this example, the controller to which the sensor is connected is able to access directly to the server via the network. The controller sends the data of the sensor to the transmission line, the server which received the data records it in the storage, and the data in the storage becomes the reference data of the Web Application.

When the communication on the sensor module side has the purpose of power consumption reduction and security consideration, a protocol conversion module called a Gateway may be interposed in the communication line. Figure 1 (b) shows the IoT system via the Gateway. The controller connected to the sensor passes the data to the Gateway once, and the Gateway sends the data to the server via the Internet.

The IoT module connected to the Internet may be hacked by a malicious third party via the Internet. The most vulnerable points are wireless and Internet communication lines. In Fig 1 (c), in order to encrypt a communication line, a system using SSL or HTTPS of MQTT(Message Queuing Telemetry Transport) is shown. In this figure, MQTT Publisher is placed on the sensor side, the gateway is set as MQTT Broker, and the server is set as MQTT Subscriber.

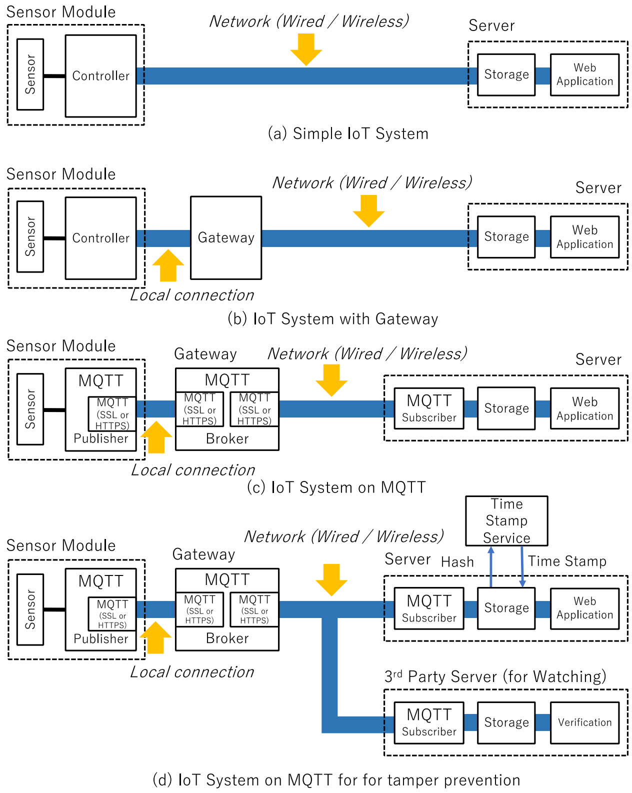
Now that many types of IoT modules have been put into practical use, there are many IoT modules that are operated as they are at shipment due to a misconfiguration of these devices. In 2016, MIRAI Botnet scans these default passwords for devices connected to the Internet, performs DDoS attacks using the compromised devices, and caused enormous damage. There are many variants of that MIRAI now.

Many IoT modules are vulnerable to various types of security threats. As a reason, there are many cases where the user is not near the IoT module, and since they are not supervised for a long time, in many cases it is not noticed that the user is under attack for a long time. In the case of wireless communication, eavesdropping is extremely easy.

In addition, components of the IoT module often have low performances in terms of their power consumption limitations and computing power. There seem to be some IoT modules on the market, commercialized without implementing sufficient security functions due to demands for low cost and a short deadline.

A powerful encryption algorithm is necessary as a technical measure against the information security threat of the IoT module. To that end, it is necessary to have a powerful processor capability that can calculate cryptographic algorithms in real-time [3].

As already mentioned, we need measures to prevent tampering by malicious third parties and parties in its operation.



(a) Simple IoT System

(b) IoT System with Gateway

(c) IoT System on MQTT

(d) IoT System on MQTT for for tamper prevention

Figure 1: Various IoT System configuration diagrams

As a method for preventing tampering, there is a method of taking an electronic signature and a method of taking an electronic signature by a timestamp server. In either case, since the administrator has all privilege, it is possible to perform operations related to tampering at any time. Therefore, there is a need for multiplexing data with a third party so that tampering history can be detected later. Figure 1 (d) shows the configuration diagram of the MQTT, the timestamp server, and the system multiplexing data with a third party. In the case of this configuration, it is assumed that the system after Subscribe of MQTT is multiplexed, and the third party's server is managed under a different administrator. When there is tampering on the original administrator side, it causes a difference from the multiplexed data, and the tampering can be detected. However, the configuration described the above makes the system complicated, and it is difficult to put into practical use in terms of labor and cost.

2.2 Blockchain

A blockchain is a system capable of tracing all transaction histories by a distributed consensus-building mechanism with network participants. So far, many kinds of virtual currency using a blockchain are distributed. Here we introduce Bitcoin which is the most famous blockchain and Ethereum which is a blockchain system that can execute application programs.

Bitcoin was developed based on the blockchain technology posted by a person named Satoshi Nakamoto in 2008. It started operation in 2009 and is a famous blockchain for virtual currency [4]. Bitcoin is a system composed of a blockchain node called Bitcoin client and a Bitcoin network. The transaction information issued by the Bitcoin client is sent as a transaction to the Bitcoin network, and minor, which is a kind of Bitcoin client, miners, so that the block is generated. And that block is approved from multiple nodes of the Bitcoin network [5].

Ethereum, proposed by Viralik Buterin's white paper in November 2013, is a blockchain and makes it possible to build applications by smart contract [6]. While Bitcoin is specialized in moving ownership of cryptographic currency, Ethereum is characterized by being able to create and execute distributed applications called smart contracts as well as moving cryptographic currencies [5][7].

A blockchain can be said a distributed database that realizes a "distributed ledger" that distributes and manages transactions as exchange information on a distributed network. We manage and operate a list of sequential data called "blocks" that summarizes those transactions on multiple nodes. Moreover, the validity of the block is secured by the mining processing using the distributed consensus algorithm. PoW (Proof of Work) is mainstream in the current distributed consensus algorithm. Under the agreement of the configuration node of the network, difficulty values are set and have a mechanism to adjust the mining time. In addition to PoW, PoS (Proof of Stake) and PoI (Proof of Importance) have been proposed as distributed consensus algorithms. Those methods have the effect of not consuming processor resources and power. In recent years, blockchains with different distributed consensus algorithms have also been released.

EOS which is one of the new blockchains uses the distributed consensus algorithm of DPoS (Delegated Proof of Stake) [8]. Generally, in PoS, block generators are determined by the amount of currency held, but in DPoS, block generators are determined by voting by other nodes in the blockchain network. Also, since the weight of the vote is determined by the amount of currency held on the blockchain, it cannot be determined by the block generator itself. However, if multiple voters agree on each other, it is possible to fix the block generator and also to perform the centralized operation. Therefore, new blockchains have been proposed that overcome these concerns [9]. These arguments are particularly active in the public blockchain, which aims at virtual currency functions. There is no need to limit the use and if the node is a member of a whitelist, it is not necessary to adhere to the expensive public blockchain, and it is more convenient to use it in a private blockchain or a consortium blockchain [10].

(1) Blockchain operation

A blockchain is composed of various elements such as nodes, P2P(Peer to Peer) networks, transactions, blocks, distribution ledgers, and mining. A blockchain is a virtual network configured on a physical ordinary network. The connection unit is called a node, and the nodes are physically connected by P2P. Each node has its own unique asymmetric key. The transaction information issued by the node is signed with the secret key of the node itself, and after another node’s approval, it is spread to the blockchain network via P2P. The spread transactions are grouped together in a preset time and are blocked by mining. The block information is handled as information of the distributed ledger after undergoing multiple approval processes.

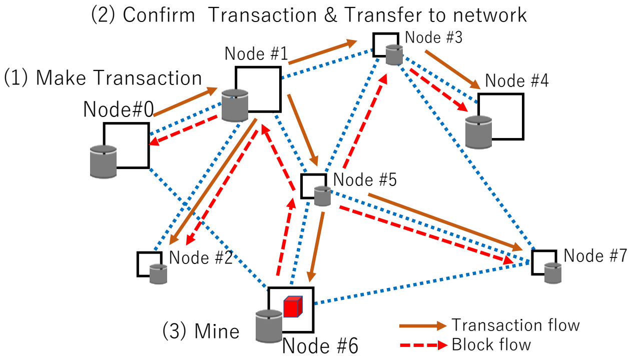


Figure 2: Process flow among blockchain nodes

Figure 2 shows an example of a flow of processing among nodes of the blockchain. First, if Node #0 issued a transaction of transaction, the information is passed to Node #1, approved, and then diffused with Node #2, #3, #5 -> #4, #6, #7. After that, Node #6 having a mining function performs Mining by using a distributed consensus algorithm, and then performs blocking. The information of the block is spread information of block information via Node # 5, approved in the entire blockchain, and is handled as a valid distributed ledger in each node.

(2) Smart contract

Whereas Bitcoin has a mechanism specialized for virtual currency trading, there is a blockchain that can handle smart contracts, which is a type of program shared on the blockchain, as well as virtual currency transactions. Ethereum is one of them, and each node can access virtual currency transaction, mainly to execute the virtual program. The creation and execution of the smart contract are treated as transactions, so their generation and execution records are stored in the blockchain and cannot be tampered with. Therefore, the reliability of the execution result of the smart contract becomes very high.

2.3 IoT + Blockchain

Focusing on the convenience of distributed management of a blockchain and the characteristics of the virtual currency, the degree of expectation for adaptation to IoT is increasing. In the field of electric power systems, there are cases where blockchains efficiently perform IoT updates on an ongoing basis. Blockchains have "a public blockchain" used in virtual currency and "a private blockchain" mainly used experimentally. This report recommends the use of a private blockchain to ensure security. Also, the number of Mining Nodes is recommended to minimize implementation considering security. Furthermore, since we disclose information by using blockchains, they recommend securing the confidentiality of data in a different way from the blockchain [10].

To cope with IoT, a mechanism is developed to cover a blockchain client program with a wrapper, and by using a network different from the blockchain, a weak data transfer of the blockchain is handled (Fig. 3) [11]. According to research to use IoT in Smart Home, the merit of information security is larger than the overhead of the processing blockchain [12].

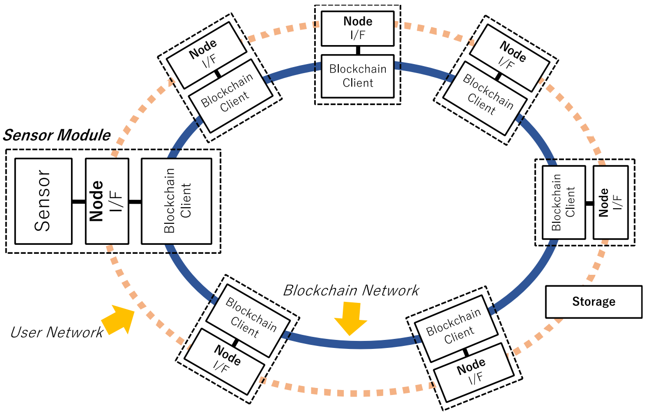


Figure 3: IoT system on blockchain with user network

In order to easily manage the configuration of the IoT module as research concretely using a smart contract, an IoT system that has the mechanism of RSA key management in Ethereum's smart contract has been reported [13].

3 IOT SYSTEM ON BLOCKCHAIN

We explained that there are new requests for information tampering prevention in the IoT system and that blockchain can be used as a means for preventing tampering. IoT systems with security requirements will have the ability to handle cryptographic algorithms at a reasonable speed. We focused on using that powerful resource for blockchain operation. However, we recognize that the penalty for latency time due to the use of the blockchain and the high cost of the huge volume of data for the blockchain. In this chapter, we will introduce the configuration of a new IoT system using a blockchain-based on the hypothesis that the bandwidth of the network including wireless will further expand, and the demand for information security will further increase.

3.1 Blockchain for Tamper Prevention

We have already explained that it is necessary for the supervisor to construct and operate a data-sharing server when realizing tamper-evident measures with the practical system configuration (Fig. 1 (d)). Implementing a new server and implementing data multiplexing in building a system takes time and labor for development and increases development cost. Furthermore, maintenance troubles and expenses including information security measures have increased, and it is clear that these configurations are not practical systems.

As techniques for tampering prevention, digital signatures and time stamps are known. These technologies can make it possible to prevent tampering by a malicious third party, but it is inevitable to prevent tampering by administrative users or users. As a method of avoiding these concerns, as already mentioned, there is a multiplex recording of data after the digital signature. That is, it requires maintenances of multiple recordings by a supervisor other than the administrator.

In the blockchain, transactions to be sent to that chain are digitally signed by the issued node and broadcasted. The node received the broadcast verifies the transaction. If the transaction is regarded as illegal, it is discarded. The approved transaction group is blocked by mining so that the data is confirmed and the record is held at each node. The advantage of the blockchain is that these series of actions have already been implemented as functions and already proven. However, even in the case of using a blockchain, it is clear that if parties such as operators and users have all nodes, it is impossible to prevent tampering from the parties. Therefore, as with the configuration in the practical system, it is necessary for the supervisor side other than the administrator to have Full Node.

As data multiplexing in practical systems has a heavy burden on system development and operation and maintenance, the use of a blockchain has a merit that a load of new system development is light. Furthermore, if major OSS (Open Source Software) blockchains are used, updates of information security, high reliabilities of the systems and lower operating costs are expected.

3.2 Block Diagram of the IoT System

Figure 4 shows a block diagram of the IoT system using a blockchain. All hardware are logically connected in the blockchain as client nodes of the blockchain. Also, the IoT Gateway composed of the smart contract has already been registered on the blockchain and can communicate with all client nodes. The configuration hardware are nodes composed of a Sensor Module, a Storage Module, a Network Gateway Module, and blockchain clients including those modules. In order to function as a blockchain, one of the client nodes has a function as a minor. in order to realize the tamper-proof function, each module communicates data as a transaction in which recording remains as a blockchain. Furthermore, in order to detect falsification of administrators and users, supervisors other than administrators and users should have one or more equivalent blockchain clients.

Generally, since the exchange information of the blockchain is handled as broadcast, the size of data that can be sent to the network is limited. This time, according to the specifications of Ethereum to be implemented, it is decided to pass data at about 1K-Bytes / sec. If more data is sent or received, it is possible to negotiate a data exchange method separately and deal with bypassing the hash value of the chunk of data to the blockchain. As an implementation method of the IoT function, an IoT Gateway is provided so that transmission and reception of IoT data can be controlled. This IoT Gateway has functions of approval of IoT module, buffering of IoT data, and distribution of IoT data. Also, by installing the IoT Gateway not by hardware implementation but by the smart contract, it is possible to avoid problems due to specific hardware or network malfunction, and realize the availability of the IoT system. Also, since this smart contract can be changed independently of hardware, scalability as an IoT system can also be secured. This implementation method is different from the conventional implementation method, and becomes a characteristic point.

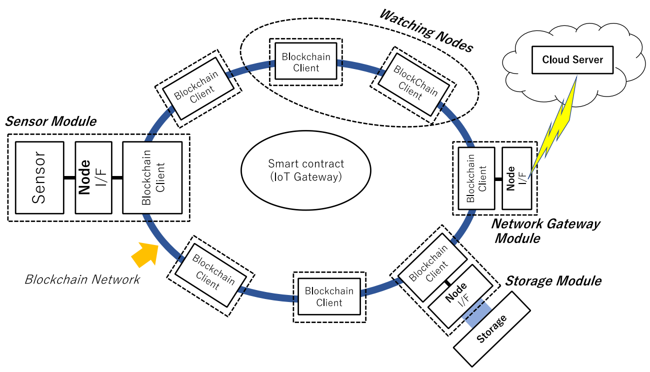


Figure 4: IoT system on a blockchain

3.3 Functions of the IoT System

This IoT system has four functions. Figure 5 shows the connection between these functions.

(1) IoT Gateway (smart contract)
The IoT Gateway has the central control function of the IoT system. In this implementation, we implemented the IoT Gateway into a smart contract of Ethereum. Individual functions include sensor module registration, activation and data buffering.

(2) Sensor Module Node
The sensor module has a function of acquiring sensor data from the sensor and sending it to the smart contract which is the IoT Gateway on the blockchain. Initially, the sensor module performs its own activation at the IoT Gateway. After approval, this sensor module will be able to send sensor data to the IoT Gateway.

(3) Network Gateway Module Node
The network Gateway module monitors events of the IoT Gateway. When an event occurs, this module receives data from the IoT Gateway and sends this data to the server on the Cloud Computer System. We will also implement a function of sending a data transmission request to the IoT Gateway according to an instruction from the cloud server.

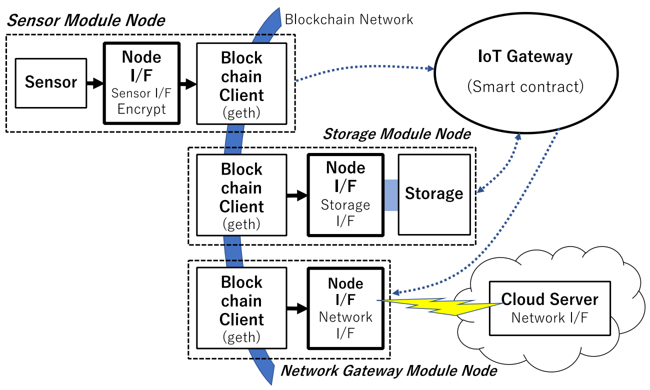


Figure 5: Various functions of IoT nodes

(4) Storage Module Node
The storage module receives the data from the IoT Gateway and stores the data in the specified channel. In addition, we will implement a function of reading data according to instructions from the IoT Gateway.

Since the function of IoT Gateway operates with the smart contract of the blockchain, when the smart contract function is called, it is executed at the timing of mining. Also, the node that has started needs to pay the execution cost of the virtual currency to the smart contract, and its function is effective for blocking a malicious third party who attacks unscrupulously.

3.4 Implementation Examples

In this section, we will examine how this IoT + blockchain system in some of the examples of systems that need tampering prevention introduced in Section 1.3.

(1) Traffic probe data
In this case, the place where the data is involved is the car that collects the information and the base to compile the data. Therefore, it is possible to create a system that prevents tampering by placing the clients of the blockchain at the automobile, the regional aggregate, the final summary and the audit site of the data. Specifically, we make the car a "Sensor Module Node" and the local area aggregate station a "Storage Module Node". Generally, the last summary office connects to the Cloud Server as a "Network Gateway Module Node". Each node exchanges data by communicating with "IoT Gateway (smart contract)" specialized for traffic probe data.

(2) Log information of car
There are two possibilities for car log information, such as the inside of a car and the system for collecting data from a car. Since there is a possibility of a mismatch between functional safety requirements and the blockchain specifications inside the car, we will examine a system to compile data from the car. It can be managed and operated with the same system as the above-described traffic alteration prevention system for traffic probe data.

(3) Information on measuring equipment
Regarding the log information of the measuring instrument, the measuring instrument is a "Sensor Module Node", and the management system of the base connecting the plurality of measuring instruments implements the "Storage Module Node". Also, the gateway system connecting the sites implements the "Network Gateway Module Node". Each node exchanges data by communicating with "IoT Gateway (smart contract)" specialized for measuring equipment.

(4) Healthcare sensor information
Sensor information on healthcare is more special in terms of cost and composition than the previous examples. In many cases, since the battery of the sensor is not large for miniaturization, chances of being connected to the network at all times may be small. A device that acquires data from the sensor is a "Sensor Module Node", and both a "Storage Module Node"

and a "Network Gateway Module Node" are installed in a smartphone or a PC. As in previous systems, each node exchanges data via the "IoT Gateway (smart contract)" specialized for healthcare sensor processing.

3.5 Latency Time Model

One of the guidelines for measuring the performance of the IoT system is latency time. If the latency time is short, it becomes possible to use the IoT system for abnormality detection and machine control. In order to shorten the latency time, it is necessary to prepare a circuit having abundant processing capability and a high-speed communication line. These conditions increase development difficulty of the IoT system, and increase development and implementation cost. Therefore, it is desirable to set the latency time condition according to the purposes and conditions of the IoT system.

The target IoT system using the blockchain treats propagation of IoT data as a transaction of a blockchain. Therefore, it is necessary to consider the latency time along the data flow of the blockchain.

Figure 6 shows the flow of the delay model of the latency time of the IoT system on the blockchain. Sensor data is obtained at the sensor module and is sent as a transaction to the blockchain network through the blockchain client (Sensing Delay). A transaction flowed to the blockchain propagate to each node as broadcast, and reach the node where mining is executed. This delay is thought to depend on the total number of nodes and the logical connection configuration of Nodes (Broadcast Delay 1). In the mining execution node, mining is executed after the interval of mining timing, and the smart contract is executed (Mining Interval, Mining Delay, Execute smart contract). Events issued by the smart contract are broadcasted in the blockchain as a transaction, reaching the receiving node (Broadcast Delay 2). Receive processing is performed at the receiving node (Receiving Delay).

In this manner, the latency time of the data propagation of the transaction in the blockchain is constituted by many routes, the route distribution algorithm due to the blockchain, and the discontinuous traffic adjustment function for blocking, it is difficult to predict with high accuracy. Therefore, an evaluation should be performed in advance in an environment similar to the system to be realized.

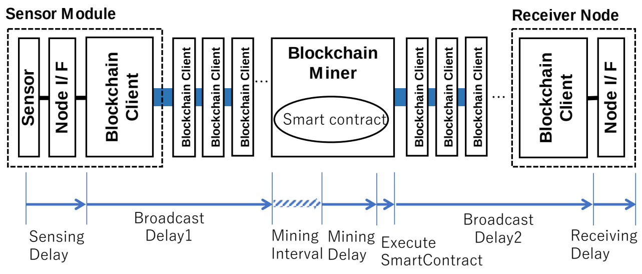


Figure 6: Latency time mode for IoT system

4 PROTOTYPE IMPLEMENTATION AND EVALUATION

In the IoT system using the blockchain designed in Chapter 3, we measured the latency time, which is one of important performance, under two conditions and evaluated the performance as an IoT system.

In the IoT system assumed this time, the function of the virtual currency of the blockchain is not important, and the main purpose is to use the information security function of the blockchain for the falsification deterrence function of IoT data. Therefore, it is not necessary to be a public blockchain with some security risk, and the prototype system and evaluation environment were implemented on a private blockchain. Also, the blockchain client program used this time is Ethereum geth v.1.8.6.

4.1 Evaluation Environment

Figure 7 shows the configuration of the environment used for this evaluation. Ethereum Private Net was constructed by implementing Ethereum 's client program (geth) on the server' s container, Note PC and IoT module. Furthermore, IoT Gateway(smart contract), which is an IoT API, is implemented in the blockchain net and it was created in advance as a transaction in the blockchain. Three types of Nodes were prepared for the evaluation. The first one is a general Node which is a contract Owner. The second one is a “Sensor Module Node” that uploads sensor data to the blockchain. The third one is an “Internet Gateway Module Node” that takes sensor data from the blockchain. Each node assigns an account address of the blockchain.

Figure 8 shows the hardware configuration of the evaluation environment. Since the evaluation at this time required a lot of nodes, we implemented the client program (geth) on the containers of server computer.

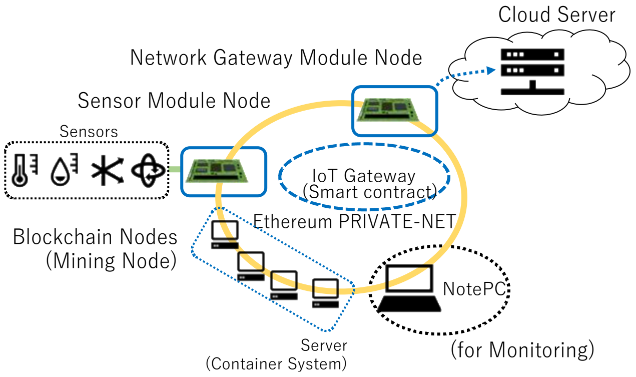


Figure 7: Overview of Logical nodes connection

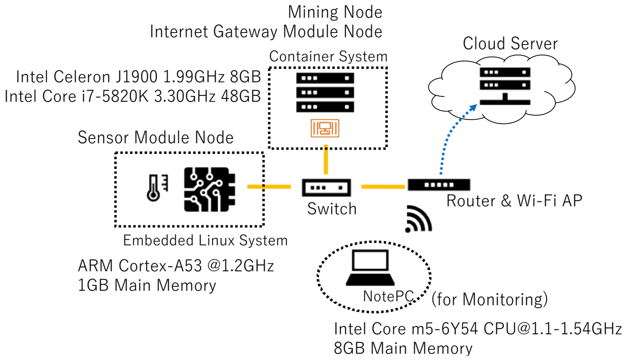


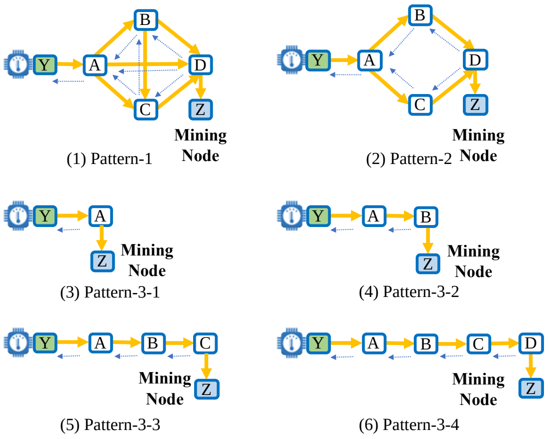
Figure 8: Physical Connection of Evaluation Environment

4.2 Evaluation Method

For evaluation, we prepared an IoT Gateway implemented on the smart contract we developed this time, Ethereum client program (geth) and JavaScript evaluation script. Each Node participates in the blockchain network by running the client program (geth). Referring to the logical function of Fig. 5, the evaluation script sends test data from the "Sensor Module Node" to the "IoT Gateway (smart contract)". The "IoT Gateway" then returns an event and the "Network Gateway Module Node" receives the event and reads the sent data. In order to accurately measure the latency time of data, this evaluation is to make the "Sensor Module Node" have the function of the "Network Gateway Module Node" and to measure the latency time of data within one node. These processes are executed for the specified number of times. Mining Node for blocking transactions in the blockchain is a Node running on a container connected to the network. Also, the execution instruction of the Mining process was manually performed.

4.3 Evaluation and Results

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



(1) Pattern-1 (2) Pattern-2 (3) Pattern-3-1 (4) Pattern-3-2 (5) Pattern-3-3 (6) Pattern-3-4

Figure 9: Various connection patterns with 1-miner

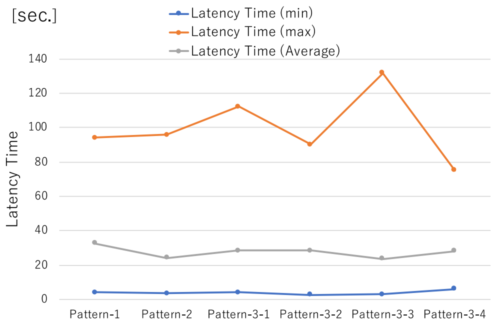
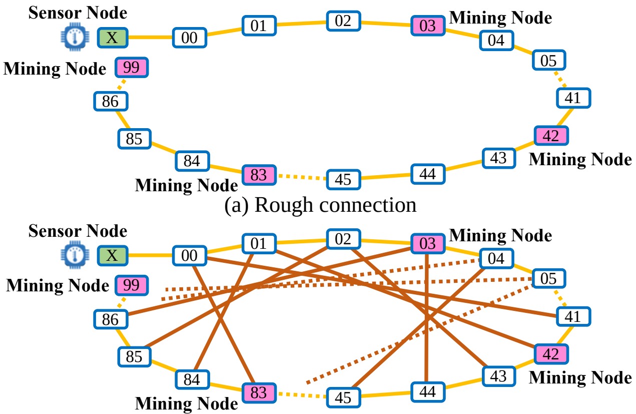


Figure 10: Latency time of each pattern with 1-miner

(1) Data transfer latency time by physical connection

This evaluation measures the data transfer latency time in "Node - smart contract - Node" by changing the physical connection form of Node in the blockchain. We made 100 data accesses in each physical connection form. In this evaluation, we used Intel Celeron J1900 1.99 GHz 8GB Main memory as a server and used a container environment. The number of blockchain nodes was evaluated at 3 to 6. Six types of physical connection were prepared. Figure 9 shows these connections. 'Y' in the green box is a Node with sensor data. 'Z' in the blue box is Mining Node. Pattern-1 indicates that the constituent Nodes A to D are mutually connected. In Pattern-2, configuration nodes A to D are connected in an annular shape. Pattern-3-1 to Pattern-3-4 are connection embodiments in which the number of Mining Nodes is changed from Node having sensor data. The yellow arrows indicate the expected direction of propagation of the issued transaction. The blue dotted arrows indicate the flow of the transaction after Mining.

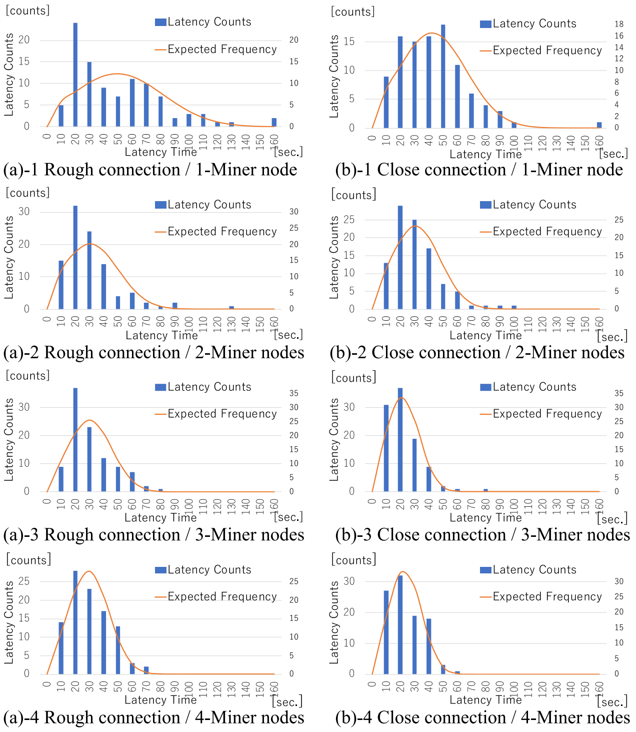
Figure 10 shows the latency time results for each connection pattern. The minimum value of the latency time was 3.86 seconds on average and the average value of the pattern was 27.59 seconds, and there was no big difference in any physical connection pattern. However, the average of the maximum values in each connection pattern is 75.09 to 131.74 seconds, which widens the value range, and the dispersion increases as the physical distance increases.



(a) Rough connection

(b) Close connection

Figure 11: 100-Nodes blockchain network



(a)-1 Rough connection / 1-Miner node

(b)-1 Close connection / 1-Miner node

(a)-2 Rough connection / 2-Miner nodes

(b)-2 Close connection / 2-Miner nodes

(a)-3 Rough connection / 3-Miner nodes

(b)-3 Close connection / 3-Miner nodes

(a)-4 Rough connection / 4-Miner nodes

(b)-4 Close connection / 4-Miner nodes

Figure 12: Latency time under various conditions

(2) Data transfer latency time by mining number

This evaluation used a container environment with Intel Core i7-5820K 3.30 GHz 48GB Main memory as a server. Since the actual usage environment was assumed, a blockchain model was prepared with the number of nodes increased to 100. Under the environment of this model, 16 sensors were assumed, and the sensor data size was 32-byte, and a high load evaluation system was prepared that could apply 128 times the sensor data load compared to the evaluation environment of the previous section. The data is sent from 16 sensors without sending delay time. The mining cycle has a cycle of about every 12 seconds because it uses the original functions of Ethereum.

In addition, two types of inter-node connection models were prepared. The first is a model (Fig. 11 (a)) in which all nodes are connected in a straight line in order to express non-uniformity of connection between nodes. The second assumes a normal blockchain connection, and prepares a model (Fig. 11 (b)) in which each node interconnects six nodes. The green box "X" in Fig. 11 (a) and 11 (b) is a sensor node and sends sensor data to the blockchain. And 16 sensors send data to one sensor node for high load. In addition, the pink box is set as one to four mining nodes assumed this time.

Figure 12 (a) shows the distribution of latency times by the number of mining nodes in a model in which 100 nodes are arranged in a straight line as a rough connection. When there is one mining node, the average latency time is 45.4 seconds, the standard deviation is 32.3, and the variation is large. Even in this model, if there are two or more mining nodes, the variations become smaller.

Figure 12 (b) shows the distribution of latency times by the number of mining nodes in a model in which 100 nodes are closely interconnected. As a result, when the number of mining nodes is one, the average latency time is 38.2 seconds, the standard deviation is 23.9, the latency time is relatively large, and the variation is also large. When the number of mining nodes was three, the average latency time was shortened to 17.1 seconds, and the standard deviation was 11.5, and stable latency time could be kept. However, with 3 and 4 mining nodes, there was no significant difference in the variation of latency time.

5 CONCLUSION

In this paper, we developed a system that uses practical blockchains to suppress falsification of IoT data, and evaluated it specifically for latency, which is one of the required performance of IoT systems. As a result, we were able to construct a system that does not require short latency time such as real-time warning and notice, which is suitable for logging application of IoT data. As a characteristic implementation method of this time, in order to use IoT data directly in the blockchain, the IoT Gateway function that controls authentication and delivery of IoT data is realized by the smart contract of the practical blockchain.

In this evaluation, we prepared an environment that applies a high load to a relatively large blockchain network. The processing performance depends on the number of connections between nodes (the number of peers) and the number of mining nodes, but it was found that IoT data could be acquired with an average latency of 17 to 46 seconds. From these results, IoT systems using this blockchain are suitable for use in IoT data logging systems without falsification of data, rather than a real-time warning or prediction system that requires short latency. These applications include traffic probe data, car log information, measuring device information, and healthcare sensor information acquisition, introduced in Section 3.4.

Moreover, in the system using the practical blockchain, the basic function and the performance have a margin, and the average latency time as IoT framework has almost no burden of data propagation due to the number of configuration nodes and the like. It was found that the impact conditions are the number of issued transactions, which is equivalent to the frequency of occurrence of sensor data, and the processing performance of mining (TPS: Transaction Per Second). Furthermore, it was found that when the density of connection between nodes is coarse or there are few mining nodes, the dispersion of latency time becomes large.

Especially in the case of the rough connection between nodes, if node connection is disconnected for some reason, it may take time until data synchronization again. Therefore, the number of peers of each node connection in the blockchain IoT system should be 3 or more.

In this experiment, we implemented the IoT system using the smart contract of Ethereum. Nowadays, many other blockchains also have smart contracts with more useful functions and capabilities. As future developments, we will study on a framework that can handle many blockchains and make it a more sophisticated IoT.

REFERENCES

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

[2] R. Patterson, “How safe is your data with the IoT and smart devices?,” Information Security, https://www.comparitech.com/blog/information-security/iot-data-safety-privacy-hackers/ (2017).

[3] A. F. Mohammed, “Security Issues in IoT,” IJSRSET Volume 3 Issue 8, http://ijsrset.com/paper/3369.pdf (2017).

[4] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” https://bitcoin.org/bitcoin.pdf, (2008).

[5] “Ethereum home page, “ https://www.ethereum.org/ (2019).

[6] V. Buterin, “A Next Generation Smart Contract & Decentralized Application Platform,” http://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf, whitepaper (2014).

[7] “Ethereum Homestead Documentation,” http://www.ethdocs.org/en/latest/ (2019).

[8] L. M. Bach, B. Mihaljevic, M. Zagar, “Comparative analysis of blockchain consensus algorithms,” Electronics and Microelectronics (MIPRO), Opatija, Croatia, pp. 1545–1550 (2018).

[9] “IOST WHITEPAPER,” https://iost.io/iost-whitepaper/ (2018).

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

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

[12] A. Dorri, S. S. Kanhere, R. Jurdak, P. Gauravaram, “Blockchain for IoT security and privacy: The case study of a smart home,” In Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE International Conference, 618–623 (2017).

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

(Received January 7, 2019)
(Revised May 6, 2019)

Tetsuo Furuichi received his B.E. degree in Electronic Engineering from Himeji Institute of Technology in 1985. He currently works for e-Cloud Computing&Co. and he is currently a Doctor-course student at Shizuoka University. He is currently interested in the embedded system, IoT, information security, and blockchain, and so on. He is currently the Registered Information Security Specialist [RISS]. He is a member of IEEE.

Tomochika Ozaki received the B.E. degree from the Nagoya University in 1988, the M.E. degree from the Nagoya University in 1990 and received the Ph.D. degree in Informatics from Shizuoka University, Japan, in 2018. In 1990, he joined Hitachi Ltd. His research interests include embedded systems, energy management systems and human machine interface. He is a member of Information Processing Society of Japan.

Hiroshi Mineno received his B.E. and M.E. degrees from Shizuoka University, Japan in 1997 and 1999, respectively. In 2006, he received his Ph.D. degree in information science and electrical engineering from Kyushu University, Japan. Between 1999 and 2002, he was a researcher in the NTT Service Integration Laboratories. In 2002, he joined the Department of Computer Science of Shizuoka University as an Assistant Professor. He is currently a Professor. His research interests include Intelligent IoT system as well as heterogeneous network convergence. He is also a member of ACM, IEICE, IPSJ, and the Informatics Society.

Regular Paper

Motif Density for Selecting Optimal Window Length in Motif Discovery

Makoto Imamura*, Mao Inoue*, Masahiro Terada*, and Daniel Nikovski**

** School of Information and Telecommunication Engineering, Tokai University, Japan
** Mitsubishi Electric Research Laboratories, USA
imamura@tsc.u-tokai.ac.jp

Abstract- Motif discovery is not only a fundamental method for finding repetitive subsequences in a longer time series, but is also used as a sub-routine in higher-level analytics including classification, clustering, visualization, and rule discovery. However, existing motif discovery algorithms depend critically on the knowledge of the correct subsequence length. Therefore, deciding an appropriate window length for subsequences is required before using those algorithms. In this work, we investigate how to decide an appropriate window length. We propose a novel index called a 'motif index' that counts the number of similar subsequence occurrences within the neighborhood in the space of subsequence, while avoiding trivial matches. We also propose a heuristic method to select an appropriate error distance for the neighborhood required as a parameter to define motif density. Furthermore, we show that motif density can decide an optimal window in the simulation data in which motifs are intentionally embedded.

Keywords: Time series data mining, Motif discovery, Window length selection, Motif density

1 INTRODUCTION

Time series motifs [1][2] are approximately repeating subsequences embedded in a time series. Motifs are one of the most important primitives in time-series data mining, and motif discovery has been used as a sub-routine in higher-level analytics, including classification, clustering, visualization and rule-discovery. Moreover, motif discovery has been applied to domains as diverse as factory operation [3], medicine [4], and seismology [5]. The notion of a motif is useful for a wide range of applications, because a repeated and frequently occurring pattern implies a latent system that occasionally produces a repeatable output. For example, this system may be an over-caffeinated heart, sporadically introducing a motif pattern containing an extra beat [6], or the system may be a factory worker, producing repetitive movement in a series of assembly operations [3].

Since the Matrix Profile [7], a fast and scalable algorithm for subsequence all-pairs-similarity-search in time series, has been introduced, it has helped to develop new innovative ideas for time-series data mining [8]. However, because a motif is defined as a pair of subsequences the distance between which is the smallest, it does not necessarily imply the frequent occurrence of a motif subsequence. That is, there are not necessarily many subsequences in the neighborhood of a motif. Furthermore, motif discovery algorithms expect that a subsequent length be chosen beforehand, which usually means in practice that users must try several possible lengths, and must confirm that the discovered motif indeed has frequent similar subsequences in a time series.

In this work, we propose a novel index called a 'motif density' that counts the number of similar subsequence occurrences within the neighborhood in the space of subsequences, ignoring trivial matches. We also propose a heuristic method to select an appropriate error distance for the neighborhood, where error is a parameter that decides the similarity level in motif density. Furthermore, we show that motif density can decide an optimal window in simulation data in which motifs have been embedded intentionally.

The rest of our paper is organized as follows. Section 2 describes the definition of a motif, and the criteria to determine the appropriateness of a subsequence as a motif. Section 3 defines motif density, based on a neighborhood of a subsequence in a set of subsequences without trivial matching subsequences. Section 4 proposes an algorithm to calculate motif density. Section 5 evaluates our proposed algorithm empirically. First, we show that motif density can decide an optimal window-length for finding motifs. Second, we evaluate a heuristic method to select an appropriate error distance for the neighborhood, required as a parameter to define motif density.

2 MOTIF CRITERIA

This section describes the commonly used definition of motif and summarizes the problem of deciding optimal window-length of a motif as motif criteria.

2.1 Our Approach

A motif is defined by using the nearest-neighbor distance in the space consisting of subsequences in a time series.

Definition: time series X

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

Definition: subsequence $X[p:q]$

A *subsequence* $s = [x_p, x_{p+1}, \ldots, x_q] = X[p:q]$ is a list which consists of continuously occurring values of *X*, starting at position *p* and ending at position *q*.

The *length* *w* of a subsequence *s* is $w = q - p + 1$, and we denote it by *length*(*s*). We also denote a subsequence $X[p:q]$ by $X_w(p)$, which means a subsequence staring at *p* with length *w*.

Definition: support of a subsequence

The *support* of a subsequence S is a set of time points [*p*:*q*] = [*p*, *p +1*,..., *q-1*, *q*], and we denote it by *support* (*s*).

Definition: subsequence space $S_w(X)$

A subsequence space is the set of all the subsequences with length w in a time series *X*. We denote it by $S_w(X)$. A subsequence space is the w-dimensional Euclidean space. Therefore, for given subsequences s_i and s_j, the distance between s_i and s_j, which we denote by dist (s_i, s_j), can be defined similarly to that in a vector space. In this paper, we use L_1 distance defined below.

$$dist(X_w(p), X_w(q)) \equiv \sum_1^w |X(p+i-1) - X(q+i-1)|$$

Definition: disjoint subsequences

Let s_i and s_j be subsequences. When *support* (s_i) and *support* (s_j) are disjoint, that is, *support* (s_1) ∩ *support* (s_2) $= \emptyset$, we say that s_i and s_j are disjoint.

Definition: Motif subsequence (1-NN)

Let w be a window-length, and let *X* be a time series. A subsequence s with length w in X is said to be *motif*, if it satisfies the below condition.

There is a subsequence s' with window-length w, such that

$$\text{dist}(s, s') = \min_{i,j} \{ \text{dist}(s_i, s_j) \mid s_i, s_j \in S_w(X) \text{ and } \text{support}(s_i) \cap \text{support}(s_j) = \emptyset \}$$

The above definition is based on the one-nearest-neighbor (1-NN) distance. We can extend this definition to k- nearest-neighbor distance by replacing the minimum with the k-th minimum in the above condition.

2.2 Challenges in Defining Motif Criteria

In this subsection, we investigate criteria to determine the appropriateness of a subsequence as a motif, which we call motif criteria. The intuitive meaning of a motif is a subsequence which has many similar subsequences in a time series, therefore we will try to define an index to measure the meaning of 'similar' and 'many' according to the intuition above. A similar sequence is measured by the distance between subsequences. For defining "many", we should count the number of similar subsequences to a motif. We call this number "occurring frequency". Challenges in defining occurring frequency are summarized in the following three points.

(1) Error dependency

When we say a sequence s_i is similar to a subsequence s, it means that *dist*(s_i, s) is small. Therefore, the threshold of an error distance parameter ϵ is required for counting similar subsequences. A naïve definition of occurring frequency of s is $|\{s_i \mid s_i \in S_w(X) \; and \; dist(s_i, s) \le \epsilon \}|$, where |A| means the number of elements of a set A. This definition of occurrence frequency requires a window-length w and an error ϵ as parameters. That is, how to decide an appropriate pair of a window-length w and an error value ϵ is the first challenge.

(2) Window-length dependency

If an error value is equal in subsequences with different lengths, the longer subsequence seems to be more appropriate than the shorter one as a motif. How to normalize by a window-length is the second challenge.

(3) Trivial match

Subsequences close to a subsequence s in a time series are similar to s, if the time series is continuous and varies slowly. We call this property "trivial match". A trivial match is described formally by the property that dist(X[p': $p' + w - 1$], X[p: $p + w - 1$]) is small, if $|p - p'| \ll w$. When we count similar subsequences, we must remove trivially matching sequences. The third challenge is how to count similar subsequences, while avoiding trivially matching subsequences.

3 MOTIF DENSITY

3.1 Our Approach

This subsection describes our approach to solving each of the problems described in the precious section.

(1) Error parameter dependency

We shall define the neighborhood of a subsequence in $S_w(X)$ for a given time series X, a window-length w and a threshold on the distance ϵ.

(2) Window-length dependency

We shall define '*motif density*' which expresses occurring frequencies normalized by window-lengths for comparing the appropriateness among motifs with different window-lengths.

(3) Trivial match

When we define the neighborhood of a subsequence s in a subsequence space $S_w(X)$, we remove trivially matching subsequences of s by using the concept of disjoint subsequences defined previously. That is, we shall define a special topology for a subsequence space generated by a time series.

3.2 Neighborhood of a Subsequence

We will define the neighborhood of a subsequence in a time series to avoid a trivial match problem.

Definition: *Disjoint neighborhood of a subsequence*

Let X, w , ϵ and s are a time series, a window-length, a positive real number, and a subsequence with length w respectively. A subset of $S_w(X)$, $D_{w,\epsilon}(s)$, is called a disjoint neighborhood of a subsequence, if it satisfies the following conditions.

(i) For every $s_i \in D_{w,\epsilon}(s)$, $dist(s_i, s) \le \epsilon$

(ii)For every $s_i, s_j \in D_{w,\epsilon}(s)$, support $(s_i) \cap$ support $(s_j) = \emptyset$.

We select a maximal one for constructing the occurring frequency of a subsequence.

Definition: Maximal neighborhood of a subsequence

Let $\mathcal{D}_{\mathrm{w},\epsilon}(\mathrm{s})$ denote a set of all of the disjoint neighborhoods of a subsequence s. A disjoint neighborhood of a subsequence s is said to be a maximal neighborhood $\mathrm{B}_{\mathrm{w},\epsilon}(\mathrm{s})$, if it has the largest number of elements in $\mathcal{D}_{\mathrm{w},\epsilon}(\mathrm{s})$. $\mathrm{B}_{\mathrm{w},\epsilon}(\mathrm{s})$ can be defined formally by the following formula.

$\mathrm{B}_{\mathrm{w},\epsilon}(\mathrm{s}) = \underset{\mathrm{D}_{\mathrm{w},\epsilon}(\mathrm{s}) \in \mathcal{D}_{\mathrm{w},\epsilon}(\mathrm{s})}{\operatorname{argmax}} |\mathrm{D}_{\mathrm{w},\epsilon}(\mathrm{s})|$, where $|\mathrm{D}_{\mathrm{w},\epsilon}(\mathrm{s})|$ means the number of elements of $\mathrm{D}_{\mathrm{w},\epsilon}(\mathrm{s})$.

We shall define the occurring frequency of a subsequence s by the number of element of $\mathrm{B}_{\mathrm{w},\epsilon}(\mathrm{s})$. The following theorem gives us how to construct a $\mathrm{B}_{\mathrm{w},\epsilon}(\mathrm{s})$ for given w, ϵ, and s.

Theorem: Construction of a maximal neighborhood.

Let X, w, ϵ and s be a time series, a window-length, a positive real number, and a subsequence with length w, respectively. $B_{w,\epsilon}(s)$, which is constructed by the below procedure, is a maximal neighborhood of s.

(step1) Select the disjoint subsequences whose distances from s are smaller than ϵ from s towards right (later time) to the end of a time series in order. We call the set of those subsequences a right disjoint set.

(step2) Select disjoint subsequences whose distances from s are smaller than ϵ from s towards left (earlier time) to the beginning of a time series in order. We call the set of those a left disjoint set.

(step 3) Let $B_{w,\epsilon}(s)$ be the union of the right and left disjoint sets.

Proof:

Let $B'_{w,\epsilon}(s)$ be one of the maximal neighborhoods of s. It is enough to prove $|B_{w,\epsilon}(s)| = |B'_{w,\epsilon}(s)|$, where $|B_{w,\epsilon}(s)|$ means the number of the elements of $B'_{w,\epsilon}(s)$.

We show only the case from s toward right to the end, because the case towards left is similar.

Let the elements of $B_{w,\epsilon}(s)$ be sorted by time ordering, we obtain

$B_{w,\epsilon}(s) = \{\ldots, s = X_w(p), X_w(p_1), X_w(p_2), \ldots, X_w(p_n)\}$
where $p < p_1 < p_2 < \cdots < p_n$.

Similarly, we obtain

$B'_{w,\epsilon}(s) = \{\ldots, s = X_w(p), X_w(p_1'), X_w(p_2'), \ldots, X_w(p_n')\}$
where $p < p_1' < p_2' < \cdots < p_n'$.

By the above construction of $B_{w,\epsilon}(s)$, p_1 is the smallest, so $p_1 \le p_1'$. In the same way, we get $p_2 \le p_2'$, because "$X_w(p_2')$ is disjoint with $X_w(p_1)$" and "$X_w(p_2)$ is the leftmost disjoint subsequence with $X_w(p_1)$". By mathematical induction, we obtain $p_i \le p_i'$ for $1 \le i \le n$, where n is$|B'_{w,\epsilon}(s)|$. This shows that $|B'_{w,\epsilon}(s)| \le |B_{w,\epsilon}(s)|$.

If $|B_{w,\epsilon}(s)| < |B'_{w,\epsilon}(s)|$, it is contrary to the maximality of $|B'_{w,\epsilon}(s)|$. Therefore, $|B_{w,\epsilon}(s)| = |B'_{w,\epsilon}(s)|$, which is what we wanted to prove.

3.3 Occurring Frequency and Motif Density

First, we define the occurring frequency of a subsequence for each window-length.

Definition: Occurring frequency

Let w , ϵ and s are a window-length, a positive real number, and a subsequence with length w, respectively.
The occurring frequency of a subsequence s is the number of the elements of a maximal neighborhood of a subsequence $B_{w,\epsilon}(s)$, that is, $|B_{w,\epsilon}(s)|$.

Next, we define motif density to normalize the difference among window-length.

Definition: Motif density

Let w , ϵ and s are a window-length, a positive real number, and a subsequence with window-length w, respectively.
The motif density of a subsequence s is $w \times |B_{w,\epsilon}(s)|$.

We regard a subsequence that has the highest motif density as the best motif among all the subsequence with various window-lengths. We show a procedure to select the best motif.

1. Give a list of window-lengths $W = [w_1, \ldots, w_i, \ldots, w_n]$.
2. Select the subsequence s_i which has the largest occurring frequency for each window-length w_i in *W*. We call the subsequence s_i the optimal motif for a window-length w_i.
3. Select the motif that has the highest motif density among the optimal motifs $[s_1, \ldots, s_i, \ldots, s_n]$ for the window-lengths *W*. We call this motif *the best motif* among optimal motifs for window-lengths *W*. We also call the window-length of the best motif *the best motif length*.

In the above procedure, the best motif length depends on an error parameter ϵ that determine the similarity level in counting occurring frequency. We call ϵ an error parameter hereafter. The error parameter in motif density is essential like a parameter k is essential in k-means clustering algorithm. We propose a method to help finding the appropriate error parameter ϵ like the Elbow method [11] for finding the appropriate number k of clusters in clustering. When we plot motif density against error parameter values, we get the graph of a monotonically increasing function. We can select an appropriate error parameter value where the rate of increase suddenly drops in the graph. This method based on the institution that a good motif has a clear boundary that divides similar subsequences from dissimilar ones after trivially matching sequences are removed.

An optimal motif for a smaller window-length than the best motif length has relatively high motif density value, because a part of a motif is also a motif. Furthermore, a motif with a smaller length might have a quickly rising motif density at very small error values.

We summaries the above considerations as three hypotheses.

Hypothesis 1: Motif density can decide the best window-length for motif discovery.

Hypothesis 2: An optimal motif for a smaller window-length than the best motif length has relatively high motif density values.

Hypothesis 3: We can select an appropriate error parameter by means of an Elbow method for the graph of a motif density functions against error parameter values.

We shall evaluate the above hypotheses in Section 5.

4 ALGORITHM

We can obtain algorithms for calculating occurring frequency and motif density by operationally interpreting the definitions and the theorem in the previous section.

Table 1 shows an algorithm that counts the occurring frequency of a given subsequence. The inputs are a time series X, a window-length w of the given subsequence s, a starting time t of s, and an error per window-length ϵ. The outputs are the occurring frequency and the motif density of the given subsequence s.

Line 01 calculates the distance between the given subsequence s and each subsequences in $S_w(X)$. Line 02 counts the number of elements that are in the right-hand side of s in the maximal neighborhood of s. Line 03 counts the number of those in the left-hand side of s. Line 04 counts the total occurrence frequency of s by adding the occurrence frequency in the right side obtained by line 02 to that in the left side obtained by line 03. Line 05 calculates the motif density of s by multiplying the window-length w and the occurring frequency obtained by line 04.

Table 2 shows an algorithm that counts the number of elements of a maximal neighborhood subsequence set whose elements are to the right of the given subsequence s. The inputs are the distance list DL obtained by line 01 in Table 1 the window-length w of a given subsequence s, a starting time t of s, and the error per window-length ϵ. The output is the number of maximal neighborhood subsequences in the right side of s.

Line 01 initializes a time cursor 'Cur' and a normalized error 'Thr'. Line 02-13 is a while-loop that chooses maximal subsequences that are in the right-hand side of the given subsequence s toward the end of the time series X. Line 03-05 is a while-loop that searches the next disjoint subsequence whose distance from s is smaller than 'Thr'. Line 06-09 increments 'Right' when the line 03-05 found a new disjoint subsequence. Line 10-12 exits while-loop 02-13 after checking all the subsequences in the right-hand side of s.

Table 3 shows an algorithm that counts the number of maximal neighborhood subsequences which are in the left-hand side of the given subsequence s. The left-hand case is reduced to the right-hand case by reversing the time series values from right to left.

Line 01 reverses the distance list 'DL' from right to left. Line 02 reverses the starting time t of s from right to left. Line 03 gets the value of the left-hand case by calling the algorithm 'CountRightOccurence' with reversed arguments.

Table 1. CountOccurringFrequency Algorithm.

```
Algorithm: CountOccurringFrequency ($X$, $w$, $t$, $\epsilon$)
[Input]   $X$: Given time series
          $w$: Length of a given subsequence $s$
          $t$: Stating time of a given subsequence $s$
          $\epsilon$: Error per window-length
[Output]  OF: Occurring frequency of s for $w$ and $\epsilon$
          MD: Motif density of s
01  DL = distanceListFromS($X$, $t$, $w$);
02  OFR = countRightOccurence (DL, $t$, $w$, $\epsilon$)
03  OFL = countLeftOccurence (DL, $t$, $w$, $\epsilon$)
04  OF = OFR + OFL;
05  MD = OF * $w$;
06  return (OF, MD);
```

Table 2. CountRightOccurence.

```
Algorithm: countRightOccurence (DL, $t$, $w$, $\epsilon$)
[Input]   DL: Distance list
          $w$: Window-length of a given subsequence $s$
          $t$: Stating time of $s$
          $\epsilon$: Error per window-length
[Output]  Right: the number of maximal neighborhood
                 subsequences to the right of $s$.
01  Cur = t+1;  Thr = $\epsilon * W$;
02  while Cur <= length(DL)
03    while DL(Cur) > Thr or Cur <= length(DL)
04      Cur = Cur + 1;
05    end
06    if DL(Cur) <= Thr
07      Right := Right + 1;
08      Cur := Cur + $w$ – 1;
09    end
10    if Cur > length(X)
11      break;
12    end
13  end
14  return Right;
```

Table 3. CountLeftOccurence Algorithm.

```
Algorithm: countLeftOccurence (DL, $w$, $t$, $\epsilon$)
[Input]   DL: Distance list
          $w$: Window-length of a given subsequence $s$
          $t$: Stating time of a given subsequence $s$
          $\epsilon$: Error per window-length
[Output]  Left: the number of maximal neighborhood
                subsequences in the left of $s$.
01  DL_rev = fliplr(DL);
02  t_rev = length($X$) – $t$ + 1;
03  Left = countRightOccurence (DL_rev, t_rev, $w$, $\epsilon$)
```

5 EXPERIMENTAL EVALUATION

We evaluate the three hypotheses described in section 4.

5.1 Window Length Selection

This subsection evaluates the two hypotheses below in two simulated time series in which motif subsequences are intentionally embedded.

Hypothesis 1: The best window-length can be decided by selecting the one that has the highest motif density values.

Hypothesis 2: A maximal motif for a smaller window-length than the best motif length has relatively high motif density values.

(1) Experiment on data set 1

First, we will show that motif density can be used to decide the best motif length (15) by selecting the window-length that has the highest motif density among optimal motifs with window-lengths 5,9,15, and 31.

Figure 1 is a simulated time series that combines sine curves with length (period) 15 samples per one cycle, and random subsequences with various lengths. In Fig. 1, the horizontal axis means time points in the time series, and the vertical axis means the values of the time series. Sine curves with length 15 are intentionally embedded as motifs. We call this time series data set 1.

Data set 1 is obtained by alternatively arranging ‘a noisy sine curve whose length of one cycle is 15’ and ‘a random subsequence that has a random length between 1 and 15’ for twenty times. Each value in a random subsequence follows a random uniform distribution whose values are between -1 and 1. The noise included in a sine curve follows a random uniform distribution whose values are between -0.02 and 0.02.

Figure 2 shows the motif density trend graph for each window-length in the case that an error per window-length parameter (we call it as EPA hereafter) is 0.01. In each graph of Fig. 2, the horizontal axis means time points in the time series, and the vertical axis means the motif density of each subsequence starting at each time point. A procedure how to decide an EPL will be described in the next subsection. The top graph is a motif density trend for window-length 5. The second, third, and fourth trend graphs from the top to the bottom are those for window-lengths 5, 9, 15, and 31 respectively. The third trend graph for window-length 15 has highest motif density values at the times when motif patterns start. The trend graphs for lengths 5 and 9 have times at which sub-patterns of the optimal motifs with length 15 have relatively high motif density values and longer peak durations than those of length 15. The reason for this observation is in the fact that the best motif pattern includes motifs with smaller window-lengths. They also support hypothesis 2.

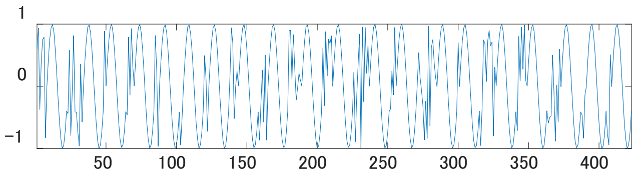


Figure 1: A time series with a motif of length 15 samples.

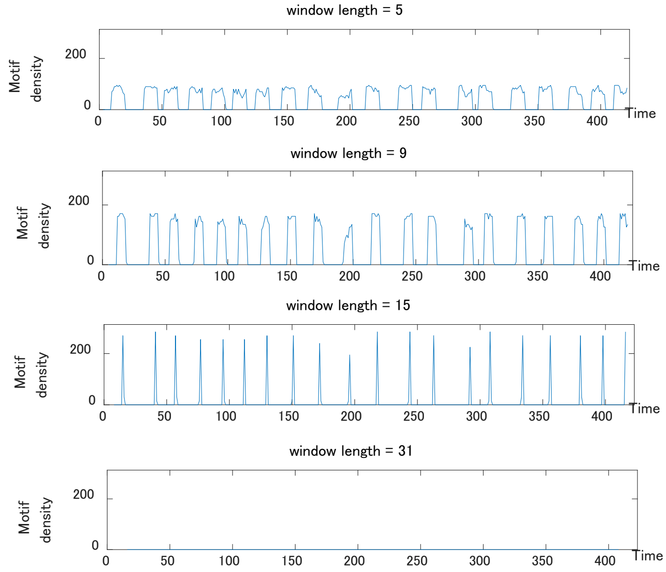


Figure 2: Motif density trend for each window-length (in case of EPL 0.01).

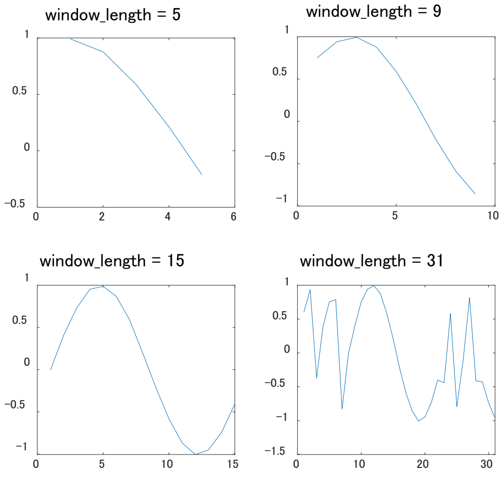


Figure 3: Optimal motif for each window-length.

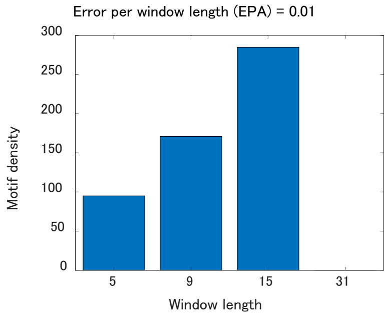


Figure 4: Highest Motif density of the optimal motif for each window-lengths (in case of EPL 0.01).

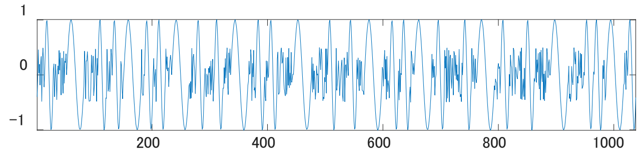


Figure 5: A time series with length 15 and 31 motifs.

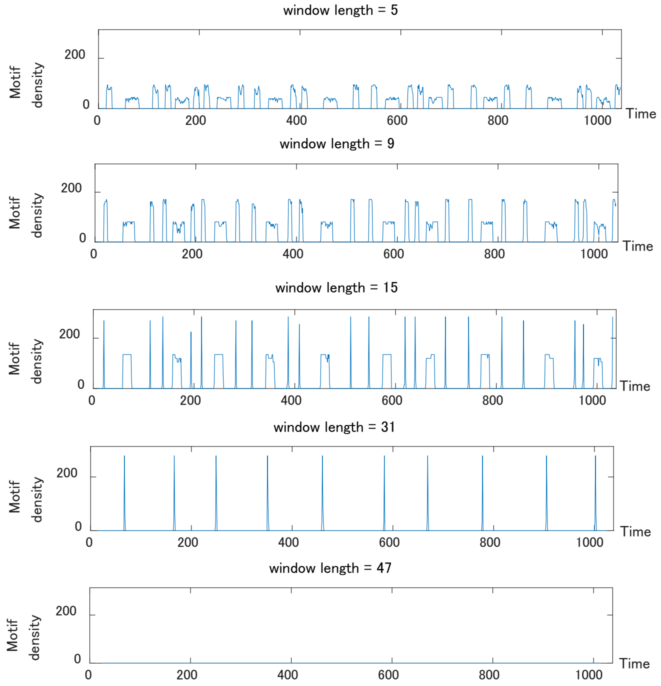


Figure 6: Motif density of each times for each window-lengths (in case of EPL 0.01).

Figure 3 shows each optimal motif in each window-length. The optimal motif in length 15 is also the best motif in the sense that it has the highest motif density as will be shown in Fig. 4. In each graph of Fig. 3, the horizontal axis means time points of each optimal motif subsequence in time series. The vertical axis means the values of each optimal motif. The best motifs for window-lengths 5 and 9 are the sub-patterns of the best motif with length 15. The optimal motif with length 31 is a subsequence including the optimal motif with length 15.

Figure 4 shows the motif density value for each optimal motif with each window-length in case of EPL 0.01. In Fig. 4, the horizontal axis means the length of each optimal motif, and the vertical axis means the motif density of each optimal motif. The window-length that has the highest motif density is 15. It supports hypothesis 1 that "motif density can be used to decide the best window-length". It also supports hypothesis 2, "a maximal motif for a smaller window-length 5, 9 than the best motif length 15 has relatively high motif density value".

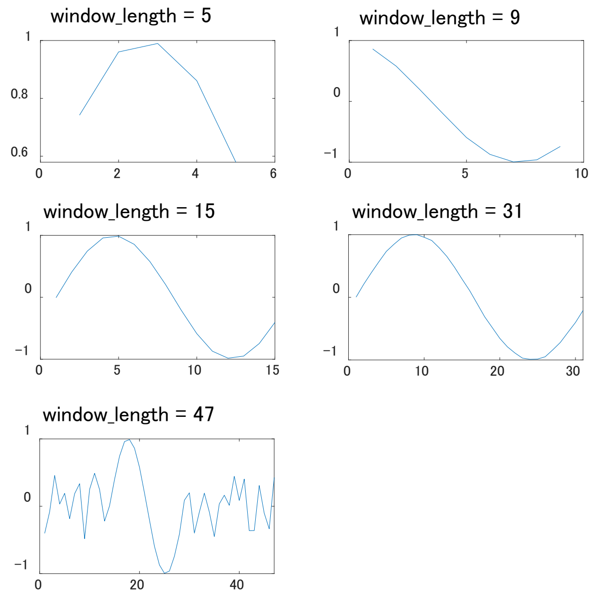


Figure 7: Best motifs for each window-lengths.

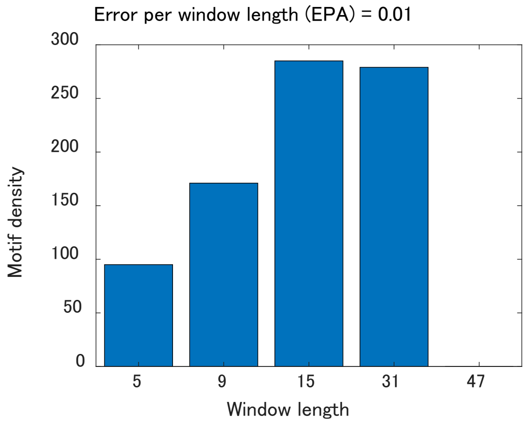


Figure 8: Highest Motif density of the optimal motif with each window-length (in case of EPL 0.01).

(2) Experiment on data set 2

Next, we show that motif density can be used to decide the best motif length (15 and 31) in time series in which two motifs with length 15 and 31 are intentionally embedded.

Figure 5 is a simulated time series that combines sine curves with length 15 and 31 with random subsequences of various lengths. The horizontal axis and the vertical axis in Fig. 5 have the same meanings as those in Fig. 1. Sine curves with length 15 are intentionally embedded motifs. We call this time series data set 2.

Data set 2 is obtained by alternatively arranging 'a random subsequence which has random length between 1 and 31', 'a noisy sine curve whose length of one cycle is 15' and 'a noisy sine curve whose length of one cycle is 31'. In data set 2, a 5-subsequence pattern in which a random subsequence, a sine curve with length 15, a random one, a sine curve with length 31 and a random one are arranged in this order repeat for 10 times. Those random subsequences and the noise of sine curves follow the same random uniform distributions in data set 1.

Figure 6 shows the motif density trend for each window-length for the case of a EPL of 0.01. The horizontal axis and the vertical axis in Fig. 6 have the same meanings as those in Fig. 2. The trend graphs are for window-lengths 5, 9, 15, 31, and 47, from the top to the bottom, respectively. As in the first experiment, the trend graphs for lengths 15 and 31 have the high peaks of motif density at the times when motif patters occur. The trend graphs for 5, 9, and 15 have relatively high motif density values at times when sub-patterns of the motif patterns with length 15 or 31 occur.

Figure 7 shows each optimal motif in each window-length. The optimal motifs in window-lengths 15 and 31 are the best motifs in the sense that they have larger motif densities as will be shown in Fig. 8. (and have been intentionally embedded). The horizontal axis and the vertical axis in Fig. 7 have the same meanings as those in Fig. 3. The optimal motifs of window-lengths 5 and 9 are sub-patterns of the best motifs with window-length 15 or 31. The optimal motif with window-length 47 is a subsequence including the best motif with window-length 31.

Figure 8 shows each motif density trend for each window-lengths. The horizontal axis and the vertical axis in Fig. 8 have the same meanings as those in Fig. 4. It shows that 15 and 31 are the top 2 window-lengths. This supports hypothesis 1. It also shows that window-lengths smaller than 15 have relatively high motif density values. This supports hypothesis 2.

5.2 Error Parameter Selection

This subsection evaluates the hypothesis 3 below.

Hypothesis 3: we can select an appropriate error parameter by means of an Elbow method for the graph of a motif density functions for window-lengths

(1) Experiment on data set 1

Figure 9 shows each error dependency graph of the motif density for each optimal motif with window-length 5, 9, 15, and 31. In Fig. 9, the horizontal axis means EPL values, and the vertical axis means the motif density of at each EPL value.

In data set 1, the window-length of intentionally embedded motifs is 15. The range of EPL for the top graph (a) is from 0 and 2, and that for the bottom one (b) is from 0 to 0.015. The graph (a) shows that when EPL is over 0.8, there are no differences among motif densities for all the window-lengths even though the best motif length is 15. The graph (b) shows that motif densities for window-lengths 5,9 and15 rise quickly at EPL of 0.005, and increase while EPL is from 0.005 to 0.01 and then become constant from EPL values of 0.01. Therefore, 0.01 is an elbow point for window-lengths 5, 9 and 15. On the other hand, the motif density for window-length 31 has constant value 0 for EPL ranging from 0 to 0.015. This observation shows that an appropriate EPL is 0.01 for finding the best motif length shown in the previous subsection. That is, this observation supports hypothesis 3 in case of data set 1.

We compare motif density trend with different EPLs in order to understand the intuitive meaning of EPL. Figure 10 shows the motif density trend for optimal motifs with window-length 5, 9, 15, and 31. The horizontal axis and the vertical axis in Fig. 10 have the same meaning as those in Fig. 2.

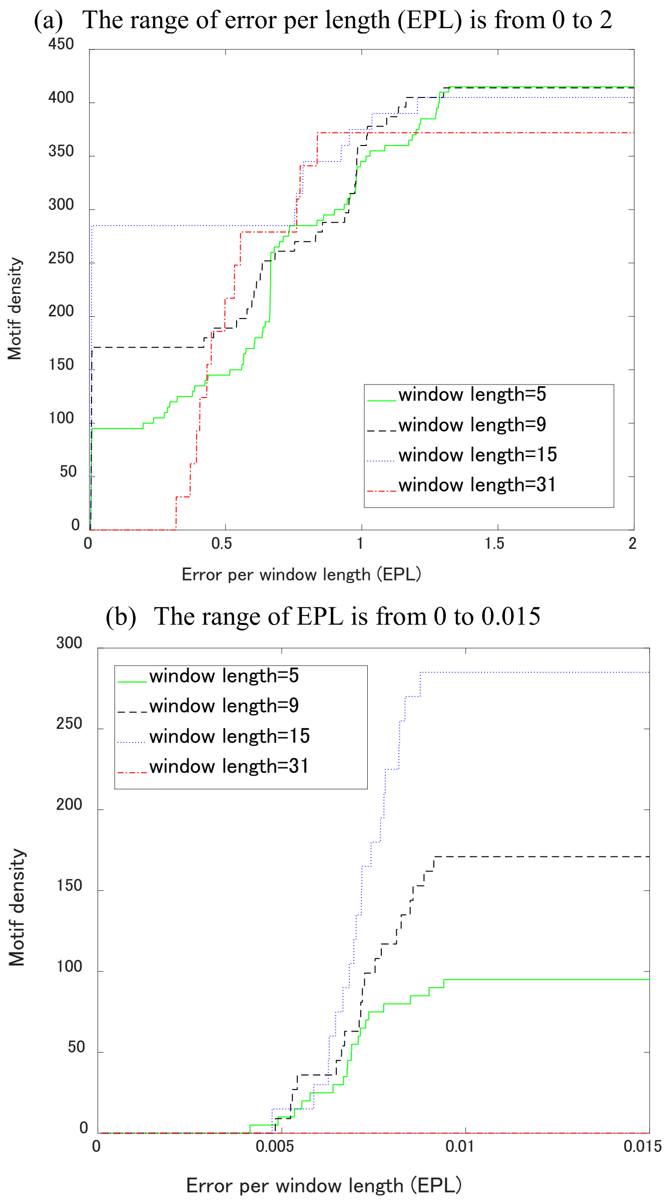
The EPL of the top graph (a), that of the middle one (b) and that of the bottom one (c) are 0.01, 0.1 and 1, respectively. In the case of EPL equal to 0.01, the graph for the best motif length (15) has sharp peaks when similar subsequences occur. On the other hand, in the case of EPL=0.1, the graph for it has only blunt peaks. Furthermore, in the case of EPL=1, there seems to be no peaks. The graphs for smaller window-lengths (5, 9) than the best motif length (15) have similar trends to that for 15. Motif densities for 5 and 9 have relatively high values, because subsequences of a motif are motifs. That is, if X(i:i+14) is a motif , X(i:i+4), X(i+1:i+5), …, and X(i+10:i+14) are also motifs. This is why motif density for window-length 5 and 9 have less sharp peaks than those for window-length 15. This observation also supports hypothesis 2.

(2) Experiment on data set 2

Figure 11 shows each error dependency graph of the motif density for each optimal motif with window-length 5, 9, 15, 31, and 47. The horizontal axis and the vertical axis in Fig. 11 have the same meaning as those in Fig. 9. In data set 2, the window-lengths of intentionally embedded motifs are 15 and 31. The range of EPL for the top graph (a) is from 0 and 2, and that for the bottom one (b) is from 0 to 0.015. Graph (a) shows that when EPL is over 1, there are no differences among motif densities for all the window-lengths, even though the best motif lengths are 15 and 31. Graph (b) shows that motif densities for window-lengths 5,9, and15 rise quickly at EPL=0.005 and increase while EPL ranges from 0.005 to 0.01, and then become constant from about EPL=0.01. Therefore, 0.01 is an elbow point for window-lengths 5, 9, 15, and 31. On the other hand, the motif density for window-length 47 has a constant value 0 for EPL ranging from 0 to 0.015. This observation shows that an appropriate EPL is 0.01 for finding the best motif length shown in the previous subsection. That is, this observation supports hypothesis 3 in the case of data set 2.

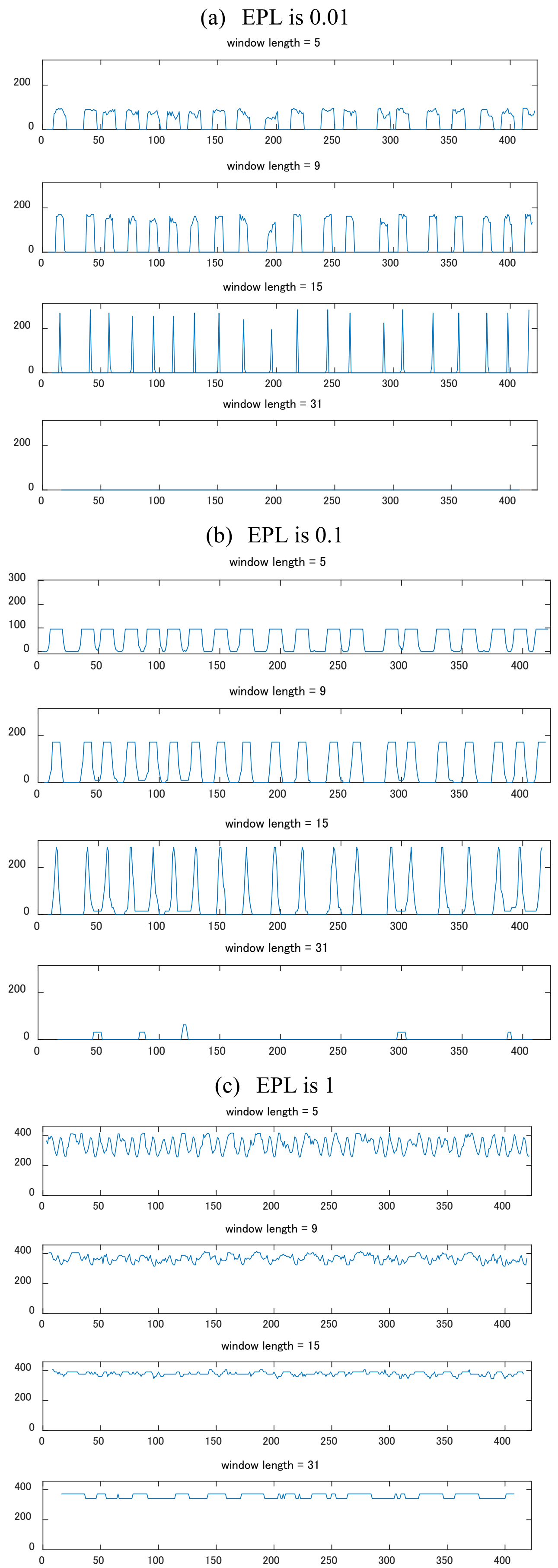
As with experiment 1, we investigate density trend graphs with different EPLs. Figure 12 shows the motif density trend for optimal motifs with window-length 5, 9, 15, 31, and 47. The horizontal axis and the vertical axis in Fig. 12 have the same meaning as those in Fig. 2.

The EPL of the top graph (a), that of the middle one (b), and that of the bottom one are 0.01, 0.1 and 1, respectively. In the case of EPL=0.01, the graph for the best motif lengths 15 and 31 have sharp peaks when similar subsequences occur. The blunt peaks in the graph for window-length 15 correspond to the occurrences of the subsequences of the best motifs with window-length 31. On the other hand, in the case of EPL=0.1, the graphs for window-length 15 and 31 have only blunt peaks. Furthermore, in the case of EPL=1, they have no peaks. As with the experiment on data 1, this observation supports hypothesis 2.


(a) The range of error per length (EPL) is from 0 to 2

(b) The range of EPL is from 0 to 0.015

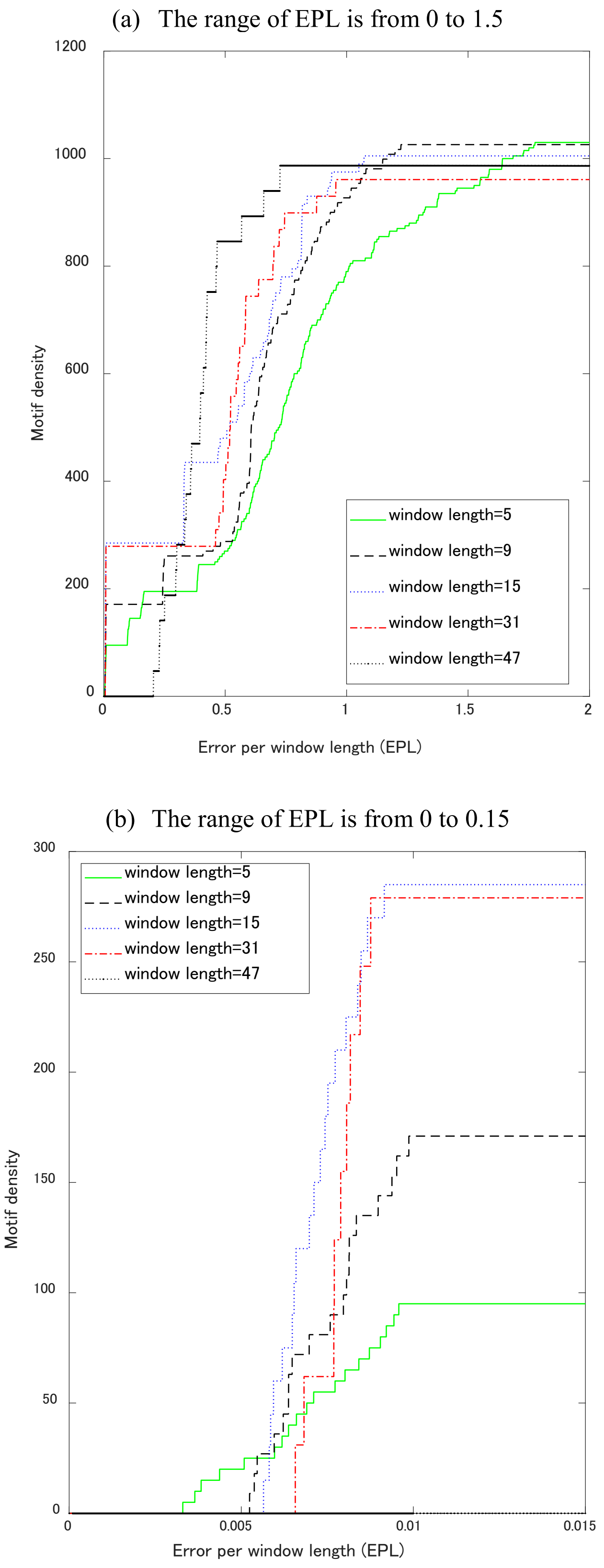
Figure 9: Error dependency of motif density (data 1).


(a) EPL is 0.01

(b) EPL is 0.1

(c) EPL is 1

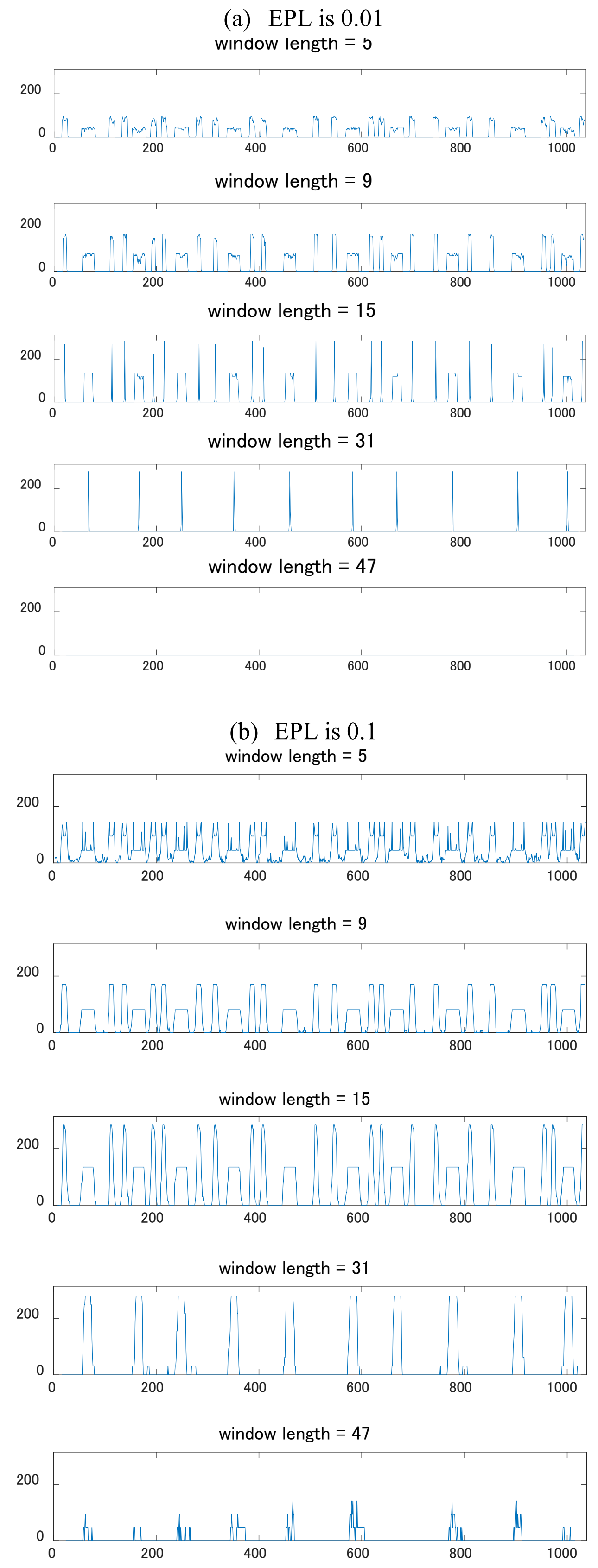
Figure 10: Motif density trends (data1).



(a) The range of EPL is from 0 to 1.5

(b) The range of EPL is from 0 to 0.15

Figure 11: Error dependency of motif density (data 2).



(a) EPL is 0.01

(b) EPL is 0.1

(c) EPL is 1

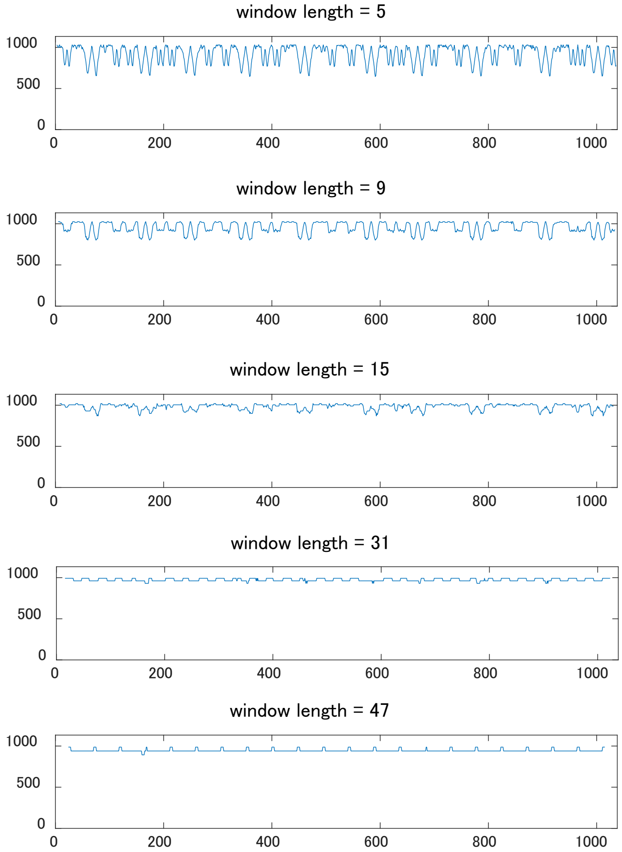


Figure 12: Motif density trends (data2).

6 CONCLUSIONS

We proposed a novel index called 'motif density' together with a selection method to find an appropriate EPL required for defining motif density. The core idea of motif density is in considering a special topology in a subsequence space generated by a time series for avoiding trivial matching and handling different window-lengths. Furthermore, we showed that motif density can decide an optimal window-length in simulated data.

In this paper, we treated the problem of finding one isolated motif in a time series. From a theoretical point of view, it remains as future work how to define and find a sequence of motifs. From an experimental point of view, we plan to apply our algorithms to more complex simulated data, as well as real data.

This work is supported by JSPS KAKENHI Grant Number 17K00161.

REFERENCES

[1] P. Patel, E. Keogh, J. Lin, and S. Lonardi: "Mining Motifs in Massive Time Series Databases", IEEE ICDM pp. 370-377 (2002).

[2] A. Mueen, E. Keogh, Q. Zhu, S. Cash, and M. B. Westover: "Exact Discovery of Time Series Motifs", SDM pp. 473-484 (2009).

[3] T. Maekawa, D. Nakai, K. Ohara, and Y. Namioka: "Toward practical factory activity recognition: unsupervised understanding of repetitive assembly work in a factory", UbiComp pp. 1088-1099 (2016).

[4] Z. Syed, C. M. Stultz, M. Kellis, P. Indyk, and J. V. Guttag: "Motif discovery in physiological datasets: A methodology for inferring predictive elements", TKDD Vol. 4, No.1: 2:1-2:23 (2010).

[5] Y. Zhu, Z. Zimmerman, N. S. Senobari, C. M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh: "Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins", IEEE ICDM, pp. 739-748 (2016).

[6] W. R. Lovallo, M. F. Wilson, A. S. Vincent, B. H. Sung, B. S. McKey, and T. L. Whitsett: "Blood Pressure Response to Caffeine Shows Incomplete Tolerance After Short-Term Regular Consumption", Hypertension vol. 43, No. 4 p.760-765 (2004).

[7] C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh: "Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets", IEEE ICDM, pp. 1317-1322 (2016).

[8] Y. Zhu, M. Imamura, D. Nikovski, and E. Keogh: "New Primitive for Time Series Data Mining", IEEE ICDM, pp. 695-704 (2017).

[9] E. Keogh, J. Lin, and W. Truppel: "Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research", IEEE ICDM, pp. 115-122 (2003).

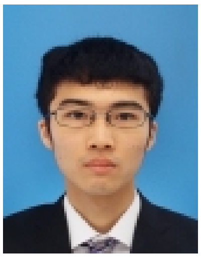
[10] T. Idé: "Why Does Subsequence Time-Series Clustering Produce Sine Waves?", PKDD, pp. 211-222 (2006).

[11] S. Raschka and V. Mirjalili: "Python Machine Learning Second Edition", Packt Publishing, p. 357-358 (2017).

(Received March 6, 2019)
(Revised August 24, 2019)

Makoto Imamura He received a M.E. degree from Kyoto University of Applied Mathematics and Physics in 1986 and a Ph.D. degree from Osaka University of the Information Science and Technology in 2008. From 1986 to 2016, he worked for Mitsubishi Electric Corp and he is presently a Professor at Tokai University. His research interests include machine learning, model-based design and PHM (Prognostics and Health Management). He is a member of IEEE.

Mao Inoue He received a B.E. degree from Tokai University, Japan in 2018. He is a master course student of the school of Information and Telecommunication Engineering at Tokai University. His research interests include IoT systems and data analytics.



Masahiro Terada He received a B.E. degree from Tokai University, Japan in 2019. He is a master course student of the school of Information and Telecommunication Engineering at Tokai University. His research interests include IoT systems and data analytics.

Daniel Nikovski He received a PhD in robotics from Carnegie Mellon University in 2002, and is presently the group manager of the Data Analytics group at Mitsubishi Electric Research Laboratories. He has worked on probabilistic methods for reasoning, learning, planning, and scheduling, and their applications to hard industrial problems. He is a member of IEEE.

Regular Paper

Regression Verification for C Functions with Recursive Data Structure — Using SAW —

Kozo Okano[†], Rin Karashima[†], Satoshi Harauchi[‡], and Shinpei Ogata[†]
[†]Faculty of Engineering, Shinshu University, Japan
[‡]Mitsubishi Electric, Japan
okano@cs.shibshu-u.ac.jp, ogata@cs.shinshu-u.ac.jp

Abstract - Programs are usually revised to improve performance. In such cases, programmers have to ensure that the revised program preserves the behavior of the previous version of the program. Regression testing is performed to check whether both the revised version and the previous version have the same behavior. It, however, requires much time and large number of test-cases. Tools based on formal method might reduce the costs. They ensure that two given programs output the same results for the same inputs based on a logical analysis of their source code and they perform effective path search using SAT/SMT solvers. Software Analysis Workbench (SAW), a novel tool based on formal methods, can check whether two given functions in C act in the same behavior (conformance verification). SAW, however, has a limitation that it cannot check functions dealing with data structures. This paper proposes a new technique for conformance verification on C functions with data structures using SAW. The technique is based on a kind of bounded model checking. We limit the size of data structures which are generated by recursive definitions, in order to limit the space to search. This paper also reports results on performance evaluation that shows our proposed method works for standard data structures.

1 INTRODUCTION

C programs are widely used especially in embedded systems which have a limitation of resources and they are often revised to improve performance. In such cases, programmers have to ensure that the revised program preserves the behavior of the previous version of the program. Usually, regression testing is performed to check whether both of the revised version and the previous version have the same behavior. Thus, it is tedious work and requires large number of test-cases.

Formal technique approach might help to reduce such costs. Based on formal approaches several tools have been developed. Such techniques exhaustively check whether two given programs have always the same output for every same input. Thus, these tools will find potential bugs or confirm the conformance with adequate efficiency. We call this kind of verification formal conformance verification (FCV).

Recent tools, however, do not fully support programs dealing with dynamic data structures especially recursive data structures. Such a program sometimes suffers a halting problem in computability theory.

In this paper, we firstly propose a method for FVC for a program with recursive data structures. In order to avoid the halting problem, the method is based on the bounded model verification technique [1], [2]. We also perform experimental evaluations using Software Analysis Workbench (SAW) [3]. SAW is a recent formal verification tool. We use SeaHorn [4] to compare with SAW. SeaHorn is also a recent formal verification tool for the C language.

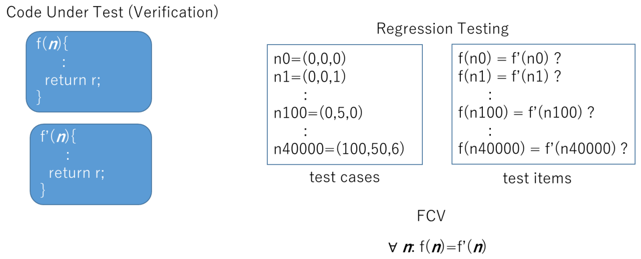


Figure 1: Regression Testing versus FCV

The rest of this paper is organized as follows. Section 2 gives the preliminaries, and Section 3 describes the proposed method. Section 4 describes the experimental evaluation, and Section 5 discusses the results. Finally, Section 6 summarizes this paper.

2 PRELIMINARY AND RELATED WORK

In general, testing is the historical and popular method for checking the quality of source code. For conformance testing, regression testing is widely used. Regression testing, however, has a high time cost and workload. It also has the disadvantage that the verification becomes incomplete in most cases.

Figure 1 shows the difference between the conventional regression testing and FCV approach. As a given Code Under Test (CUT), or CUV Code Under Verification, let us assume that two functions f and f' exist. We want to check that for any input $\boldsymbol{n}$, $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ holds.

In regression testing, we have to prepare a sufficient number of test cases (in Fig. 1, we have 40000 cases) and check all cases by executing a test driver. As we can see, regression testing requires a substantial amount of time and it does not cover whole range of input cases.

2.1 Formal Conformance Verification

The following two theorems are well-known results in Computation Theory [5].

Theorem 1 (Termination Problem (Halting Problem))
The Termination Problem is undecidable.

In other words, $\forall f$ and $\forall \boldsymbol{n} \in \mathbb{Z}^{|\boldsymbol{n}|}$, whether $f(\boldsymbol{n})$ always terminates or not, is undecidable.

Theorem 2
The general conformance checking problem is undecidable.

In other words, $\forall f,\ f'$ and $\forall \boldsymbol{n} \in \mathbb{Z}^{|\boldsymbol{n}|}$, whether $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ always holds or not, is undecidable.

If we restrict the condition, the general conformance checking problem can be decidable. Thus, the restricted conformance checking problem is decidable.

Theorem 3
The restricted conformance checking problem is decidable.

$\forall f,\ f'$ and $\forall \boldsymbol{n} \in \mathbb{Z}_t^{|\boldsymbol{n}|}$, whether $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ always holds or not, is decidable, provided that both f and f' terminate for any $\boldsymbol{n} \in \mathbb{Z}_t^{|\boldsymbol{n}|}$, where $\mathbb{Z}_t$ is a whole set of t-bit integers for some fixed parameter t.

Proof 3
The size of $\mathbb{Z}_t$ is 2^t. Therefore, the size of $\mathbb{Z}_t^{|\boldsymbol{n}|}$ is at most $2^{t|\boldsymbol{n}|}$. The assumption guarantees that we think of only functions f and f' that always terminate for any $bmn \in \mathbb{Z}_t^{|\boldsymbol{n}|}$. Thus, we can compute the result of $f(\boldsymbol{n})$ and $f'(\boldsymbol{n})$ in finite steps $\alpha(\boldsymbol{n})$ and $\alpha'(\boldsymbol{n})$, respectively. In conclusion, we can decide if $f(\boldsymbol{n}) = f'(\boldsymbol{n})$ always holds, in finite steps $2^{t|\boldsymbol{n}|} \cdot \max(\alpha(\boldsymbol{n}), \alpha(\boldsymbol{n}))$. □

In FCV, we check the logical expression $\forall \boldsymbol{n} : f(\boldsymbol{n}) = f'(\boldsymbol{n})$, where functions f and f' are expressed in some logical clauses derived from CUV.

Note that $\boldsymbol{n}$ is usually a vector of bounded integers, such as a 32-bit integer, thus, the number of check cases is finite.

Form Theorem 3, if we suppose that functions f, and f' always terminate then the expression can be efficiently checked using SAT/SMT solvers[6]–[12].

SAW and SeaHorn [4] are tools appearing recently to efficiently check all inputs cases.

2.2 SAW

Software Analysis Workbench (SAW) [3] is developed by Galois inc. It is an open source software which verifies code written in C or Java using a compiler that generates LLVM, or JVM (Java Virtual Machine). Some recent formal verification techniques [13]–[15] use JVM and LLVM as their targets. An LLVM file is compiled from a C, C++, or Objective-C source file. LLVM is a virtual machine instruction set (intermediate representation) and usually used for code optimization in compilers. It, therefore, supports a three-address code scheme and the Static Single Assignment form, which facilitate static analysis for optimizing compiled code. LLVM has pointer types as well, which is mandatory for compilers of C-family languages.

SAW supports equivalence checking between two C functions given in the LLVM format. Both symbolic execution and equivalence checking functions are provided as commands of a script language used in SAW. SAW also supports property checking and has been successfully applied to security domains such as cryptographic protocol analysis.

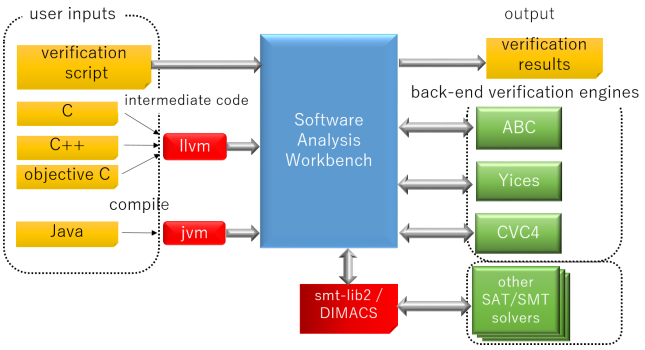


Figure 2: The architecture of SAW

Figure 2 shows the architecture of SAW.

We summarize the features of SAW.

- It uses its own verification script called SAWScript, which is a kind of functional programming languages.
- It has several verification packages that support specific fields:
  - llvm_extract
  - llvm_symexec
  - llvm_verify
  - crucible_llvm_verify
- It has a build-in solver, ABC [17], and can also use three SAT/SMT solvers, Z3 [6], Yices [7], and CVC4 [8].
- It can generate proof constraints with a form of AIG (And-Inverter Graphs) [16], and smtlib2 [18]. Using the format file, other external solvers can be available.

Package llvm_symexec is the original package used by SAW. Crucible_llvm_verify package is provided recently, and supports pointers and data structures in C.

2.2.1 Verification Examples using SAW

The following example shows the verification process for two functions that output a value twice of the input values (See Listing 1, 2, and 3).

Listing 1: Twice Program

```
// reference function
unsigned int reference_function(unsigned int x){
  return x * 2;
}

// implementation
unsigned int implementation_function(unsigned int x){
  return x << 1;
}
```

Listing 1 has two functions reference_function() and implementation_function(). Function reference_function() just outputs the value of the input multiplied by two, while implementation_function() outputs arithmetic left shift of the input value by 1 bit. Though the codes differ from each other, those functions output the same value for any value of the same input. Proof scripts Listing 2 and 3 prove by different approaches that the two functions are equivalent.

Listing 2: Verification script for twice program (llvm_symexec)

```
// llvm_symexec
// .bc is llvm format
load <- llvm_load_module "add.bc";

// reference_function
// variable x is defined as 32bit integer
x <- fresh_symbolic "x" {| [32] |};
// alloc is used when pointer is used
let alloc_ref  = [];
//
let input_ref  = [("x", x, 1)];
//
let output_ref = [("return", 1)];

t1 <-
  llvm_symexec load "reference_function"  alloc_ref
  input_ref output_ref true;

// implementation_function
let alloc_imp  = [];
let input_imp  = [("x", x, 1)];
let output_imp = [("return", 1)];

t2 <-
  llvm_symexec load "implementation_function" alloc_imp
  input_imp output_imp true;

// verification
thm <- abstract_symbolic {{ t1 == t2 }};
result <- prove z3 thm;
//
print result;
```

Listing 3: Verification script for twice program (crucible_llvm_verify)

```
// crucible_llvm_verify
// add.bc
load <- llvm_load_module "add.bc";

// reference_function
let add_setup = do {
//
    x <- crucible_fresh_var "x" (llvm_int 32);
//
    crucible_execute_func [crucible_term x];
//
    crucible_return (crucible_term {{ x << 2 : [32] }});
};

crucible_llvm_verify load "reference_function" [] false
  add_setup abc;
```

Listings 4 and 5 show the results, respectively.

Listing 4: Result (SAW:llvm_symexec)

```
$saw llvm_symexec.saw
Loading file "llvm_symexec.saw"
Running reference_function
Finished running reference_function
Running implementation_function
Finished running implementation_function
Valid
```

Listing 5: Result (SAW:crucible_llvm_verify)

```
$saw crucible_llvm_verify.saw
Loading file "crucible_llvm_verify.saw"
Proof succeeded! @reference_function
Running reference_function
```

Messages “Valid” and “Proof succeeded! @ reference_function” show that the two functions have the same behaviors, for llvm_symexec and crucible_llvm_verify, respectively.

Listings 6, 7, and 8 show the case that FCV outputs counter-examples.

Listing 6: Wrong implemented code

```
// Reference Function
unsigned int reference_function(unsigned int x){
  return x * 2;
}

// Implementation
// (llvm_symexec)
unsigned int implementation_function(unsigned int x){
  if(x == 10){
    return x * 3;
  }
  return x << 1;
}
```

Listing 7: Result (SAW:llvm_symexec)

```
$saw llvm_symexec.saw
Loading file "llvm_symexec.saw"
Running reference_function
Finished running reference_function
Running implementation_function
Finished running implementation_function
prove: 1 unsolved subgoal(s)
Invalid: [x = 10]
```

Listing 8: Result (SAW:crucible_llvm_verify)

```
$saw crucible_llvm_verify.saw
Loading file "crucible_llvm_verify.saw"
Subgoal failed: @reference_function safety assertion:
 literal equality postcondition
SolverStats {solverStatsSolvers = fromList ["ABC"],
 solverStatsGoalSize = 60}
——————Counterexample——————
("x",10)
user error ("crucible_llvm_verify"
(crucible_llvm_verify.saw:8:1-8:21):
Proof failed.)
```

For these cases, when x equals to 10, the behavior differs. SAW correctly shows the counter-examples. This is the most useful advantage of SAW.

2.2.2 ABC

A user of SAW can analyze the LLVM using symbolic execution. The result of the execution is stored in an AIG (And-Inverter Graphs) [16]. AIG data can be verified by a theorem prover called ABC [17]. ABC is especially suited for equivalence checking [19] between two functions represented in AIG. ABC is the default solver for SAW.

2.3 SeaHorn

SeaHorn [4] also verifies C program code using LLVM. SeaHorn has the following features.

- It is easy to use because a programmer can directly write assertions in the code. The notation is based on a simple notion of Design by Contract [20].
- Learning times for the tool are shorter than for other formal based tools.

2.3.1 Verification Example using SeaHorn

Listing 9 shows an example of verification using SeaHorn.

Listing 9: Verification script for twice program(seahorn)

```
#include "seahorn/seahorn.h"
extern int nd(void);

// code under verification
unsigned int implementation_function(unsigned int x){
  return x << 1;
}

int main(){
  int x, val;

  x = nd();
  val = nd();
  val = implementation_function(x);

  // assrtion
  sassert(val == x * 2 );
}
```

Here, function nd stands for non-deterministically, and returns an arbitral value. sassert(P) states that P is true.

Listing 10 shows the result of the verification.

Listing 10: Result (SeaHorn)

```
$sea pf double.c --show-invars
  -- omit --
unsat
Function: main
main@entry: true
main@entry.split: true
```

Note that SeaHorn usually checks unsatisfiability. In other words, unsat is printed if and only if the assertion holds in SeaHorn.

2.4 SAT/SMT solvers

Recently a number of efficient SAT (SATisfiability problem) solvers are emerging and these solvers prove many constraint based problems. A number of satisfiability problem is usually given as a set of clauses, where each clause is a logical disjunction of Boolean variables. The set of clauses is evaluated as a logical conjunction of clauses. Therefore, the set can be evaluated as satisfiable or unsatisfiable. Satisfiability problem is known as NP-complete. However, SAT solvers efficiently proves most of instances.

SMT (Satisfiability Modulo Theories) is an extension of SAT. Each Boolean variable is substituted with inequality expressions over integers or reals. Several classes are known for SMT. Some of these classes are decidable and there are tools which can efficiently proves the satisfiability of these expressions.

2.4.1 Z3

Z3 [6] is one of the famous SMT solvers developed by Microsoft Research. In SMT-COMP, an SMT solver competition, it has excellent results every year. It is one of the built-in SMT solvers by SAW.

2.4.2 CVC4

CVC4 [8] is one of the CVC (Cooperating Validity Checker) series used by the theorem proving system, SVC developed by Stanford University. At SMT-COMP 2017, it won in many divisions.

2.4.3 Yices

Yices [7] is an SMT solver developed at SRI and was upgraded as Yices 2 in 2009. It also had excellent results at SMT-COMP 2017. It is one of the built-in SMT solvers in SAW.

2.4.4 MathSAT

MathSAT [9] is an SMT solver which supports a wide range of theories, such as equality and uninterpreted functions, linear arithmetic, bit-vectors, and arrays. It also supports the computation of Craig interpolants, extractions of unsatisfiable cores, and the generation of models and proofs.

2.4.5 SMT-RAT

SMT-RAT [10] is an SMT solver that can perform parallel processing written in C ++. Since It is not built in as standard in SAW, it is necessary to output a file such as smtlib2 to use it.

2.4.6 minisat

Minisat[11] is one of the representative SAT solvers. It has the minimum set of functions as a SAT solver, and its source code is about 2000 lines. Because SAW does not built in as standard, it is necessary to apply minisat after outputting in Conjunctive Normal Form (CNF) or AIG [16] format file format.

3 OUR PROPOSED METHOD

Program codes with recursive data structures have a loop structure which has a termination condition. The termination condition depends on the recursive data structures, thus we cannot put a bound of the number of iterations to a fixed finite value. For this reason, verification on such a program code faces the so-called termination problem.

In other words, verification on program code with recursive data structures is essentially undecidable.

In order to overcome this problem, in practice, we usually approximate the problem by introducing the idea of bounded verification.

Our proposed method also uses bounded verification by bounding the size of recursive data structures.

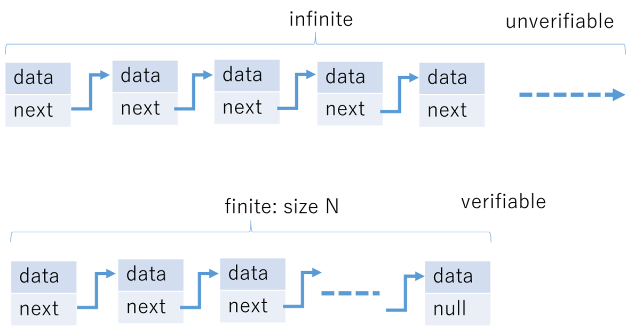


Figure 3: Bounded Model Checking

The upper half of Fig. 3 shows that verification does not end due to infinite size of the list, while the lower half of Fig. 3 shows that verification terminates due to the specifying of a limit of size.

3.1 The Concept

Bounded verification usually terminates iterations of a loop body by a fixed value. We limit the size of recursive data structures. This is essentially the same idea of the usual bounded model checking approaches.

For example, let us consider a linear list. We fixed the size of the list namely n. We produce a verification script for every pattern of the data structure with in size n (See Fig. 3).

Then we perform each formal verification for the produced scripts.

For example, if we choose 100 as n, then we perform formal verification with size 1 to 100 of the linear list. If all of the verification passed, we strongly assume that the program is valid for any size of a list.

The scheme has the advantage that we can choose any feasible value of n, but as we observe, the verification time becomes large as n becomes large.

For every pattern, we perform each formal verification for the produced script.

We use the above idea with the crucible_llvm_verify package for SAW, and evaluated the effectiveness of our proposed method.

In a similar way, for fixed value n, we produce every pattern of binary trees with size 1 to n, where n is the number of nodes in the binary tree. Figure 4 shows every pattern of the binary trees with a size of 3.

3.2 Verification Targets

We perform experiments for the following three data structures.

- Two-level nests
- Linear lists with recursive definition
- Binary trees with recursive definition

In this section, we show every program under verification.

3.2.1 Two-level nests

The program for calculating the summation of 32-bit integers in the parent and children nodes of a two-level nest is shown in Fig. 5.

Listing 11 shows the data structure.

Listing 11: Two-level nest

```
typedef struct s {
    int a;
} s_t;
//parent
typedef struct t {
    int x;
    s_t n;
} t_t;
//reference function
int f_ref(t_t *p) {
    return (p->n).a + p->x;
}
//implementation
int f_imp(t_t *p) {
    return  p->x + (p->n).a;
}
```

The difference between the reference function and the implementation is trivial. We simply change the left and the right terms.

3.2.2 Linear List

The program in Fig. 6 calculates the summation of 32-bit integers in every cell of a linear list.

Listing 12 shows the data structure of the program.

Listing 12: Linear List

```
struct NODE{
  uint32_t val;
  struct NODE* next;
};
typedef struct NODE* node_t;
// reference function
int linear1(node_t node){
  if(node->next == NULL){
    return node->val;
  }else{
    return node->val + linear1(node->next);
  }
}
// implementation
int linear2(node_t node){
  if(node->next != NULL){
    return node->val + linear2(node->next);
  }else{
    return node->val;
  }
}
```

The difference between the reference function and the implementation is in the form of the if statement.

3.2.3 Binary Trees

The program in Fig. 7 calculates summation of 32-bit integers in every node of a binary tree.

Listing 13 shows the data structure of the program.

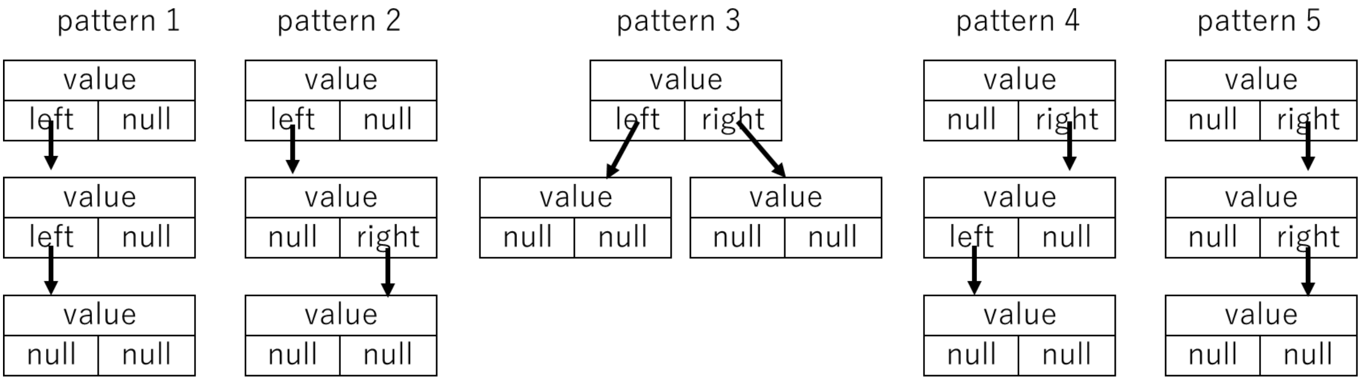


Figure 4: Patterns of Binary Trees

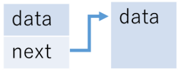


Figure 5: Two level nest

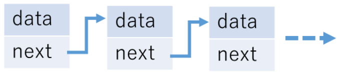


Figure 6: Linear List

Listing 13: Binary Tree

```
// node
struct BTREE {
    uint32_t val;
    struct BTREE* left;
    struct BTREE* right;
};
// reference function
uint32_t pre_order(struct BTREE* tree){
    if(tree == NULL){
        return 0;
    }
    return pre_order(tree->left) +
    pre_order(tree->right) + tree->val ;
}
// implementation
uint32_t in_order(struct BTREE* tree){
    if(tree == NULL){
        return 0;
    }
    return in_order(tree->left) +
    tree->val + in_order(tree->right);
}
```

The difference between the reference function and the implementation is the order of traversal. Thus, the difference is not trivial.

4 EXPERIMENTAL EVALUATION

The experiment environment is summarized as follows.

- OS: Windows 10 64-bit
- CPU: Intel Core i7-4500U CPU @ 1.80GHz 2.39GHz

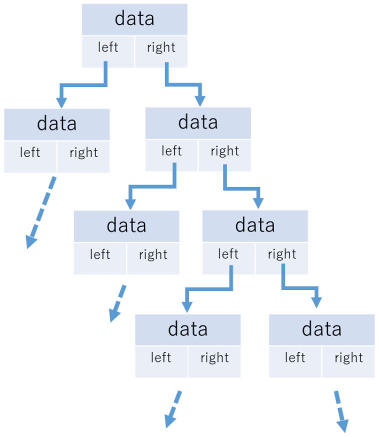


Figure 7: Binary Tree

- Memory: 8.00 GB
- Docker
  - version: 18.01.0-ce
  - Memory: 4096 MB
  - The number of CPUs: 2
- SAW: 0.2 (2018-01-31)
  - LLVM: 3.8.0
  - Z3: 4.5.0
  - Yices: 2.5.2
  - minisat: 2.2.0
  - SMT-RAT: 2.1.0
- SeaHorn: 0.1.0-rc3
  - LLVM: 3.6.0

4.1 Comparison of SMT Solvers: EXP 1

We evaluate the verification results and CPU execution times using the built-in function “Output a file for SAT/SMT solver” of SAW. For ABC, we use a proof package named crucible_llvm_verify. For other SAT/SMT solvers, we use a proof package named llvm_symexec.

4.2 Comparison of SAW and SeaHorn: Exp 2

We evaluate the verification results and CPU execution times using SAW and SeaHorn. We use a proof package named crucible_llvm_verify.

4.3 THE RESULTS

For all results, T/O specifies that verification time is over 3,600 sec. The '–' symbol shows a failure of verification. The unit of time is second.

4.3.1 Comparison of SAT/SMT solvers

Table 1 summarizes the verification of Two-level nest with varying SAT/SMT solvers.

Table 2 summarizes the verification of Linear Lists of size 100 with varying SMT/SAT solvers.

Table 3 summarizes the verification of Binary Trees of depth 5 with varying SAT/SMT solvers.

4.3.2 Comparison with SAW and SeaHorn

Table 4 summarizes verification of Two-level nests using SAW and SeaHorn.

Table 5 summarizes the verification of Linear Lists of size 100 using SAW and SeaHorn.

Table 6 summarizes the verification of Binary Trees of depth 5 using SAW and SeaHorn.

Table 1: Results for Two-level nests

SMT/SAT	ABC	Z3	Yices	minisat	SMTRAT
CPU time	1.06	1.23	1.47	1.24	T/O

Table 2: Results for Linear List

SMT/SAT	ABC	CVC4	Z3	Yices	Mathsat
CPU time	6.981	1.132	59.110	307.307	–

Table 3: Results for Binary Tree

SMT/SAT	ABC	CVC4	Z3	Yices	Mathsat
pattern 1	0.685	0.502	0.484	0.471	0.571
pattern 2	0.682	0.530	0.483	0.467	0.575
pattern 3	0.673	0.491	0.478	0.469	0.604
pattern 4	0.649	0.490	0.465	0.469	0.580
pattern 5	0.646	0.489	0.461	0.501	0.570

Table 4: Results for Two-level nests (SAW and SeaHorn)

Verification Tool	SAW	SeaHorn
CPU Time	1.06	0.104

t!

Table 5: Results for Linear List (SAW and SeaHorn)

Verification Tool	SAW	SeaHorn
CPU time	1.47	–

Table 6: Results for Binary Tree (SAW and SeaHorn)

Verification Tool	SAW	SeaHorn
CPU time	0.51	–

5 DISCUSSION

We verified a variety of code structures by applying bounded verification to functions that deal with structures including recursion. We can verify two-level nest and linear list structures correctly using crucible_lvm_verify. SeaHorn can only verify two-level nests. It was not possible to verify the binary tree structure by all verification methods.

In Exp1, the verification succeeded when we used the "crucible_lvm_verify" package.

In Exp2, it was possible to verify Linear List structures with 100 elements. When we investigated the maximum number of elements that package could be handled, the number of elements was about 5000. In a realistic verification, since sufficient verification can be performed even with the list structure up to 1000 elements, the bounded verification method can be applied. For binary tree structures, we can also obtain good results.

On the other hand, verification using SeaHorn needs less verification time for Two-level nest about one tenth of SAW. Therefore, it is superior to the SAW in view of the verification time. However, when trying to verify a program dealing with recursive structure, recursive functions. it automatically skips the analysis of the structures. Thus, it is impossible to handle programs containing recursive structures. As a result, it can be said that SAW that can handle of recursive structures is superior to that of SeaHorn at the present.

6 CONCLUSION

This paper proposed a new method for Formal Conformance Verification based on bounded model checking for programs with recursive data structures. We also conducted an experimental evaluation using SAW. We showed that the proposed method works well for a simple program which deals with calculations over a linear list.

In future work, we want to improve the performance for binary trees and other complex data structures.

ACKNOWLEDGMENTS

Funding from Mitsubishi Electric Corp. is gratefully acknowledged. The authors wish to thank Mr. Maruyama for his work in the preliminary experiments. The research is being partially conducted as Grant-in- Aid for Scientific Research C (16K00094) and C (17K00111).

REFERENCES

[1] E. Clarke, A. Biere, R. Raimi, Y. Zhu: "Bounded Model Checking Using Satisfiability Solving," Formal methods in system design, Vol.19 Issue 1, pp.7-34 (2012).

[2] T. Liu, M. Nagel, and M. Taghdiri, "Bounded Program Verification using an SMT Solver: A Case Study," Proceedings of the 5th International Conference on Software Testing, Verification and Validation, pp.101-110 (2012).

[3] R. Dockins, A. Foltzer, J. Hendrix, B. Huffman, D. McNamee, and A.Tomb: "Constructing Semantic Models of Programs with the Software Analysis Workbench," Proceedings of VSTTE 2016 (2016).

[4] A. Gurfinkel, T. Kahsai, A. Komuravelli, and J. A. Navas: "The SeaHorn Verification Framework," International Conference on Computer Aided Verification, pp.343-361 (2015).

[5] J. E. Hopcroft, R. Motwani, and J. D. Ullman: "Introduction to Automata Theory, Languages, and Computation (2nd Edition)," Addison Wesley (2000).

[6] L. de Moura and N. Bjørner: "Z3: An efficient SMT solver," Proceedings of TACAS 2008, LNCS Vol. 4963, pp.337-340 (2008).

[7] B. Dutertre: "Yices 2.2," Proceedings of CAV2014, LNCS Vol.8559, pp.737-744 (2014).

[8] C. Barrett, C. L. Conway, M. Deters, L. Hadarean, D. Jovanović, T. King, A. Reynolds, and C. Tinelli: "CVC4," Proceedings of the 23rd international conference on Computer aided verification (CAV'11) pp. 171-177 (2011).

[9] A. Cimatti, A. Griggio, B. Schaafsma, and R. Sebastiani: "The MathSAT5 SMT Solver," Proceedings of TACAS 2013, LNCS Vol. 7795, pp.93-107 (2013).

[10] F. Corzilius, G. Kremer, S. Junges, S. Schupp, and E. Abraham: "SMT-RAT: An Open Source C++ Toolbox for Strategic and Parallel SMT Solving," Proceedings of the International Conference on Theory and Applications of Satisfiability Testing (SAT 2015) pp.360-368 (2015).

[11] N. Een, A. Mishchenko, and N. Sörensson: "Applying Logic Synthesis for Speeding Up SAT," Proceedings of the 10th International Conference on Theory and applications of satisfiability testing pp. 272-286 (2007).

[12] A. Biere, M. Heule, H. Van Maaren, and T. Walsh: "Handbook of Satisfiability," IOS press (2009).

[13] K. Okano, S. Harauchi, T. Sekizawa, S. Ogata, and S. Nakajima: "Equivalence Checking of Java Methods — Toward Ensuring IoT Dependability —," Proceedings of the 26th International Conference on Computer Communications and Networks, pp.1-6 (ICCCN 2017) (August 2017).

[14] C. Belo Lourenco, Si-Mohamed Lamraoui, S. Nakajima, and J. Sousa Pinto: "Studying Verification Conditions for Imperative Programs," Proceedings of 15th International Workshop on Automated Verification of Critical Systems, AVoCS'15 (2015).

[15] Si-Mohamed Lamraoui, S. Nakajima: "A Formula-based Approach for Automatic Fault Localization of Multi-fault Programs," Journal of Information Processing, Vol.24, No.1, pp.88-98 (2016).

[16] A. Darringer, W. H. Joyner, Jr., C. L. Berman, and L. Trevillyan: "Logic synthesis through local transformations," IBM Journal of Research and Development, Vol.25 (4), pp.272-280 (1981).

[17] R. Brayton and A. Mishchenko: "ABC: An Academic Industrial-Strength Verification Tool," LNCS Vol.6174, pp.24-40 (2010).

[18] C. Barrett, A. Stump and C. Tinelli: "The SMT-LIB Standard Version 2.0," (2010).

[19] A. Kuehlmann, V. Paruthi, F. Krohm, and M. K. Ganai: "Robust boolean reasoning for equivalence checking and functional property verification," IEEE Transaction on CAD, vol.21 (12), pp.1377-1394 (2002).

[20] B. Meyer: "Object-Oriented Software Construction, second edition," Prentice Hall (1997).

(Received April 4, 2019)

Kozo Okano received his BE, ME, and PhD degrees in Information and Computer Sciences from Osaka University in 1990, 1992, and 1995, respectively. From 2002 to 2015, he was an Associate Professor at the Graduate School of Information Science and Technology of Osaka University. In 2002 and 2003, he was a visiting researcher at the Department of Computer Science of the University of Kent in Canterbury, and a visiting lecturer at the School of Computer Science of the University of Birmingham, respectively. Since 2015, he has been an Associate Professor at the Department of Computer Science and Engineering, Shinshu University. His current research interests include formal methods for software and information system design. He is a member of IEEE, IEICE, and IPSJ.

Rin Karashima is a graduate student of Shinshu University. His areas of intereset include formal verification and STAMP/STPA.

Satoshi Harauchi
received BE and ME degrees in Information Sciences from Kyoto University in 1996 and 1998, respectively. Since 1998, he has been at the Advanced technology R&D center of Mitsubishi Electric Corporation and is currently interested in software engineering for social infrastructure systems. He is a member of IEICE and JSASS.

Shinpei Ogata
is an Assistant Professor of the Graduate School of Science and Technology in Shinshu University, Japan. He received a PhD from Shibaura Institute of Technology, Japan in 2012. His current research interests include model-driven engineering for information system development. He is a member of IEEE, ACM, IEICE, and IPSJ.

<u>**Regular Paper**</u>

A Distributed Internet Live Broadcasting System for Multi-Viewpoint Videos

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

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

Abstract—With the recent popularization of omnidirectional cameras, multi-viewpoint live videos are now often broadcast via the Internet. Multi-viewpoint live broadcasting services allow viewers to change their viewpoints arbitrarily. To reduce the computational load of video processes such as effect additions, various distributed Internet live broadcasting systems have been developed. These systems are designed for single-viewpoint live videos, in which the screen images (images to be watched by viewers) are the same for all viewers. However, in multi-viewpoint Internet live broadcasting services, the screen images differ according to the viewpoint selected by the viewer. Thus, one of the main research challenges for multi-viewpoint Internet live broadcasting is how to reduce the computational load of adding effects under different screen images. In this paper, we propose and develop a distributed multi-viewpoint Internet live broadcasting system. To distribute the computational load of video processes, our proposed system adopts ECA (event, condition, action) rules. For the systems using ECA rules, it is difficult to determine whether effects should be added on the server side or the player side. To determine this to reduce the computational load effectively, we classify ECA rules.

Keywords: Streaming Delivery, Internet Live Broadcasting, Multi-viewpoint Camera

1 INTRODUCTION

With the recent popularization of omnidirectional cameras, multi-viewpoint live videos are often broadcast through the Internet. In multi-viewpoint Internet live broadcasting services, viewers can arbitrarily change their viewpoints. For example, major live broadcasting services such as YouTube Live and Facebook provide 360° videos in which each user can select their desired viewpoint. In recent Internet live broadcasting services, viewers or broadcasters have been able to add video or audio effects to the broadcast videos. To reduce the computational load including them for adding such effects, a number of distributed Internet live broadcasting systems have been developed [1], [2].

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

In this paper, focusing on this challenge, we propose and develop a distributed multi-viewpoint Internet live broadcasting system. In our proposed system, video effects that can be shared among viewers are added by some distributed processing servers (i.e., on the server side). Video effects that cannot be shared among viewers are added by video players (i.e., on the player side). In such systems, it is difficult to determine whether it is better to add effects on the server side or player side. To determine this so as to effectively reduce the computational load, we use grouped rules. Moreover, we develop a distributed multi-viewpoint Internet live broadcasting system adopting our proposed rules system.

The remainder of this paper is organized as follows. In Section 2, we introduce some related work. We describe the design and the architecture of our proposed system in Section 3. Evaluation results are presented in Section 4 and discussed in Section 5. Finally, we conclude this paper in Section 6.

2 RELATED WORK

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

MediaPaaS encodes, re-encodes, and delivers video using a server machine provided by cloud computing services [2]. Different from MediaPaas, our proposed system establishes load distributions using PIAX [3], a P2P agent platform. The system has multiple servers to broadcast videos, and once a client (video recording terminal) connects to a server to broadcast its recorded video, one of the servers is randomly selected by the load distribution server. The loads caused by broadcasting videos are distributed among the servers. In [1], we confirmed that the video processing time for encoding and distributing videos can be reduced by distributing the processing load to some servers.

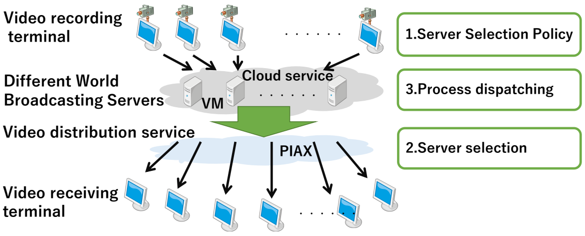


Figure 1: System architecture of the different world broadcasting system

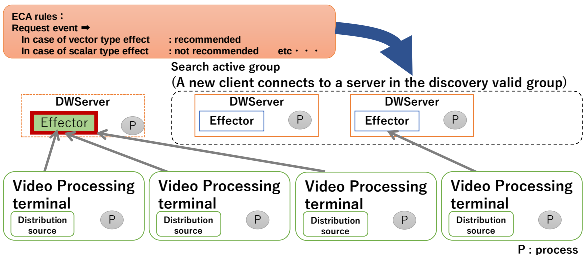

ECA rules :
Request event ➡
In case of vector type effect : recommended
In case of scalar type effect : not recommended etc・・・
Search active group
(A new client connects to a server in the discovery valid group)
DWServer
Effector
P
DWServer
Effector
P
DWServer
Effector
P
Video Processing terminal
Distribution source
P
Video Processing terminal
Distribution source
P
Video Processing terminal
Distribution source
P
Video Processing terminal
Distribution source
P
P : process

Figure 2: Load distribution mechanism using PIAX

An Internet live broadcasting system that allows the viewing of recently recorded videos (playback) was proposed in [4]. Several methods have been proposed for reducing the delay time for the distribution of videos in live Internet broadcasting. SmoothCache 2.0 [5], video data from other peers are cached and distributed among a P2P network. As a result, the communication load and delay times are reduced. Dai et al. also proposed a distributed video broadcasting system using P2P networks to reduce delay times [6]. In the HD method proposed in [7], communication traffic is reduced by simultaneously transmitting image data to a number of viewers using one-to-many broadcasting with one-to-one communication. Even in our proposed system, these delay reduction methods can be applied when delivering videos, but our current research considers the addition of video or audio effects.

Gibbon et al. proposed a system that performs video processing by transferring the data captured by a camera to a computer with high processing capabilities [8]. Ting et al. proposed a system that directly stores images captured by computers with low processing power in external storage devices, such as cloud storage [9]. However, these systems target stored video data and cannot be applied to live Internet broadcasting.

J. Bae et al. proposed a concept of blocks to classify processing flows into several patterns. A block is a minimal unit that specifies the behaviors represented in a process model [10]. A. Frömmgen et al. proposed a learning algorithm of complex nonlinear network nodes by genetic algorithm and ECA rules in [11]. As described in these papers, it is important to learn effective sequences to execute ECA rules. These are not focused on multi-viewpoint image processing. We propose a model focused on image processing with multi-viewpoint image processing.

3 DISTRIBUTED INTERNET LIVE BROADCASTING SYSTEM

In this section, we explain our previously developed cloud-based live broadcasting system using ECA (event, condition, action) rules. After that, we explain our proposed multi-viewpoint Internet live broadcasting system.

3.1 Different World Broadcasting System

3.1.1 Summary

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

Processing loads on different world broadcasting servers can be distributed by considering the load distribution when selecting a server. In conventional systems, load distribution is established by connecting processing servers via a load balancing mechanism such as a load balancer. In this method, when the load distribution mechanism needs to switch to another server while the video is being transmitted, the connection is interrupted. For this reason, it is difficult to switch servers while keeping smooth video plays. Therefore, in the different world broadcasting system, the load balancing mechanism selects a different world broadcasting server based on the requests.

3.1.2 System Architecture

The system architecture of the different world broadcasting system is shown in Fig. 1. There are three types of machine. The first is the client, which has cameras and records live videos. The second is the different world broadcasting servers, which execute processes for videos such as encoding, decoding, or video effect additions. The third type is the viewer, which plays the live videos. Each client selects a different world broadcasting server that executes the desired video effect, and transmits the video effect library and the recorded video to the different world broadcasting server. The different world broadcasting server is a VM of the cloud service that executes video processing on the video transmitted from the clients according to their requests. The video processed by the different world broadcasting server is delivered to the viewers via the video distributions service. In the system, viewers receive the processed video after selecting the server or channel of the video distributions service.

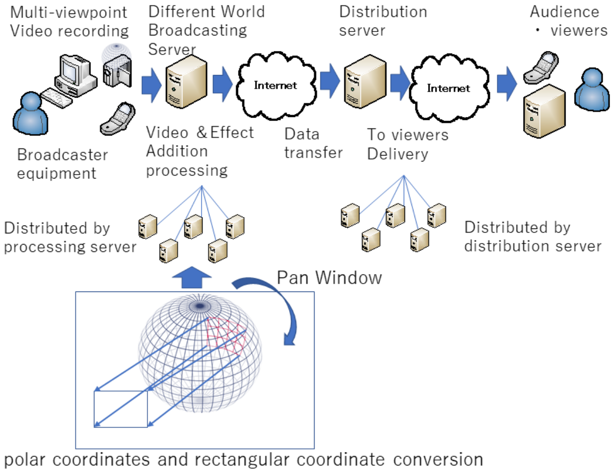

Multi-viewpoint Video recording
Different World Broadcasting Server
Distribution server
Audience · viewers
Internet
Internet
Broadcaster equipment
Video & Effect Addition processing
Data transfer
To viewers Delivery
Distributed by processing server
Distributed by distribution server
Pan Window
polar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

Figure 3: An overview of our designed distributed multi-viewpoint Internet live broadcasting system

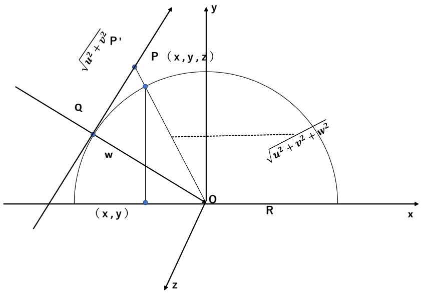
3.1.3 Load Distribution Mechanism

Figure 2 shows the load distribution mechanism of our developed system. The client software and the client side PIAX system are installed in the client. The different world broadcasting server software and the server side PIAX system is installed on the different world broadcasting servers. PIAX [3] is an open-source, Java-based platform middleware that enables efficient server resource searches using the resource search function of the overlay network. The PIAX systems used by the client and the different world broadcasting servers connect with each other via the overlay network. The client side PIAX system searches the overlay network according to the client software requests. The system selects a different world broadcasting server from the list, and then the client side PIAX system returns the IP address of the selected server and listens to the stated port number of the server software. The client software then establishes a connection with the different world broadcasting server and starts transmitting the video. New connections from the client are controlled based on the load state of the different world broadcasting server.

3.2 Extension of Different World Broadcasting System

Figure 3 shows an overview of our designed multi-viewpoint Internet live broadcasting system. As shown in the figure, our system converts the coordinates of videos from polar to rectangular when different world broadcasting servers execute processing. After that, the video images are delivered to viewers.

In this section, we first explain image conversion of multi-viewpoint videos and our design of ECA rules for multi-viewpoint videos. Then show some examples of ECA rules.

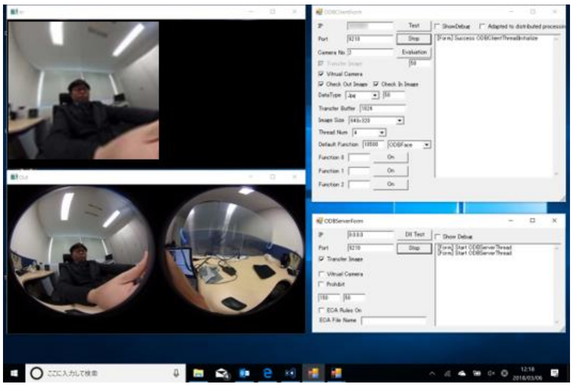


$$|\mathrm{OP'}|^2 = |\mathrm{QP'}|^2 + |\mathrm{QO}|^2 = (u^2 + v^2) + w^2$$

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

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

Figure 4: An image of the coordinate conversion

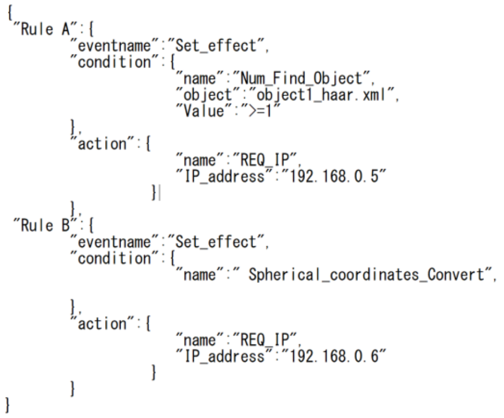

ODBClientForm
IP
Test
□ ShowDebug
□ Adapted to distributed processing
Port 8210
Stop
Camera No 2
Evaluation
☑ Transfer Image 50
☑ Virtual Camera
☑ Check Out Image ☑ Check In Image
DataType Jpg 50
Transfer Buffer 1024
Image Size 640x320
Thread Num 4
Default Function 10500 ODBFace
Function 0 On
Function 1 On
Function 2 On
[Form] Success ODBClientThreadInitialize

ODBServerForm
IP 0.0.0.0
DB Test
□ Show Debug
Port 8210
Stop
☑ Transfer Image
□ Virtual Camera
□ Prohibit
150 50
□ ECA Rules On
ECA File Name
[Form] Start ODBServerThread
[Form] Start ODBServerThread

Figure 5: Server software and client software

3.2.1 Image Conversion of Multi-Viewpoint Videos

With omnidirectional cameras, it is not realistic to take dozens of omnidirectional images from a certain viewpoint and synthesize them on a computer to create a panoramic image. Instead, we create multi-viewpoint videos from panoramic images. The lower left part of Fig. 5 shows two panoramic images (front and back) for a multi-viewpoint video. These panoramic images were obtained from cameras using fisheye lenses. It is necessary to convert these images into a planar image. There are many methods that obtain wide images from car-mounted fisheye lenses and correct the distortion [12]. Figure 4 shows how to obtain a wide image from a panoramic image in our proposed system. As shown in this figure, the wide images are obtained by assuming an imaginary hemispherical border for the panoramic images. The converted wide image is shown in the upper left part of Fig. 5. The conversion transforms virtual hemispherical polar coordinates into rectangular coordinates using the equation (1). In our proposed system, the distributed processing of polar/rectangular coordinates is performed using a different world broadcasting server.



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

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

Figure 6: Examples of ECA rules

Table 1: Events in Communication

Event Name	Description
Receive_Data	Occurs when receives data.
Finish_Transmission	Occurs when data transmission finishes.
Computer_Request	Occurs when recommend server request.
Change_Server	Occurs when DWS server is busy

Table 2: Variables for Conditions in Communication

Variable Name	Description
Data[]	Received data
Transmission_Result	Result of transmission
Turn-around-avg	Turn around time average
T-around-avg-diff	Turn around time average previous differential

Table 3: Actions in Communication

Action Name	Description
Dispatch	Launch Dispatcher

3.2.2 Example of ECA Rules Set

In multi-viewpoint Internet live broadcasting services, the screen images differ according to the viewpoint selected by the user. Thus, the processes for adding effects are usually executed on the users' computers. On the other hand, general processes for Internet live broadcasting such as video encoding, video distribution are executed on the broadcaster's computer or the distribution servers. This means that processes for distributed multi-viewpoint Internet live broadcasting systems have some types. We design three types for ECA rules. One is the effect type that is related to video effects. The viewers' computers are suitable for the execution of this type because they do not need to transmit video data to others. Other one is the communication type and the rules in this type is executed on the computers performing communications. The last one is the processing type. The DWB servers are suitable for the execution of this type because they execute these processes.

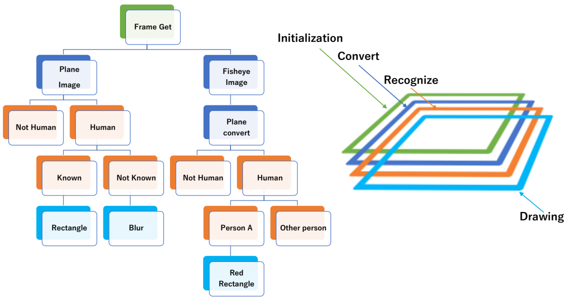


Figure 7: Examples of hierarchical ECA rules

3.2.3 Design of ECA Rules for Multi-Viewpoint Videos

Video effects have various procedures. For example, the face detection process is generally performed before the mosaic effect is applied to the detected face. The "Timer" or "Message" functions of the ECA rules in the proposed system can define such procedures. If the procedure is defined in order-dependent ECA rules, the system needs to execute the rules according to the sequence. Otherwise, if the ECA rules do not depend on the processing request, the system can execute the rules concurrently. This reduces the processing time compared with order-dependent ECA rules. In the current system, it is impossible to process ECA rules in parallel. The parallel processing of cloud computing services is left as a future task. Lists of events, conditions, and actions are described in our previous research [1]. We list some of them in Tables 1-3. Figure 6 shows an example of two ECA rules. In this example, the servers with IP addresses 192.168.0.5 and 6 are assigned as initial machines for the video processing requests from video recording terminals for the condition named "Num_Find_Object" and "Spherical_coordinates_Convert."

In cases where the processes of ECA rules have a sequence, the system should execute the processes in the order of the sequence. For example, Fig. 7 shows an example of the sequences of ECA rules. Some example sequences follow:

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

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

3.2.4 Implementation of Proposed System

We developed a distributed live Internet broadcasting system using Microsoft Azure as a cloud service. The different world broadcasting servers run on the VMs provided by

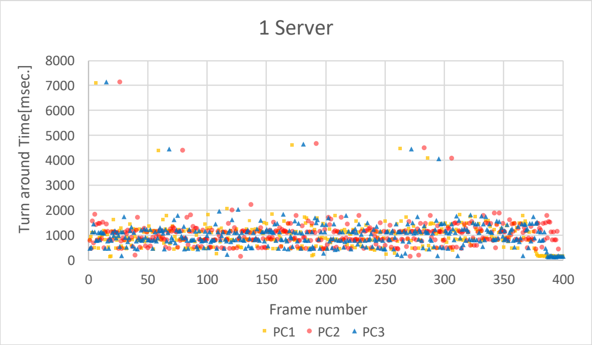


Figure 8: Turnaround times under one cloud server

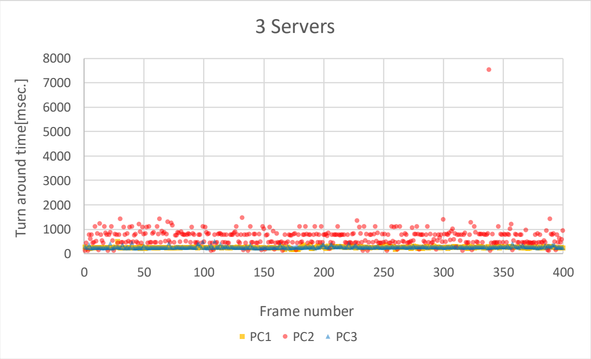


Figure 9: Turnaround times under three cloud servers

Table 4: Specifications of Microsoft Azure VMs

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

Table 5: Specifications of Client PCs

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

Azure. Each VM is logically connected through a virtual network, which is one of the services provided by Microsoft Azure. Figure 5 shows a screenshot of the server software and client software. When starting the process of adding video effects, it provides an interface of the different world broadcasting server software. Using the client software, we can visually check the result of applying the selected effects. The client software holds the IP address of different world broadcasting servers from which video processing can be requested. If the “Apply distributed processing” checkbox in the client software dialog box is selected, the client software requests the different world broadcasting server to execute the video processing specified by the pull-down menu of the initial IP address.

4 EXPERIMENTAL EVALUATION

We evaluated the performances of our implemented system.

4.1 Experimental System

In this evaluation, a different world broadcasting server was running on a VM provided by the Microsoft Azure service. Table 4 lists the specifications of the VM and OS. We used five different VMs for different world broadcasting servers. Open CV, parallelized by Intel’s Parallel Computing Library TBB [13], was used as a library for executing video processing on different world broadcasting servers. The clients were PCs installed at Osaka University. Table 5 lists the specifications of the client PCs. We attached a full omnidirectional camera to only one PC. These PCs communicated with different world broadcasting servers via different home optical networks to avoid network congestion.

4.2 Evaluation Environment

We used a Theta S (RICHO Co., Ltd.) omnidirectional camera for evaluating our proposed system. Each video frame was encoded in JPEG format, transmitted, and received as a USB virtual camera. Image conversion and rule processing were realized by different world broadcasting. In the evaluation, we measured the time from the start of generating the original image data to the time that the processed image data were obtained. To confirm the efficiency of the proposed system, the video processing time, including the processing time of the ECA rule and the turnaround time, was measured as evaluation items.

This includes the following four items:

a) Preprocessing time during which the client receives data (same as the time from the end of reception of previous frame data to the start of the next frame data transmission).

b) Communication time, while different world broadcasting servers receive frame data.

c) Processing time on different world broadcasting servers.

d) Communication time during which the client receives frame data from a different world broadcasting server.

The video processing time is defined as the time from the start of the video processing, excluding the video data reception time, to the end of the processing.

To select an available different world broadcasting server, we used the PIAX overlay network described in subsection 3.1. When a different world broadcasting server overloads, the server sends a notification to the PIAX process on the server side and waits until the load has decreased. The turnaround time of the evaluation was measured in two cases. The first case is a concentrated case in which three clients request video processing from one of three different world broadcasting servers. The second case is a completely distributed scenario in which each of the client requests is sent to a different server.

As the video image processing for the evaluation experiments, face detections are executed on the processing servers after the coordinate conversions.

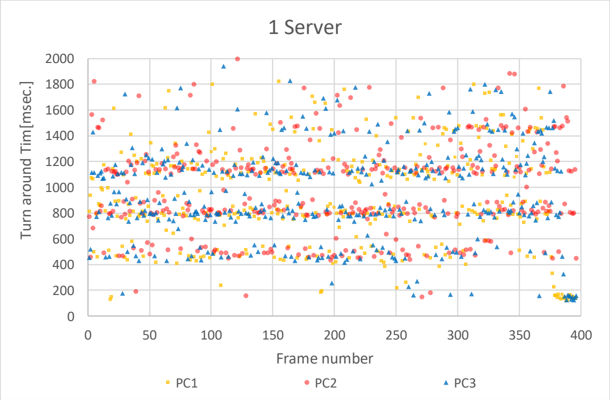


Figure 10: Turnaround time under one server

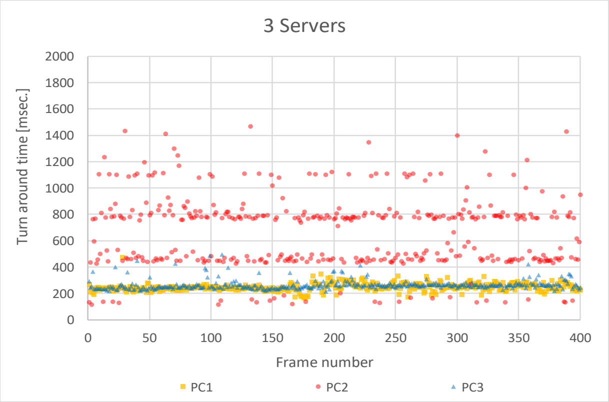


Figure 11: Turnaround time under three servers

4.3 Influence of the Number of Servers

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

Figures 8 and 9 show the evaluation results of the turnaround time under the evaluation environment described in Section 4.1. The horizontal axis is the recorded frame number, and the vertical axis is the turnaround time. In Fig. 8, which shows the case where the load is concentrated on a single different world broadcasting server, the turnaround times gradually increase. Figure 9 shows the case where the image processing requests are distributed to three different world broadcasting servers. In this case, the turnaround time is less than 1500 ms, and the processing delay is around 7500 ms.

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

We also measured the turnaround time required to change the processing server to the recommended different world broadcasting server by PIAX. The average time required to process a query for determining the recommended different world broadcasting server was 16 ms.

As a result, we confirmed that the processing requests are allocated to the different world broadcasting servers based on the ECA rule, and the load is distributed. Moreover, we confirmed that the turnaround time might fluctuate, even for VMs with similar hardware performance, under the effects of communication delays.

4.4 Influence of Computational Load

We measured the turnaround times, changing the loads of DWB servers. To change the loads, we gradually increased the load caused by human face detection every one frame and measured the turnaround time. The results are shown in Figures 10 and 11.

The turnaround times were measured to determine whether the load is concentrated on one virtual server or not. The turnaround times were approximately 1000 ms in this experiment. We have measured the turnaround time for single-viewpoint videos in previous research [14]. The turnaround times were approximately 16 ms. Comparing with this result, the turnaround times for multi-viewpoint videos are longer because the data amount is larger.

5 DISCUSSION

5.1 Fluctuation of Turnaround Times

In our evaluation experiments, even when the calculation load was distributed among the three servers, video processing was sometimes concentrated on only one server. We used two networks for evaluation. (NTT's FLET'S Hikari and K-Opticom's eo light). When requests are concentrated on one different world broadcasting server, the turnaround time is relatively long. When requests from clients are concentrated on a single server, the processing load is distributed to the different world broadcasting server.

Moreover, the video processing involved detecting faces in the video using the specified effect described in the ECA rule. Results using the test rules are shown in Figures 10 and 11, which confirm the fluctuations in turnaround time among cloud computing service VMs. This is caused by actual server performance fluctuations due to differences in the cloud environment of the network distance. Such issues should be considered when the user configures the system.

5.2 Effectiveness of ECA Rules

In previous research, we implemented a distributed Internet live broadcasting system using a cloud service and evaluated its performance. In the installed system, the processing of additional effects is performed using the VM provided by the cloud service. By determining which processing should be allocated to the VM using the ECA rule, it is possible to flexibly change the computer that executes the processing. As a result of our previous evaluations, we confirmed that the turnaround time of the effect adding process could be reduced. As an ECA rule for load balancing to be given to client software, the effect selection made by the client software is set as an event, and a list of corresponding enquiries is set in advance as an IP address. As a result, the server selection is performed automatically and smoothly in the process of adding special video effects.

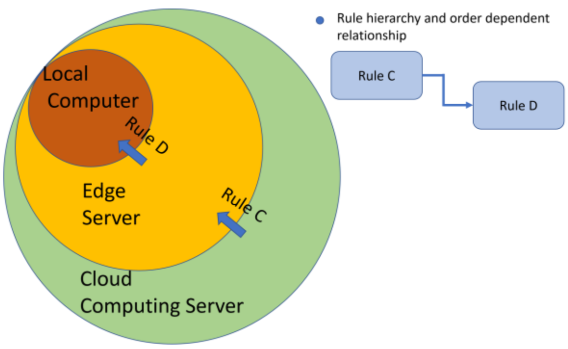


Figure 12: Image of the hierarchical rules

In cases where the processes of an ECA rule have sequences, the system should execute the processes in the order of the sequence. Otherwise, the system can execute them in parallel.

In this paper, we have proposed grouped three-stage rules. After the rules have been prepared, the location for their processing is selected to be either: (1) a local client, (2) edge computing, or (3) cloud computing. An image of the grouped rules is shown in Fig. 12, and the example rules are shown in Fig. 13. In this rule system, the different world broadcasting server that adds the video effects changes as the performance of the current server varies.

6 CONCLUSION

In this research, we have proposed and developed a multi-viewpoint distributed live Internet (different world) broadcasting system. One of the main research challenges for multi-viewpoint Internet live broadcasting systems is how to reduce the computational load required to add effects under different screen images. Our proposed system adopts ECA rules for executing video processes. In this research, we focused on which computer executes the rules, we classified the rules into three types. Each type has a suitable computer for its execution. By classifying ECA rules to these types, our proposed system establishes appropriate execution of rules for video processes.

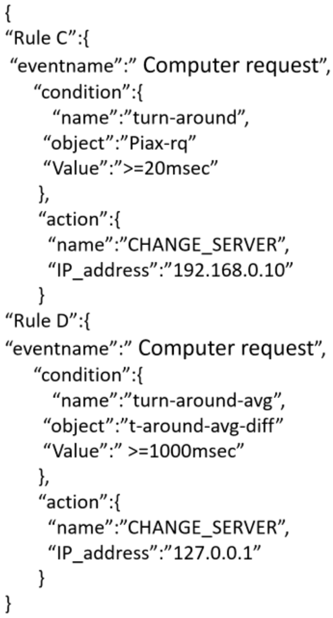
In future work, we plan to exploit edge computing environments in which computers on the edge of the Internet can execute video processes. This could reduce the processing time because edge computers have short turnaround times.

ACKNOWLEDGMENT

This research was supported by a Grants-in-Aid for Scientific Research (C) numbered JP17K00146 and JP18K11316, and by I-O DATA Foundation.

REFERENCES

[1] S. Matsumoto, Y. Ishi, T. Yoshihisa, T. Kawakami, and Y. Teranishi, "Different Worlds Broadcasting: A Distributed Internet Live Broadcasting System with Video and Audio Effects," in Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 71-78 (2017).



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

Figure 13: Example of ECA rules

[2] S. Matsumoto, Y. Ishi, T. Yoshihisa, T. Kawakami, and Y. Teranishi, "A Design and Implementation of Distributed Internet Live Broadcasting Systems Enhanced by Cloud Computing Services," in Proc. of International Workshop on Informatics (IWIN), pp. 111-118 (2017).

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

[4] Y. Gotoh, T. Yoshihisa, H. Taniguchi, and M. Kanazawa, "Brossom: a P2P Streaming System for Webcast," Journal of Networking Technology, Vol. 2, No. 4, pp. 169-181 (2011).

[5] R. Roverso, R. Reale, S. El-Ansary, and S. Haridi, "Smooth-Cache 2.0: CDN-quality adaptive HTTP live streaming on peer-to-peer overlays," in Proc. of ACM Multi-media Systems Conference (MMSys), pp. 61-72 (2015).

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

[7] T. Yoshihisa and S. Nishio, "A division-based broadcasting method considering channel bandwidths for NVoD services," IEEE Transactions on Broadcasting, Vol. 59, No. 1, pp. 62-71 (2013).

[8] D. Gibbon and L. Begaja, "Distributed processing for big data video analytics," IEEE ComSoc MMTC E-Letter, Vol. 9, No. 3, pp. 29-31 (2014).

[9] W.-C. Ting, K.-H. Lu, C.-W. Lo, S.-H. Chang, and P.C. Liu, "Smart Video Hosting and Processing Platform for Internet-of-Things," in Proc. of IEEE International Conference on Internet of Things (iThings), pp. 169-176 (2014).

[10] J. Bae, H. Bae, S. Kang, and Y. Kim, "Automatic Control of Workflow Processes Using ECA Rules," IEEE Transaction On Knowledge and Data Engineering, Vol. 16, No. 8, pp. 1010-1023 (2004).

[11] A. Frömmgen, R. Rehner, M.Lehn, and A. Buchmann, "Fossa: Using Genetic Programming to Learn ECA Rules for Adaptive Networking Applications," IEEE Conference on Local Computer Networks (LCN), pp.197-200 (2015).

[12] J. Jeong, H. Kim, B. Kim, and S. Cho, "Wide Rear Vehicle Recognition Using a Fisheye Lens Camera Image" IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), pp. 691-693 (2016).

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

[14] S. Matsumoto, Y. Ishi, T. Yoshihisa, T. Kawakami, and Y. Teranishi, "A Distributed Internet Live Broadcasting System Enhanced by Cloud Computing Services," International Journal of Informatics Society (IJIS), Vol. 10, No. 1, pp. 21-29 (2018).

(Received April 14, 2019)

Satoru Matsumoto received his Diploma's degrees from Kyoto School of Computer Science, Japan, in 1990. He received his Master's degrees from Shinshu University, Japan, in 2004. From 1990 to 2004, he was a teacher in Kyoto School of Computer Science. From 2004 to 2007, he was Assistant Professor of The Kyoto College of Graduate Studies for informatics. From 2007 to 2010, he was Assistant Professor of Office of Society Academia Collabo-ration, Kyoto University. From 2010 to 2013, he was Assistant Professor of Research Institute for Economics & Business Administration, Kobe University. From 2015 to 2016, he was a specially appointed assistant professor of Cybermedia Center, Osaka University. From April 2016 to September 2016, he became a specially appointed researcher. Since November 2016, he became an assistant professor. His research interests include distributed processing systems, rule-based systems, and stream data processing. He is a member of IPSJ, IEICE, and IEEE

Tomoki Yoshihisa received the Bachelor's, Master's, and Doctor's degrees from Osaka University, Osaka, Japan, in 2002, 2003, 2005, respectively. Since 2005 to 2007, he was a research associate at Kyoto University. In January 2008, he joined the Cybermedia Center, Osaka University as an assistant professor and in March 2009, he became an associate professor. From April 2008 to August 2008, he was a visiting researcher at University of California, Irvine. His research interests include video-ondemand, broadcasting systems, and webcasts. He is a member of the IPSJ, IEICE, and IEEE.

Tomoya Kawakami received his B.E. degree from Kinki University in 2005 and his M.I. and Ph.D. degrees from Osaka University in 2007 and 2013, respectively. From 2007 to March 2013 and from July 2014 to March 2015, he was a specially appointed researcher at Osaka University. From April 2013 to June 2014, he was a Ph.D. researcher at Kobe University. Since April 2015, he has been a assistant professor at Nara Institute of Science and Technology. His research interests include distributed processing systems, rule-based systems, and stream data processing. He is a member of the IPSJ and IEEE.

Yuuichi Teranishi received his M.E. and Ph.D. degrees from Osaka University, Japan, in 1995 and 2004, respectively. From 1995 to 2004, he was engaged Nippon Telegraph and Tele-phone Corporation (NTT). From 2005 to 2007, he was a Lecturer of Cybermedia Center, Osaka University. From 2007 to 2011, He was an associate professor of Graduate School of Information Science and Technology, Osaka University. Since August 2011, He has been a research man-ager and project manager of National Institute of Information and Communications Technology (NICT). He received IPSJ Best Paper Award in 2011. His research interests include technologies for distributed network systems and applications. He is a member of IPSJ, IEICE, and IEEE.

Regular Paper

Search of Elliptic Curves Suitable for Signature

Masaaki Shirase[†]

[†]School of Systems Information Science, Future University Hakodate, Japan
shirase@fun.ac.jp

Abstract - Elliptic curve signatures as ECDSA have features that the processing is faster and the signature length is shorter than those of RSA signatures with same security. The use of elliptic curve signatures was decided for the V2X communication with limited bandwidth. However, higher processing speed is required.

In an elliptic curve signature using 256-bit prime p, thousands of modular multiplications $X \cdot Y \bmod p$ performed according to a signature algorithm are dominant. Therefore, how to speed up multiplications and $\bmod p$ computations is one of the objectives of researches on elliptic curve signature implementations. One of speeding up method of reduction $\bmod p$ is to use a special form of prime called pseudo Mersenne prime such that $p = 2^n - k$, where k is a small value. However, in an elliptic curve signature, computation of $\bmod l$ with another integer l, which is the order of a base point, is also required although the number is a few.

In this paper, the authors give a program to construct elliptic curves such that reduction $\bmod l$ can be computed as mod a pseudo Mersenne prime. The program found elliptic curves $638y^2 = x^3 + 10x^2 + x \bmod p = 2^{256} - 58097$, $82y^2 = x^3+18x^2+x \bmod p = 2^{256}-507225$, and $3805y^2 = x^3 + 18x^2 + x \bmod p = 2^{256} - 979077$ [1].

Keywords: Elliptic Curve, Elliptic Curve Signature, Modular Multiplication, Pseudo Mersenne Prime

1 INTRODUCTION

Recently, elliptic curve signatures as ECDSA are often used in the TLS communication and block chains. In Europe, it was decided to use an elliptic curve signature in the V2X communication [5]. Elliptic curve signature compared with RSA signature has the advantage that signature generation/ verification is faster and signature length is shorter with same security. However, signature verification in the V2X communication requires further speeding up.

Elliptic curve is a cubic curve given by

$$y^2 = x^3 + ax + b \text{ (Weierstrass form), or}$$
$$By^2 = x^3 + Ax^2 + x \text{ (Montgomery curve),}$$

where x and y are variables. A remarkable feature of elliptic curve is that an operation $+$ is defined[2] in the set of points on E, and the set forms a group for the operation [15].

Dominant processes of the operation $+$, which is explained in Sec.2.2, is modular multiplications

$$X \cdot Y \bmod p \tag{1}$$

for $X, Y \in \mathbb{F}_p = \{0, 1, 2, \cdots, p-1\}$, where p is typically a 256-bit prime. The calculation of (1) is divide into

$$Z \leftarrow \underbrace{X}_{256 \text{ bit}} \cdot \underbrace{Y}_{256 \text{ bit}}, \tag{2}$$
$$W \leftarrow \underbrace{Z}_{512 \text{ bit}} \bmod \underbrace{p}_{256 \text{ bit}}. \tag{3}$$

Dominant operation of elliptic curve cryptosystems (ECCs) including elliptic curve signatures is thousands of modular multiplications (1). Therefore, it is important to speed up (2) and (3) to speed up processes of elliptic curve signatures. As explained in Sec.2.4, using Montgomery curve rather than Weierstrass form reduces the number of modular multiplications required for signature generation and verification. Moreover, when a coefficient A of Montgomery curve is 6, 10, 14, and 18, the number of modular multiplications is further reduced.

Karatsuba method, use of high speed multiplier, and parallel implementation speed up integer multiplications. Montgomery reduction [8] that can be applied to arbitrary odd number p is a famous method for reduction $\bmod p$. Also, when p is a pseudo Mersenne prime written as $p = 2^n - k,\ k < 2^{n/2}$, reduction $\bmod p$ can be very efficient.

Curve25519 [1] is a Montgomery curve

$$E_{25519} : y^2 = x^3 + 486662x^2 + x$$

for a pseudo Mersenne prime $p = 2^{255} - 19$. Curve25519 is secure and suitable for high-speed implementation of ECCs and then it has attracted attention in recent years. Some of NIST curves [11], which are elliptic curves for ECCs standardized by NIST, also use pseudo Mersenne prime.

In public key encryptions as ECElGamal and key agreements as ECDH using elliptic curves, required reduction is $\bmod p$ only. On the contrary, in elliptic curve signatures, required reductions are not only $\bmod p$ but also $\bmod l$, where l is the order of a base point.

The order l of a base point on Curve25519 and NIST curves is not pseudo Mersenne prime. Hence, when implementing a signature using these curves, we have to use Montgomery reduction to compute reduction $\bmod l$. However, when implementing an elliptic curve signature by hardware and applying the high-speed reduction for $\bmod p$ and Montgomery

[1] This work was supported by JSPS KAKENHI Grant Number 16K00188.

[2] The operation $+$ is conventionally used for the operation, however, it is different from ordinary addition.

reduction for $\bmod l$ in-creases the hardware scale. Applying Montgomery reduction for both of $\bmod p$ and $\bmod l$ increases computation time. Therefore, it is desirable to be able to compute $\bmod l$ by same way of $\bmod p$.

The purpose of this paper is to make a program to find Montgomery curves such that $\bmod\ l$ can be computed by same way of $\bmod p$ for pseudo Mersenne prime and $A = 6$, 10, 14, and 18, and to give examples of such curves.

Sec.2 explains the definition of elliptic curve, operation $+$, scalar multiplication, coordinate system, secure elliptic curve, and Curve25519. Sec.3 introduces ECDSA that is the most popular elliptic curve signature. Sec.4 introduces efficient reduction methods. Sec.5 is the contribution of this paper. Sec.5 proposes a requirement for elliptic curves to be suitable for ECDSA. Then, Sec.5 makes a program to find elliptic curves that meet the requirement, and gives examples of such curves. Sec.**??** concludes this paper and gives future work.

2 ELLIPTIC CURVE

Sec.2 introduces subjects of elliptic curves required for this paper. For details for subjects of Secs.2.1, 2.2, and 2.3, refer to [15] or [4].

2.1 Definition of Elliptic Curve

Elliptic curve is a cubic curve given by

$$E : y^2 = x^3 + ax + b \text{ (Weierstrass form)} \tag{4}$$

or

$$E : By^2 = x^3 + Ax^2 + x \text{ (Montgomery curve)} \tag{5}$$

with variables x, y. When used in cryptosystems, Montgomery curve (5) is often selected because it can reduce the cost of cryptographic processes.

For a prime p, the set $\mathbb{F}_p$ is defined as

$$\mathbb{F}_p = \{0, 1, 2, \cdots, p-1\}$$

Then, elliptic curve E can be considered on $\mathbb{F}_p$. We consider an elliptic curve

$$E' : y^2 = x^3 + 2$$

on $\mathbb{F}_5$ as an example. For $x, y \in \mathbb{F}_5 = \{0, 1, 2, 3, 4\}$, when

$$y^2 = x^3 + 2 \bmod 5$$

is satisfied, (x, y) is regarded as a point in E' on $\mathbb{F}_5$. For example, $(2, 0)$ is a point in E' on $\mathbb{F}_5$ because

$$0^2 = 2^3 + 2 \bmod 5$$

is satisfied. $(1, 1)$ is not a point in E' on $\mathbb{F}_5$ because

$$1^2 \neq 1^3 + 2 \bmod 5.$$

Executing a program (written for PARI/GP [12]) of Fig. 1, we see that all points in E' on $\mathbb{F}_5$ are

$$\{(2, 0), (3, 2), (3, 3), (4, 1), (4, 4)\}.$$

```
\\Finding points on elliptic curve
{
  p=5;
  for(x=0,p-1,
    for(y=0,p-1,
      if((y^2-(x^3+2))%p==0,
        print([x,y]);
      );
    );
  );
}
```

Figure 1: Program for finding $\mathbb{F}_p$ points on E'

The set adding this set with the point at infinity $\mathcal{O}$ [3] is written as $E'(\mathbb{F}_5)$.

$$E'(\mathbb{F}_5) = \{(2, 0), (3, 2), (3, 3), (4, 1), (4, 4), \mathcal{O}\} \tag{6}$$

The order of $E(\mathbb{F}_p)$, $\#E(\mathbb{F}_p)$, is defined as the number of points in $E(\mathbb{F}_p)$. Hence, the order of $E'(\mathbb{F}_5)$ is 6. The trace of $E(\mathbb{F}_p)$ is defined as an integer t such that

$$\#E(\mathbb{F}_p) = p + 1 - t.$$

When a Montgomery curve $E : By^2 = x^3 + Ax^2 + x$ on $\mathbb{F}_p$ has the trace t, another Montgomery curve $E' : B'y^2 = x^3 + Ax^2 + x$ has the trace t or $-t$. In other words, we have

$$\#E'(\mathbb{F}_p) = \#E(\mathbb{F}_p) \text{ or } 2p + 2 - \#E(\mathbb{F}_p).$$

When $\#E'(\mathbb{F}_p) = 2p + 2 - \#E(\mathbb{F}_p)$, E' is called the twist of E.

Let E be an elliptic curve and L the order of $E(\mathbb{F}_p)$. Then, it is known that

$$p + 1 - 2\sqrt{p} \leq L \leq p + 1 + 2\sqrt{p} \tag{7}$$

is held (Hasse's theorem). That means L is close to p. Conversely, for a prime p and an integer L satisfying (7), Weierstrass form elliptic curve E exists such that $\#E(\mathbb{F}_p) = L$. On the other hand, when E is a Montgomery curve, the order is always a multiple of 4.

2.2 Operation $+$

Let E be an elliptic curve given by Weierstrass form (4) or Montgomery curve (5). Then, the operation $+$ in $E(\mathbb{F}_p)$ is defined as follows.

1. For any $P \in E(\mathbb{F}_p)$,

$$P + \mathcal{O} = \mathcal{O} + P^4.$$

2. In the case of $P = (x_1, y_1), Q = (-x_1, y_1) \in E(\mathbb{F}_p)$,

$$P + Q = \mathcal{O}.$$

[3] When considering an elliptic curve in the real plane, $\mathcal{O}$ is intuitively (∞, ∞) and then $\mathcal{O}$ is called "the point at infinity." When also considering an elliptic curve in $\mathbb{F}_p$, $\mathcal{O}$ is called the point at infinity [16, IV.1]. Although $\mathcal{O}$ is a special point, it can be dealt with as normal one in projective coordinate system. Refer to [16, Appendix A] for details.

[4] $\mathcal{O}$ plays a role of zero element in the operation.

Table 1: Computation of $P_i + P_j$ on $E'(\mathbb{F}_5)$

	P_0	P_1	P_2	P_3	P_4	P_5
P_0	P_0	P_1	P_2	P_3	P_4	P_5
P_1	P_1	P_0	P_4	P_5	P_2	P_3
P_2	P_2	P_4	P_3	P_0	P_5	P_1
P_3	P_3	P_5	P_0	P_2	P_1	P_4
P_4	P_4	P_2	P_5	P_1	P_3	P_0
P_5	P_5	P_3	P_1	P_4	P_0	P_2

3. In the case of $P = (x_1, y_1), Q = (x_2, y_2) \in E(\mathbb{F}_p), x_1 \neq x_2$, and $P \neq Q$, $P + Q = (x_3, y_3)$ is computed as

$$\left.\begin{aligned} \lambda &= \frac{y_1 - y_2}{x_1 - x_2}, \\ x_3 &= \lambda^2 - x_1 - x_2, \\ y_3 &= \lambda(x_1 - x_3) - y_1. \end{aligned}\right\} \quad (8)$$

4. In the case of $P = Q = (x_1, y_1) \in E(\mathbb{F}_p)$, $P + Q = (x_3, y_3)$ is computed as

$$\left.\begin{aligned} \lambda &= \begin{cases} \dfrac{3x_1^2 + a}{2y_1} & E\text{: Weierstrass} \\ \dfrac{3x_1^2 + Ax_1 + 1}{2By_1} & E\text{: Montgomery,} \end{cases} \\ x_3 &= \lambda^2 - 2x_1, \\ y_3 &= \lambda(x_1 - x_3) - y_1, \end{aligned}\right\} \quad (9)$$

where a, A, and B are coefficients of Eqs. (4) and (5). Eqs.(8) and (9) are called addition formula and doubling formula, respectively.

For any $P, Q, R \in E(\mathbb{F}_p)$, the following are held.

1. $(P + Q) + R = P + (Q + R)$
2. $P + \mathcal{O} = \mathcal{O} + P = P$
3. For $P = (x_1, y_1)$ and $Q = (-x_1, y_1)$, $P + Q = \mathcal{O}$.
4. $P + Q = Q + P$

That the above holds means that $E(\mathbb{F}_p)$ forms a group. This property is very significant, and it is also used for cryptosystems.

Let $E'(\mathbb{F}_5)$ of (6) be written as

$$E'(\mathbb{F}_5) = \left\{ \begin{array}{l} P_0 = \mathcal{O}, P_1 = (2, 0), P_2 = (3, 2), \\ P_3 = (3, 3), P_4 = (4, 1), P_5 = (4, 4) \end{array} \right\}. \quad (10)$$

Then, all results of $P_i + P_j$ are given by Table 1.

2.3 Scalar Multiplication

For a base point $P \in E(\mathbb{F}_p)$ and a natural number n, additions of n terms of P,

$$nP = P + P + \cdots + P$$

is called scalar multiplication. For $P \in E(\mathbb{F}_p)$, the order of P is defined as the smallest positive integer l such that $lP = \mathcal{O}$.

For the order L of $E(\mathbb{F}_p)$ and the order l of $P \in E(\mathbb{F}_p)$, the followings are held (Lagrange's theorem).

1. l is a divisor of L.
2. $LP = \mathcal{O}$.

Let $E'(\mathbb{F}_5)$ be of (10). Then, we see that $2P_1$, $3P_1$, $4P_1$, $\cdots$ are

$$\begin{aligned} 2P_1 &= P_1 + P_1 = \mathcal{O}, \\ 3P_1 &= 2P_1 + P_1 = \mathcal{O} + P_1 = P_1, \\ 4P_1 &= 3P_1 + P_1 = P_1 + P_1 = \mathcal{O}, \\ 5P_1 &= 4P_1 + P_1 = \mathcal{O} + P_1 = P_1, \\ 6P_1 &= 5P_1 + P_1 = P_1 + P_1 = \mathcal{O}, \end{aligned}$$

and $2P_4, 3P_4, 4P_4, \cdots$ are

$$\begin{aligned} 2P_4 &= P_4 + P_4 = P_3, \\ 3P_4 &= 2P_4 + P_4 = P_3 + P_4 = P_1, \\ 4P_4 &= 3P_4 + P_4 = P_1 + P_4 = P_2, \\ 5P_4 &= 4P_4 + P_4 = P_2 + P_4 = P_5, \\ 6P_4 &= 5P_4 + P_4 = P_5 + P_4 = \mathcal{O}. \end{aligned}$$

Hence, the order of $P_1 \in E'(\mathbb{F}_5)$ is 2, and the order of $P_4 \in E'(\mathbb{F}_5)$ is 6. Also we see Lagrange's theorem holds.

Algorithm 1 (Binary method) and Algorithm 2 (Montgomery reduction) are algorithms for computing a scalar multiplication. Let n be a k-bit integer, and

$$n = (n_{k-1}, n_{k-2}, \cdots, n_0)_2$$

be the binary representation of n. Then, Algorithm 1 takes k doubling formulas and $k/2$ addition formulas on average, and Algorithm 2 takes k doubling formulas and k addition formulas.

```
Algorithm 1 (Binary method)
Input: $P \in E(\mathbb{F}_p), n = (n_{k-1}n_{k-2}\cdots n_0)_2 \in \mathbb{N}$
Output: $nP \in E(\mathbb{F}_p)$
$Q \leftarrow P$
for $i = k - 2$ down to $0$
    $Q \leftarrow 2Q$
    if $n_i = 1$ then $Q \leftarrow Q + P$
end for
return $Q$
```

```
Algorithm 2 (Montgomery ladder)
Input: $P \in E(\mathbb{F}_p), n = (n_{k-1}n_{k-2}\cdots n_0)_2 \in \mathbb{N}$
Output: $nP \in E(\mathbb{F}_p)$
$Q_0 \leftarrow \mathcal{O},\ Q_1 \leftarrow P$
for $i = k - 1$ down to $0$
    if $n_i = 0$ then $Q_1 \leftarrow Q_0 + Q_1,\ Q_0 \leftarrow 2Q_0$
    if $n_i = 1$ then $Q_0 \leftarrow Q_0 + Q_1,\ Q_1 \leftarrow 2Q_1$
end for
return $Q_0$
```

2.4 Coordinate System

Addition formula (8) and doubling formula (9) needs a division, whose cost is high, to compute λ. Hence, we would like to compute both formulas without division.

For a point (x, y) in the xy coordinate system, another coordinate system that uses X, Y, Z satisfying $x = X/Z,\ y = Y/Z$ to represent (x, y) as (X, Y, Z) is called the projective coordinate system. In the projective coordinate system, addition formula and doubling formula can be computed without division. Hence, the projective coordinate system (or its variants) is generally used in cryptographic implementations. Algorithms 3 and 4 are addition and doubling formulas for Weierstrass form in the projective coordinate system. In these algorithms, Roman face means temporal variables.

Algorithm 3
(Addition for Weierstrass on projective coordinates)

Input: $P = (X_1, Y_1, Z_1), Q = (X_2, Y_2, Z_2) \in E(\mathbb{F}_p)$
Output: $P + Q = (X_3, Y_3, Z_3) \in E(\mathbb{F}_p)$

```
$\mathtt{Y1Z2} \leftarrow Y_1 \cdot Z_2$
$\mathtt{X1Z2} \leftarrow X_1 \cdot Z_2$
$\mathtt{Z1Z2} \leftarrow Z_1 \cdot Z_2$
$\mathtt{u} \leftarrow Y_2 \cdot Z_1 - \mathtt{Y1Z2}$
$\mathtt{uu} \leftarrow \mathtt{u}^2$
$\mathtt{v} \leftarrow X_2 \cdot Z_1 - \mathtt{X1Z2}$
$\mathtt{vv} \leftarrow \mathtt{v}^2$
$\mathtt{vvv} \leftarrow \mathtt{v} \cdot \mathtt{vv}$
$\mathtt{R} \leftarrow \mathtt{vv} \cdot \mathtt{X1Z2}$
$\mathtt{A} \leftarrow \mathtt{uu} \cdot \mathtt{Z1Z2} - \mathtt{vvv} - (\mathtt{R} + \mathtt{R})$
$X_3 \leftarrow \mathtt{v} \cdot \mathtt{A}$
$Y_3 \leftarrow \mathtt{u} \cdot (\mathtt{R} - \mathtt{A}) - \mathtt{vvv} \cdot \mathtt{Y1Z2}$
$Z_3 \leftarrow \mathtt{vvv} \cdot \mathtt{Z1Z2}$
```

Algorithm 4
(Doubling for Weierstrass on projective coordinates)

Input: $P = (X_1, Y_1, Z_1) \in E(\mathbb{F}_p), a$ of Eq.(4)
Output: $2P = (X_3, Y_3, Z_3) \in E(\mathbb{F}_p)$

```
$\mathtt{XX} \leftarrow X_1^2$
$\mathtt{ZZ} \leftarrow Z_1^2$
$\mathtt{w} \leftarrow a \cdot \mathtt{ZZ} + 3\mathtt{XX}$
$\mathtt{s} \leftarrow 2(Y1 \cdot Z1)$
$\mathtt{ss} \leftarrow \mathtt{s}^2$
$\mathtt{sss} \leftarrow \mathtt{s} \cdot \mathtt{ss}$
$\mathtt{R} \leftarrow Y1 \cdot \mathtt{s}$
$\mathtt{RR} \leftarrow \mathtt{R}^2$
$\mathtt{B} \leftarrow (\mathtt{X1} + \mathtt{R})^2 - \mathtt{XX} - \mathtt{RR}$
$\mathtt{h} \leftarrow \mathtt{w}^2 - 2\mathtt{B}$
$X_3 \leftarrow \mathtt{h} \cdot \mathtt{s}$
$Y_3 \leftarrow \mathtt{w} \cdot (\mathtt{B} - \mathtt{h}) - 2\mathtt{RR}$
$Z_3 \leftarrow \mathtt{sss}$
```

Let $P = (X, Y, Z)$ be a point on a Montgomery curve in the projective coordinate system. Then, X and Z coordinates of nP can be computed from X and Z coordinates of P using Algorithm 2 (Montgomery ladder) [9]. In this case, Algorithms 5 and 6 are used. For details for these algorithms, refer to [4]. By the way, although omitted, reduction $\bmod p$ is required at every step of Algorithms 3, 4, 5, and 6. For details for them, refer to [2].

Table 2: The cost of Algorithms 3,4,5,6

	Cost	Purpose
Algorithm 3	$14M + 7add$	Addition for Weierstrass
Algorithm 4	$11M + M_a + 12add$	Doubling for Weierstrass
Algorithm 5	$6M + 6add$	Addition for Montgomery
Algorithm 6	$4M + M_{A'} + 4add$	Doubling for Montgomery

Table 3: Cost of scalar multiplication of nP, where n is k-bit

Used curve	Used algorithms	Cost
Weierstrass	Algorithms 1,3,4	$18kM + kM_a + 18k\ add$
Montgomery	Algorithms 2,5,6	$10kM + kM_{A'} + 10k\ add$

We will estimate the cost of theses algorithms in the number of modular multiplications because the cost of modular addition/subtraction is much smaller than one of modular multiplication. Let M, M_a, $M_{A'}$, and add denote a modular multiplication in $\mathbb{F}_p$, a modular multiplication in $\mathbb{F}_p$ with a constant a, a modular multiplication in $\mathbb{F}_p$ with a constant A', and a modular addition/subtraction in $\mathbb{F}_p$, respectively, where a is of (4), and $A' = (A + 2)/4$ for A of (5). Then, the cost of the algorithms are given by Table 2. Note that $M_{A'}$ is for $A' \cdot \mathtt{C}$ at step 7 of Algorithm 6. Also, the cost of scalar multiplication are given by Table 3. We see that the cost of a scalar multiplication on a Montgomery curve is less than one on Weierstrass form.

2.5 Secure Elliptic Curve

The security of ECCs including digital signature depends on the maximum prime factor l of the order $L = \#E(\mathbb{F}_p)$, not the size of p [13]. Attack time against ECCs is roughly proportional to $\sqrt{l}$ and then the larger l is, the more secure ECCs are. Therefore, we have to select elliptic curve E such that

$$L = \#E(\mathbb{F}_p) \text{ has a big prime factor} \tag{11}$$

for ECCs. Also, it is desirable that

$$\begin{gathered} 2p + 2 - L \text{ that is the order of the twist of } E \\ \text{has a big prime factor} \end{gathered} \tag{12}$$

according to [3]. Moreover, we have to select E such that

$$L \neq p, p \pm 1 \tag{13}$$

according to [6], [14].

2.6 Curve25519

Curve25519 is a Montgomery curve

$$E_{25519} : y^2 = x^3 + 486662x^2 + x$$

Algorithm 5
(Addition for Montgomery on XZ coordinates)

```
Input: $Q_1 - Q_0 = (X_1, Z_1), Q_0 = (X_2, Z_2), Q_1 = (X_3, Z_3)$
Output: $Q_0 + Q_1 = (X_4, Z_4) \in E(\mathbb{F}_p)$
A ← $X_2 + Z_2$
B ← $X_2 - Z_2$
C ← $X_3 + Z_3$
D ← $X_3 - Z_3$
DA ← D · A
CB ← C · B
$X_4 \leftarrow Z_1 \cdot (\mathtt{DA} + \mathtt{CB})^2$
$Z_4 \leftarrow X_1 \cdot (\mathtt{DA} - \mathtt{CB})^2$
```

Algorithm 6
(Doubling for Montgomery on XZ coordinates)

```
Input : $R = (X_1, Z_1), A' = (A + 2)/4$ for $A$ in Eq.(7)
Output: $2R = (X_4, Z_4) \in E(\mathbb{F}_p)$
A ← $X_1 + Z_1$
AA ← $\mathtt{A}^2$
B ← $X_1 - Z_1$
BB ← $\mathtt{B}^2$
C ← AA − BB
$X_4 \leftarrow \mathtt{AA} \cdot \mathtt{BB}$
$Z_4 \leftarrow C \cdot (\mathtt{BB} + A' \cdot \mathtt{C})$
```

with $p = 2^{255} - 19$ [1]. The order $L = \#E_{25519}(\mathbb{F}_p)$ is

$$L = 2^2 \cdot l,$$
$$l = 2^{252} + 27742317777372353535851937790883648493,$$

where l is a 253-bit prime. Curve25519 meets the security requirement in Sec.2.5.

Curve 25519 has been applied in many cryptographic libraries such as NaCl [10], and Curve 25519 was added to Special Publication 800-186, which specifies the approved elliptic curve used by the US federal government by NIST in 2017.

3 ECDSA

ECDSA is a digital signature using an elliptic curve. ECDSA consists of *system parameter* held by all users, *key generation* for generating each user's (private key, public key), *signature generation* for generating a signature using a user A's secret key, and *signature verification* for verifying the signature using A's public key.

System parameter
A sufficiently large (e.g. 256-bit) prime p, an elliptic curve E such that $E(\mathbb{F}_p)$ meets the security requirements (11), (12), and (13), and a base point $G \in E(\mathbb{F}_p)$ of which the order is l are selected. Also a hash function $H : \{0,1\}^* \to \{0, 1, 2, ..., l-1\}$ is selected. (p, l, E, G, H) is the system parameter.

Key generation
User A choose $s \in [1, l-1]$ at random, and computes $Y = sG$ (scalar multiplication) in $E(\mathbb{F}_p)$. Then, s and Y are A's private key and public key, respectively.

Signature generation
User A generates a signature of a message $m \in \{0,1\}^*$ as follows.

1. Computing $m' = H(m)$.
2. Choosing $r \in [1, l-1]$ at random, and compute
$$U = \underbrace{rG}_{\text{scalar mul. on } E(\mathbb{F}_p)} = (u_x, u_y),$$
$$u = u_x \bmod l.$$
3. Using the secret key s to compute
$$v = r^{-1}(m' + su) \bmod l.$$
4. The pair (u, v) is the signature of m.

Signature verification
A recipient of the message m with signature (u, v) verifies the signature as follows.

1. Computing $m' = H(m)$.
2. Computing $d = v^{-1} \bmod l$.
3. Computing $U' = \underbrace{(m'd)G}_{\text{scalar mul. on } E(\mathbb{F}_p)} + \underbrace{(ud)Y}_{\text{scalar mul. on } E(\mathbb{F}_p)}$.
4. Computing $u' =$ (the x coordinate of U') $\bmod l$.
5. If $u = u'$ then the signature is accepted, and if $u \neq u'$ then it is rejected.

Thus, the dominant processes of ECDSA is scalar multiplications in $E(\mathbb{F}_p)$ [5]. Signature generation of ECDSA takes one scalar multiplication and signature verification of ECDSA takes two scalar multiplications. Therefore, we see that

in order to speed up processes of ECDSA,
it is important to speed up scalar multiplication.

As seen Table 3, using not Weierstrass form but Montgomery curve reduces the number of modular multiplications required for a scalar multiplication.

4 MODULAR REDUCTION

In order to speed up ECCs including ECDSA, it is important not only to reduce the number of modular multiplications but also to reduce the cost of one modular multiplication. This section introduces efficient reduction methods.

4.1 Montgomery Reduction

The Montgomery reduction (Algorithm 7) [8] is a method for efficiently calculating $X \bmod N$ for general odd number N and X given in Montgomery representation[6].

[5] The dominant process of not only ECDSA but also all ECCs is scalar multiplications.

[6] For Montgomery representation, refer to [8] or [4].

Algorithm 7 (Montgomery Reduction)

Input: odd number N of n bits, $R = 2^n$,
$N' = (-N^{-1}) \bmod R$, natural number $u < RN$
Output: $u \bmod N$ in Montgomery representation

```
1. t ← $u$
2. k ← t$N' \bmod R$
3. t ← t + k$N$
4. t ← $t/R$
5. if t ≥ $N$ then t ← t − $N$
6. return t
```

4.2 Reduction modulo Pseudo Mersenne Prime

When a prime p is written as

$$p = 2^n - k, \ k < 2^{n/2},$$

it is called pseudo Mersenne prime. For pseudo Mersenne prime p, reduction $\bmod p$ can be computed at high speed using by Algorithm 8 [7]. Notice that $u/2^n$ in step 1 and $v/2^n$ in step 3 are integer divisions and then they are performed by shift operations.

Algorithm 8 (Reduction mod pseudo Mersenne prime)

Input: prime $p = 2^n - k$ $(k < 2^{n/2})$,
integer $0 \le u \le (p-1)^2$
Output: $u \bmod p$

```
1. u0 ← $u \bmod 2^n$, u1 ← $u/2^n$
2. v ← u1 · $k$ + u0
3. v0 ← v $\bmod 2^n$, v1 ← v$/2^n$
4. $w$ ← v1 · $k$ + v0
5. if $w \ge p$ then $w \leftarrow w - p$
6. return $w$
```

5 CONTRIBUTIONS

5.1 Program to Search Elliptic Curve Suitable for ECDSA

The purpose of this paper is to make a program to search for elliptic curves that is secure and suitable for high-speed implementation of ECDSA (especially by hardware implementation), and to give examples of such an elliptic curves. Specifically, we will search curves that meet the following requirements.

Elliptic Curve Requirements to Search

1. According to Table 3, scalar multiplication in Montgomery curve takes fewer modular multiplications than scalar multiplication in Weierstrass form. Hence, Montgomery curve is selected.

2. According to Table 3 again, scalar multiplication in Montgomery curve with $A' = 1, 2, 3, 4, 5$, that is, the coefficient $A = 2, 6, 10, 14, 18$, requires fewer modular multiplications than other As [7]. Therefore, we take $A = 2, 6, 10, 14, 18$.

3. Prime p is typical 256-bit. Moreover, p is a pseudo Mersenne prime $p = 2^n - k$ because of efficient reduction $\bmod p$. For convenience of execution time, set the range of k to $k \le 2^{20}$.

4. To meet the security requirement (11), the order $L = \#E(\mathbb{F}_p)$ is as $L = 4l, 8l, 16l$, where l is a prime. Notice there is a point $P \in E(\mathbb{F}_p)$ whose order is l.

5. To meet the security requirement (12), $L' = 2p + 2 - L$ is as $L' = 4l', 8l', 16l'$, where l' is a prime.

6. To meet the security requirement (13), curves such as $L = p, p \pm 1$ is removed.

7. L is written as $L = 2^n - k'$, $k' < 2^{n/2}$. Then, a reduction $\bmod l$ is computed by the algorithm proposed in Sec.5.2, which is as efficient as reduction mod a pseudo Mersenne prime.

Note Curve25519 also meets the requirements 1, 4, 5, and 7, and Curve25519 adopts not 256-bit prime but 254-bit for a prime field. Curce25519 does not consider the requirements 2 and 7. Also refer to Sec.5.3.

The authors made a program as Fig. 2 in PARI/GP to search elliptic curves meeting the requirements. The program is straightforward and then it may be easy for some readers to make a similar program. But, giving the program makes all readers (especially PARI/GP users) generate good curves. Notice that the program output only a prime p, a coefficient A of Montgomery curve, the order L of $E(\mathbb{F}_p)$, and the order l of a base point. At the moment, it is necessary to manually find another coefficient B of Montgomery curve, generate a base point whose order is l, and check $L \ne p, p-1$.

This program is briefly explained. The line

```
e=ellinit([0,A,0,1,0]);
```

sets (Montgomery) elliptic curve $E : y^2 = x^3 + Ax^2 + x$ to `e`. The function `ellap(e,p)` outputs the trace of $E(\mathbb{F}_p)$. Thus, `Num1` is the order of $E(\mathbb{F}_p)$, and `Num2` is the order of the twist of $E(\mathbb{F}_p)$. `isprime` is a prime decision function. `write` is an output function to text.

By this program, the following elliptic curves are found.

1. $p = 2^{256} - 58097$,
   $E_{S1} : 638y^2 = x^3 + 10x^2 + x$,
   base point $P = (11, 2)$,
   $L = 2^{256} - k'$, where k' is 125-bit integer
   $k' = 25181363380428710453079967399017869328$,
   $l = L/16$, which is 252-bit prime.

[7] Selecting an elliptic curve with appropriate coefficients to reduce high cost multiplication into low cost addition is one of fast implementation techniques of ECCs [4, Sec. 13.2.1c]. For Montgomery curve, if $A' = 1$ then the multiplication $A' \cdot$ `C` at step 7 of Algorithm 4 is free. If $A' = 2$ then the multiplication is performed by an addition `C` + `C`. As well, if $A' = 3$ or 4 then the multiplication is performed by two additions.

```
\\Checking pseudo Mersenne prime
check_mer(p)={
  local(n,c,k);
  n=0;
  c=0;
  while(c==0,
    n++;
    if(2^(n-1)<=p && p<2^n,c=1);
  );
  if(2^n-p < sqrt(2^n),
    k=floor(log(2^n-p)/log(2)+1);
    return([n,k]),
    return(0));
}

\\Main program
{
  count=0;
  for(a=0,4,
    A=4*a+2;
    e=ellinit([0,A,0,1,0]);
    for(k=1,2^20,
      print([a,k,count]);
      p=2^256-k;
      if(isprime(p)==1,
        t=ellap(e,p);
        num1=p+1-t;
        num2=p+1+t;
        Num1=num1;
        Num2=num2;
        check1=0;
        check2=0;
        if(num1%2==0,num1=num1/2;check1=1);
        if(num1%2==0,num1=num1/2;check1=2);
        if(num1%2==0,num1=num1/2;check1=3);
        if(num1%2==0,num1=num1/2;check1=4);
        if(num2%2==0,num2=num2/2;check2=1);
        if(num2%2==0,num2=num2/2;check2=2);
        if(num2%2==0,num2=num2/2;check2=3);
        if(num2%2==0,num2=num2/2;check2=4);
        isprime_num1=isprime(num1);
        isprime_num2=isprime(num2);
        if(t!=0 && isprime_num1==1
                && isprime_num2==1
                && (check_mer(Num1)!=0
                || check_mer(Num2)!=0),
          if(check_mer(Num1)!=0,
            count++;
            write("ijis.txt",k","A","Num1",
            "num1);
          );
          if(check_mer(Num2)!=0,
            count++;
            print("A="A);
            write("ijis.txt",k","A","Num2",
            "num2);
          );
        );
      );
    );
  );
}
```

Figure 2: Proposed program to find elliptic curves suitable for ECDSA

2. $p = 2^{256} - 507225$,
   $E_{S2} : 82y^2 = x^3 + 18x^2 + x$,
   base point $P = (2, 1)$,
   $L = 2^{256} - k'$, where k' is 127-bit integer
   $k' = 134184981501621384111934924743103436264$,
   $l = L/8$, which is 253-bit prime.

3. $p = 2^{256} - 979077$,
   $E_{S3} : 3805y^2 = x^3 + 18x^2 + x$,
   base point $P = (20, 2)$,
   $L = 2^{256} - k'$, where k' is 126-bit integer
   $k' = 67240641251824776802983670794157366424$,
   $l = L/8$, which is 253-bit prime.

For the convenience of time, the authors set the search range to $k < 2^{20}$, however, if the search range is expanded, more appropriate elliptic curve may be found.

5.2 Proposed Modular Reduction

This section proposes an algorithm (Algorithm 9) [8] similar to Algorithm 8 for computing a reduction $\bmod\ l$ for a prime l such that $L = 2^m l$ is written as $L = 2^n - k,\ k < 2^{n/2}$.

Notice $\mathtt{v}$, $\mathtt{v0}$ and k are multiples of 2^m. Thus, $\mathtt{w}$ is also a multiple of 2^m. As well,

$$\mathtt{x} \text{ is a multiple of } 2^m. \tag{14}$$

From step 2 to 7 is same as Algorithm 8 and then we see

$$\mathtt{x} = 2^m u \bmod\ 2^m l. \tag{15}$$

By (14) and (15), we see $\mathtt{x}/2^m = u \bmod\ l$.

Note that $\mathtt{v}/2^n$ in step 2, $\mathtt{w}/2^n$ in step 4, and $\mathtt{x}/2^n$ in step 7 are integer divisions and then they are performed by shift operations, and $2^m u$ in step 1 is also performed by a shift operation.

```
Proposed Algorithm 9
Input : integer $l$ such that $2^m l = 2^n - k\ (k < 2^{n/2})$,
        integer $0 \le u \le (l-1)^2$
Output : $u \bmod\ l$
1. $\mathtt{v} \leftarrow 2^m u$
2. $\mathtt{v0} \leftarrow \mathtt{v} \bmod\ 2^n,\ \mathtt{v1} \leftarrow \mathtt{v}/2^n$
3. $\mathtt{w} \leftarrow \mathtt{v1} \cdot k + \mathtt{v0}$
4. $\mathtt{w0} \leftarrow \mathtt{w} \bmod\ 2^n,\ \mathtt{w1} \leftarrow \mathtt{w}/2^n$
5. $\mathtt{x} \leftarrow \mathtt{w1} \cdot k + \mathtt{w0}$
6. if $\mathtt{x} \ge 2^m l$ then $\mathtt{x} \leftarrow \mathtt{x} - 2^m l$
7. $y \leftarrow \mathtt{x}/2^m$
8. return $y$
```

5.3 Searched Elliptic Curves v.s. Curve25519

Elliptic curves searched in this paper and Curve 25519 meet the security requirement in Sec.2.5. Both of them adopt pseudo Mersenne prime and then reductions $\bmod\ p$ for them are efficiently computed. However, computation of reduction $\bmod\ p$ may be more efficient for Curve 25519 on CPU with small words because k of $p = 2^n - k$ is smaller,

[8] Although the authors do not know whether Algorithm 9 is already known, it may be already known because Algorithm 9 is almost same as Algorithm 8.

The coefficient of searched elliptic curves are $A = 10, 18$ $(A' = 3, 5)$, on the other hand, one of Curve 25519 is $A = 486662$ $(A' = 121666)$. Notice that when $A' = 2$, a product with A' is reduced to an addition. Hence, the cost of a scalar multiplication with Curve25519 is $10kM + kM_{A'} + 10k\, add$, and one with the searched curve is $10kM + 11k\, add$ by Table 3. In general add is smaller than $M_{A'}$.

A reduction $\bmod\ l$ can be computed with Algorithm 9 for searched curves, on the other hand, it cannot be for Curve25519. Therefore, ECDSA adopting searched curves is expected to be faster and implemented more efficiently (when it is implemented hardware or CPU with small words) compared with the ECDSA adopting Curve25519.

6 CONCLUSION

This paper searched three elliptic curves suitable for ECDSA. In these curves, not only the reduction $\bmod\ p$ but also the reduction $\bmod\ l$ can be computed at high speed, where p is of $\mathbb{F}_p$ and l is the order of a base point. and doubling is faster because of a coefficient of curves $A = 10$ or 18. ECDSA adopting the searched curves has the same security as ECDSA adopting Curve 25519, and it can process faster.

The authors would like to evaluate implementation results as future work. Also, they would like to extend the search range of the suggested program in Fig. 2 and to execute it.

REFERENCES

[1] D. J. Bernstein, “Curve25519: New Diffie-Hellman speed records,” PKC 2006, LNCS 3958, pp. 207–228, Springer (2006).

[2] D. J. Bernstein and T. Lange, Explicit-formulas database, https://hyperelliptic.org/EFD/.

[3] D. J. Bernstein and T. Lange, Safecurves: Choosing safe curves for elliptic-curve cryptography, https://safecurves.cr.yp.to.

[4] H. Cohen and G. Frey, editors, Handbook of Elliptic and Hyperelliptic Curve Cryptography, Chapman and Hall/CRC (2005).

[5] ETSI, Etsi ts 103 097 v1.1.1 intelligent transport systems (its); security; security header and certificate formats (2013).

[6] G. Frey and H.-G. Rück, “A remark concerning m-divisibility and the discrete logarithm in the divisor class group of curves,” Matheematics of Computation, Vol. 62, No. 206, pp. 865–874 (1994).

[7] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, Handbook of Applied Cryptography, CRC Press (1997).

[8] P. L. Montgomery, “Modular multiplication without trial division,” Matheematics of Computation, Vol. 44, No. 170, pp. 519–521 (1985).

[9] P. L. Montgomery, “Speeding the pollard and elliptic curve methods of factorization,” Matheematics of Computation, Vol. 48, No. 177, pp. 243–264 (1987).

[10] Nacl: Networking and cryptography library, https://pari.math.u-bordeaux.fr.

[11] NIST, Nist: Fips 186-2 digital signature standard, https://csrc.nist.gov/csrc/media/publications/fips /186/3/archive/ 2009-06-25/documents/fips_186-3.pdf.

[12] PARI/GP, https://pari.math.u-bordeaux. fr.

[13] S. Pohlig and M. Hellman, “An improved algorithm for computing logarithms over $GF(p)$ and its cryptographic significance,”IEEE Transactions on Information Theory, Vol. 24, No. 1, pp. 106–110 (1978).

[14] H.-G. Rück, “On the discrete logarithm in the divisor class group of curves,” Matheematics of Computation, Vol. 68, No. 226, pp. 805–806 (1999).

[15] J. H. Silverman, The arithmetic of elliptic curves, Springer-Verlag New York (1985).

[16] J. Tate and J. H. Silverman, Rational points on elliptic curves. Springer-Verlag (1992).

(Received October 21, 2018)
(Revised April 8, 2019)

Masaaki Shirase Masaaki Shirase received the M.S. and Ph.D. degrees in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2003 and 2006, respectively. He is currently an Associate Professor in the School of Systems Information Science at Future University Hakodate. His research interests are algorithm and implementation of cryptography.

Submission Guidance

About IJIS

International Journal of Informatics Society (ISSN 1883-4566) is published in one volume of three issues a year. One should be a member of Informatics Society for the submission of the article at least. A submission article is reviewed at least two reviewer. The online version of the journal is available at the following site: http://www.infsoc.org.

Aims and Scope of Informatics Society

The evolution of informatics heralds a new information society. It provides more convenience to our life. Informatics and technologies have been integrated by various fields. For example, mathematics, linguistics, logics, engineering, and new fields will join it. Especially, we are continuing to maintain an awareness of informatics and communication convergence. Informatics Society is the organization that tries to develop informatics and technologies with this convergence. International Journal of Informatics Society (IJIS) is the journal of Informatics Society.

Areas of interest include, but are not limited to:

Internet of Things (IoT)
Smart Cities, Communities, and Spaces
Big Data, Artificial Intelligence, and Data Science
Network Systems and Protocols
Computer Supported Cooperative Work and Groupware
Security and Privacy in Information Systems
Intelligent Transportation System
Distributed Computing
Multi-media communication
Information systems
Mobile computing
Ubiquitous computing

Instruction to Authors

For detailed instructions please refer to the Authors Corner on our Web site, http://www.infsoc.org/.

Submission of manuscripts: There is no limitation of page count as full papers, each of which will be subject to a full review process. An electronic, PDF-based submission of papers is mandatory. Download and use the LaTeX2e or Microsoft Word sample IJIS formats.

http://www.infsoc.org/IJIS-Format.pdf

LaTeX2e

LaTeX2e files (ZIP) http://www.infsoc.org/template_IJIS.zip

Microsoft Word^{TM}

Sample document   http://www.infsoc.org/sample_IJIS.doc

Please send the PDF file of your paper to secretariat@infsoc.org with the following information:

Title, Author: Name (Affiliation), Name (Affiliation), Corresponding Author. Address, Tel, Fax, E-mail:

Copyright

For all copying, reprint, or republication permission, write to: Copyrights and Permissions Department, Informatics Society, secretariat@infsoc.org.

Publisher

Address: Informatics Laboratory, 3-41 Tsujimachi, Kitaku, Nagoya 462-0032, Japan
E-mail: secretariat@infsoc.org

CONTENTS