IIEEJ Transactions on

Image Electronics and Visual Computing

Special Issue on Extended Papers Presented in IEVC2019 Part II

Special Issue on CG & Image Processing Technologies for Automation, Labor Saving and Empowerment

Vol. 8, No. 2 2020

The Institute of Image Electronics Engineers of Japan

Editorial Committee of IIEEJ

Contact Information

The Institute of Image Electronics Engineers of Japan (IIEEJ)
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan
Tel : +81-3-5615-2893　Fax : +81-3-5615-2894
E-mail : hensyu@iieej.org
http://www.iieej.org/ (in Japanese)
http://www.iieej.org/en/ (in English)
http://www.facebook.com/IIEEJ (in Japanese)
http://www.facebook.com/IIEEJ.E (in English)

IIEEJ Transactions on
Image Electronics and Visual Computing
Vol.8 No.2 December 2020

CONTENTS

Published two times a year by the Institute of Image Electronics Engineers of Japan（IIEEJ）
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan
Tel:+81-5615-2893 Fax:+81-5615-2894 E-mail:hensyu@iieej.org http://www.iieej.org/

Upon the Special Issue on
Extended Papers Presented in IEVC2019 PartⅡ

Editor: Prof. Naoki KOBAYASHI
Saitama Medical University

The Institute of Image Electronics Engineering of Japan (IIEEJ) regularly holds International academic events named “Image Electronics and Visual Computing (IEVC)” since 2007, on every two or two and half years. The 6th International Conference on Image Electronics and Visual Computing (IEVC2019) was held in Bali, Indonesia on August 21-24, 2019. From this time, the name of this event was changed from International Workshop to International Conference, to promote the event more worldwide and more attractive for speakers and attendees. The conference was successfully held with 109 presentations and 165 participants (including 39 foreigners from more than 10 countries).

There were two paper categories in IEVC2019: general paper and late breaking paper (LBP), and in general paper, there were two tracks: Journal track (JT) and Conference track (CT). In IEVC2019, 33 JT papers, 52 CT papers and 24 LBP were submitted.

Journal track is a newly introduced one and has the advantage to be able to publish the paper on the journal (IIEEJ Trans. on IEVC) in the “Special Issue on Journal Track in IEVC2019” on December 2019 issue, by submitting full paper version (8 pages) together with conference paper version to be peer-reviewed in advance. Actually seven papers were adopted in the “Special Issue on Journal Track in IEVC2019” published on December 2019.

The special issue on “Extended Papers Presented in IEVC2019” to be published on June 2020 was openly called for all paper categories in IEVC2019, and five papers, including the JT papers not to be in time for the publication schedule of December 2019 issue, were adopted in June 2020 issue. In this Part Ⅱ of the “Extended Papers Presented in IEVC2019”, two general papers and one system development paper have passed the review process to be in time for the publishing schedule.

Finally, I would like to give great thanks to all the reviewers and editors for their time and efforts towards improving the quality of papers. I would also like to express my deepest appreciation to the editorial committee members of IIEEJ and the staff at IIEEJ office for various kinds of support.

IIEEJ **Paper** *-- Special Issue on Extended Papers Presented in IEVC2019 Part II --*

A Screen Shake Determination Method Based on 2D Motion Histogram Analyses by Using Group Transition and Maximum Group Ratio in Gaze Areas

Mei KODAMA†(*Fellow*)

†Hiroshima University

<**Summary**> If videos have screen shake information, it is one of the important issues to prevent viewers from VIMS (visually induced motion sickness). So far there are two major approaches to prevent them. First approach is a visually sickness information extraction method by using bio-metric information, and since it is necessary to extract it after the physical condition becomes poor, the processing delay is inevitable. On the other hand, second approach is a motion information extraction method by using image processing in videos. However, it is reported that the processing time becomes longer when the detailed motion such as global motion estimation is analyzed. Thus, a screen shake determination method, which had used the block matching method as a simple motion analysis, motion direction histograms, and this similarity, had been proposed. However, there is still the problem that the accuracy of detecting screen shake decreases, when the amount of screen shake is small in the conventional method. It cannot extract the direction information of screen shake. To solve the problems, this paper proposes a novel screen shake determination method based on 2D motion histogram analyses. In particular, there are three features: the use of gaze areas, the group transition analysis of maximum frequency, and the maximum group ratio analysis in this method. A new evaluation value Ev is defined in consideration of both the accuracy of no-swing and pseudo swing images. Simulation experiments show that the Ev in the proposed method is at most 4.02 smaller than that in the conventional method for the small screen shake. Therefore, it is revealed that the proposed method improves the accuracy of detecting the small screen shake in the conventional method and can extract the direction information of screen shake. Furthermore, it is shown that it solves the problem of setting the threshold of the histogram correlation. An adaptive method for each gaze area, and an adaptive method for each number of directions and divisions in motion vector space will deserve for consideration, but they are left for further studies.

Keywords: visually induced motion sickness (VIMS), screen shake, motion vector, histogram, group transition, group ratio

1. Introduction

Nowadays, since the number of videos viewed on the Internet is increasing, and there are many users who view them, there is the problem of video quality associated with visually induced motion sickness (hereinafter, VIMS)[1)–3)] as a biological safety problem. In addition to 2D images, 3D images, 360-degree images, and virtual space images, this problem had been studied[4)–16)]. In particular, it is one of the important issues to prevent viewers from VIMS. For example, if videos include screen shake scenes taken by amateurs, it is necessary to take measures for viewing. Specifically, it is necessary to take measures such as displaying a subtitle of warning or filtering the scene when screen shake scenes are detected, or deleting the videos included the screen shake scenes in advance. So far there are two major approaches to prevent them. The first approach[6)–9)] is a visually sickness information extraction method by using bio-metric information directly, and since it is necessary to extract it after the physical condition becomes poor, the processing delay is inevitable. On the other hand, the second approach is a motion information extraction method by using image processing in videos. However, it was reported that the processing time became longer when the detailed motion such as a global motion estimation was analyzed[15)–18)]. When obtaining the global motion vectors[17),18)] related to the screen shake, spatial image information is needed, and the processing delay occurs. Even if the high-speed calculating method of vector data are employed, a huge amount of calculation is required, compared with the general motion calculation method such

as the block matching. Thus, a screen shake determination method, which had used the block matching method as a simple motion analysis, motion direction histograms, and these similarities, had been proposed[15),16)]. This basic idea using motion histograms is an extension of the previous video identification systems based on a motion histogram analysis[19),20)] and it supports the real-time processing. However, there are still the problem that the accuracy of detecting screen shake decreases and the problem of how to set common parameters for various sequences, when the amount of screen shake is small in the conventional method. Then, to solve the problems, this paper proposes a novel screen shake determination method based on 2D motion histogram analyses. In particular, there are three features: the use of gaze areas, the group transition analysis of maximum frequency, and the maximum group ratio analysis in this method.

The previous studies are described in section 2, and the details of the proposed method are shown in section 3. The determination accuracy is evaluated by comparing the conventional method with the proposed method in the simulation experiments and the results are considered in section 4. Finally, this paper concludes in section 5.

2. Previous Study

A motion extraction method with the global motion compensation had been studied to analyze the detailed camera motions[17),18)]. However, the real-time processing is extremely difficult in these methods due to a lot of motion parameters in camera operations, e.g., pan, tilt, and so on. On the other hand, general video processing, e.g., video coding and motion analysis, generally uses a block matching method based on a parallel motion model. We need a simple motion extraction of screen shake specialized for VIMS rather than a motion extraction with high accuracy of detecting objects, and it is also required to adapt to the real-time processing.

By the way, a searching method with spatial information of a query video and a retrieval method with motion vectors of a query video had been studied as a high-speed video searching method[19),20)]. Those retrieval methods use the characteristics of the similarity of motion information. Then, considering the reduction of the processing time, the author proposed a determination method with histograms of motion vectors based on a block matching method as a general motion compensation processing[15),16)]. This method can make good use of the feature of motion vectors based on the principle of the similarity determination. In other words, it can detect the scene corresponding to screen shake by using the temporal correlation of the histograms. Then, he focused on general block matching method for extracting screen shake, and proposed a screen shake determination method to prevent VIMS. It was a histogram-based approach using motion vectors obtained by simple block matching method, which was based on a fixed block size. Since it also used the amount of the motion vectors which were separated from the motion direction histogram, the accuracy was higher than that of a determination method with only the histogram correlation. However, there was a problem that the results were not output as continuous frames. Then, the time extension method, which detected a continuous warning scene from warning candidate frames, was proposed, and it improved the accuracy of detecting VIMS. However, there were still some basic problems that the accuracy was reduced in the small screen shake, and the normal movement of camera, e.g., pan or tilt, was falsely detected as screen shake. For these undetected frames or false detection, an improved time extension method, which had used the histogram correlation with direction only and had considered the start and the end frames as the screen shake scene simply, had been studied[21)]. To simply improve the detection accuracy for VISM, there were the previous determination methods[15),16)] by using Bhattacharyya distance[22)] (hereinafter, BD) of the histograms with only the direction and the frequency change for each direction. Then, among the previous methods, this extension method with the highest accuracy is employed as the conventional method to make it easier to evaluate the accuracy of detecting screen shake under the same conditions in this study. The principle of them is simply described, the procedure of the conventional method is shown, and the problems are organized.

2.1 Principle of screen shake detection

The principle of screen shake detection had been proposed[15),16)]. In the moving pictures, which do not include scene change, since there is continuity of motion between frames due to a moving subject or a movement of a camera as features of video scene, the histogram of motion vectors in frame number $m-1$ (f_{m-1}) is generally similar to that in frame number m (f_m). On the other hand, in the moving pictures, which include scene change or screen shake information (e.g., hand shake, or shake caused by photographing environment: in a vehicle, such as car, bus, and train, and in a stadium), it is

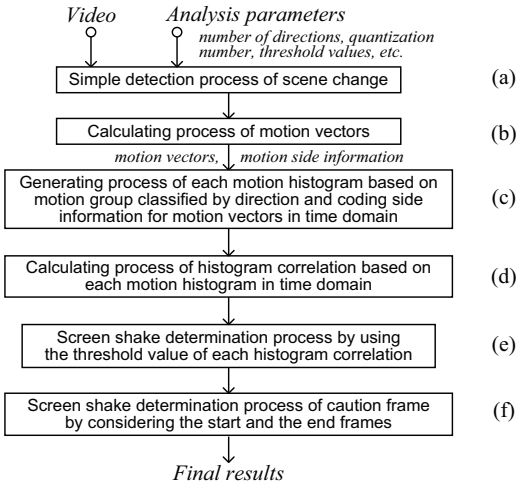


Fig. 1 Processing flow of the conventional method by using direction information

considered that the histogram in f_m is not similar to the histogram in f_{m-1}, when screen shake occurs due to external factors. Hence, the author focuses on the change of the histograms in motion vectors in the conventional method, and BD is employed as the degree of the similarity of the histogram.

2.2 Outline of the conventional method

Figure 1 shows the outline of the conventional method[21], and the procedure is shown as follows:

(a) Simple detection process is applied in order to extract a scene change frame. For example, the time correlation value, or the difference of the luminance histogram between frames can be used.

(b) Motion vectors based on a block matching method are calculated, macro block (MB) of vector detection unit in a simple method is defined as 16×16 pel and the matching search range is defined as ± 32 pel. The motion side information such as motion vectors and undetected motion is output.

(c) Each motion group histogram is generated for each direction and the motion type (coding side information) by using motion vectors for the whole screen in time domain. Here, the motion information group is defined as a group of direction (GD) when the motion vector of which norm is greater than 0 can be calculated. As the other cases in a block matching method, there are two states of motion estimation: 'no motion' information (NMI) and 'not detected' information (NDI). Two groups are defined as a group of NMI (GNMI) and a group of NDI (GNDI), respectively. A motion histogram consists of GD, GNMI, and GNDI.

(d) Each histogram correlation based on each motion histogram in time domain is calculated. Histogram correlation with BD is employed.

(e) The screen shake is determined by using the threshold value of each histogram correlation as the feature of the similarity for each frame.

(f) Finally, the caution frame with screen shake is determined by using the start and the end points extracted as candidate results in the screen shake scene. The number of frames between the start and the end frames is determined by the threshold value Th_{frame}. Thereafter, the processing of steps (c) to (f) is repeated for all input frames, and the final results are sequentially obtained.

Here, the first method consisting of steps (a) to (e) is defined as 'CM1' and the second one consisting of all steps is defined as 'CM2' as two types.

2.3 Problems

The screen shake determination method uses the change of the distribution of motion vectors obtained by each frame. A large screen shake can be extracted in the conventional method by using histograms of motion directions only, whereas a small screen shake cannot be extracted in the conventional method, which uses the determination method with BD and the frequency change additionally[15),16)]. Since the amount of motion vectors cannot be analyzed sufficiently, the shake cannot be extracted in the conventional method[21)]. In all the previous studies[15),16),21)], since both the direction and the motion amount are not used simultaneously, it cannot be found how the scene is shaking. When the input images have the object's motion and the movement of the camera with screen shake, there is also a problem that it is difficult to detect screen shake unless the amount of shake motion is large. If the threshold value of BD is set small in order to increase the determination sensitivity of BD, there is a problem that false detection increases, e.g., the object's motion may be falsely detected as screen shake. More importantly, although it is possible to detect the occurrence of shaking, it is not able to detect in which direction the screen is shaking in the previous studies[15),16),21)]. Then, as described problems, the author focuses on the accuracy improvement in the case where the amount of screen shake is small, or where screen shake is added to general camera movement. To solve the problems and to obtain the screen shake direction, this paper firstly proposes a novel screen shake determination method by using both the histogram of motion directions and the histogram of

quantized amount of motion vectors [21]. They are called quantized histograms of both direction and size of motion vectors (hereinafter, QHs) in this paper.

3. Proposed Method

For the problems described above, the author proposes a novel screen shake determination method by using 2D motion histograms to improve the decrease of accuracy of detecting small screen shake in the conventional method. In particular, the accuracy decreased in the conventional method due to a general camera motion, e.g., pan, tilt. As illustrated by **Fig. 2**, he considers that the motion added to the object motion due to screen shake is extracted. Here, in Fig. 2, the bird area surrounded by the dotted black line indicates the object in f_{m-1}, the bird area surrounded by the dotted purple line indicates the object according to the general motion in f_m, and the black background area surrounded by the red or the light green line indicates the object according to the general motion and screen shake in f_m. Figure 2 (a) shows the object motion according to the general motion, however, the object motion is extracted as the state of Fig. 2 (b) or Fig. 2 (c) due to screen shake, where I_{sx} and I_{sy} denote the horizontal and vertical screen sizes, respectively.

Then, he focuses on how the screen shaking information appears on the motion histograms. Especially, he pays attention to the sensitivity of the histograms. This paper proposes a screen shake determination method by using QHs based on gaze areas in order to improve the sensitivity of detecting histogram change. Since QHs consist of 2D histograms for the direction and the size of motion, this method can analyze not only the change of direction but also the change amount. Therefore, he also proposes a motion analysis method using group transition due to screen shake. So far only the motion direction was considered in the previous histogram, and it was difficult to obtain the direction information of screen shake. However, the proposed method uses the motion analysis with the group transition to the adjacent group in QHs, and it can extract the direction information. Additionally, gaze areas are used in order to increase the motion detection sensitivity. However, since the frequency corresponding to each bin in the histogram decreases if the gaze area becomes smaller or the division number of QHs increases, there is not only the problem that it becomes difficult to use the similarity information based on motion histogram but also the problem that the reliability of the histogram analysis decreases. That's because the motion

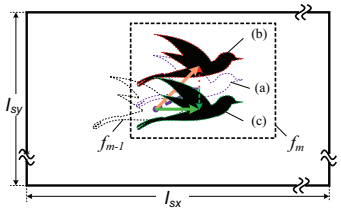
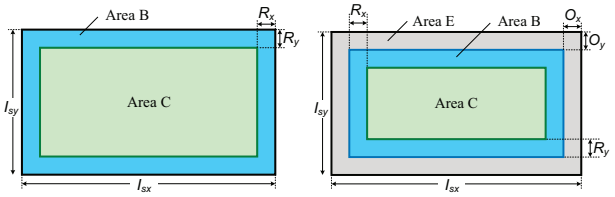


Fig. 2 Relationship among the general motion, the shake motion, and the extracted motion due to screen shake between f_{m-1} and f_m



(i) gaze area method 2 (am2) (ii) gaze area method 3 (am3)

Fig. 3 Definition of two types of gaze area methods

histogram cannot be analyzed sufficiently if the frequency of the histogram is low or it is not given as an appropriate number, as Sturages [23], Scott [24], Freedman et al. [25], and Shimazaki et al. [26] had studied in general histograms. Each problem will be evaluated in the evaluation experiments described below.

3.1 Definition of gaze areas

This method firstly employs area information separated by the edge area that does not include the object area and the central area that includes it. The conventional method targets the whole screen area (i.e., all areas of MBs); however, the proposed method does each gaze area. In the conventional method, let 'am1' be the area method that employs the whole screen and the area type (at) is given by 'All'. When a screen shake occurs, the state of the motion vector detection in each frame changes according to the amount of shake. For example, when a camera shakes from left to right, the motion is extracted as the reverse direction to the camera shake. It is difficult to extract it in the edge area of a frame. Actually, the motion information may change from the detected motion information to no motion compensation (no MC) information as coding information, i.e., not detected motion information. Therefore, due to not detected motion information, the histogram analysis may not be processed accurately. Then, he proposes two types of gaze area method as a screen segmentation method, considering screen areas where motion vectors are detected accurately. As illustrated by **Fig. 3**, they are defined as 'am2' and 'am3'. In particular, gaze areas are defined such as an edge area and a central area considering an object and a background as a simple ROI (Region Of Interest).

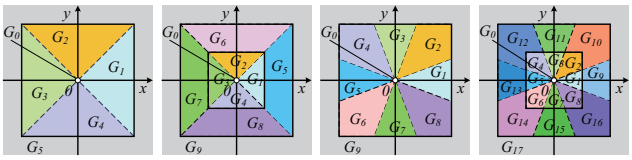


Fig. 4 Examples of division areas based on each direction and size parameter

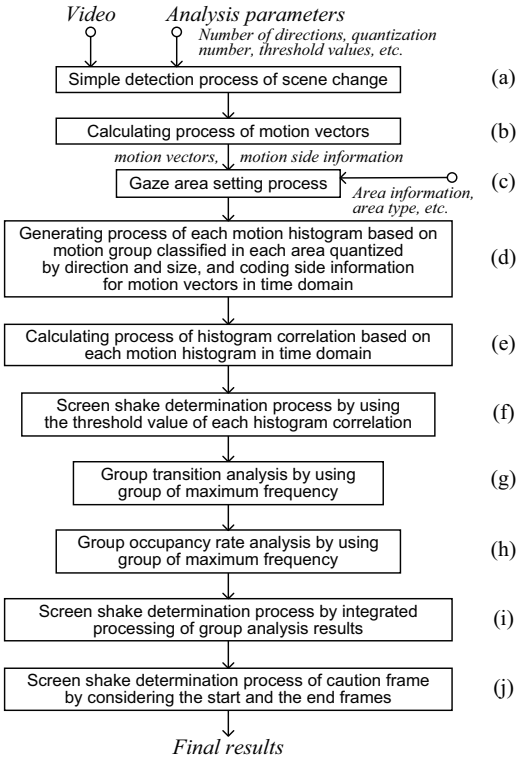


Fig. 5 Processing flow in the proposed method

In am2, a background area (Area B) and a central area (Area C) are defined for the object in a frame. In am3, an edge area (Area E), Area B, and Area C are defined. QHs for each area method are calculated. Here, I_{sx} and I_{sy} denote the horizontal and vertical screen sizes, respectively, R_x and R_y denote the width and the height of Area B, respectively, and O_x and O_y denote the width and the height of Area E, respectively.

3.2 Definition of division areas for QHs

In the conventional method, the direction information is only employed. For example, there is a problem that the change of the histogram does not appear if the screen shake is on the same axis as the motion direction of the object , or if the amount of the object's motion is large and the amount of screen shake is small. Then, to be able to analyze the motion direction and the amount of motion at the same time and identify the direction of screen shake, this method secondly employs QHs. The direction and the size divisions in motion vector space for QHs are defined. **Figure 4** depicts examples of division areas based on each direction and size parameter, where direction parameter (D_n) is set to 4 or 8, and size parameter (Q_n) is set to 1 or 2. Here, a combination of QH-parameters is expressed as $DQ(u,v)$, where $u = D_n$, $v = Q_n$. In Fig 4, x and y denote the amount of movements in the horizontal and vertical directions, respectively. The outer square shows the motion search area, and the origin coordinate shown by a white dot means a zero vector and the group of zero vectors is expressed as G_0. Each division area shows each GD defined in section 2. Here, the numbers are given in the order of rotation and small. Outside the square area indicated by gray color means undetected motion and the group is expressed as the final group number, e.g., G_5, G_9, G_{17}. QHs based on their groups are calculated.

3.3 The processing flow

The processing flow of the proposed method is shown by **Fig. 5**. It consists of ten steps, and the details are shown as follows:

(a) Simple detection process is applied in order to extract a scene change frame in the same manner as step (a) in the conventional method.

(b) Motion vectors based on a block matching method are calculated in the same way.

(c) As shown in section 3.1, three types of gaze area methods are defined as am1, am2, and am3. To utilize the motion features extracted in each area effectively, each type can be adaptively used, considering the amount of motion; however, that is a further study. If the gaze area method is not used, am1 is used simply.

(d) Each motion group histogram is generated based on motion group classified in the area quantized by direction and size of motion vectors according to each area method in time domain. Examples of quantized areas are illustrated by Fig. 4.

(e) Each histogram correlation based on each motion histogram in time domain is calculated as BD.

(f) The screen shake is determined by using the threshold value in the same way.

(g) The author focuses on the motion group transition of maximum frequency in the calculated histogram. Group transition patterns are prepared in advance. Two types are defined: ptype1 denotes motion group patterns by using the opposite directions and ptype2 denotes adjacent motion group patterns. For example, when the pattern is expressed by using a group number in Fig. 4 (i) with $DQ(4,1)$, ptype1: (1,3,1), (2,4,2),

Table 1 Experimental conditions

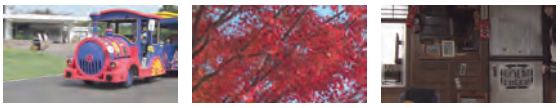
Items	Details of contents
Test sequence	three types: s1(Track train), s2(Red leaves), s3(Drama set) in ITE test sequences[27], where s3 includes one scene change frame between frame numbers 245 and 246.
Sequence type	no-swing image, pseudo swing image
Image size	504 × 280 pel
Number of frames	450 frame
Pseudo swing image	frame numbers to which pseudo shaking is added: 60–120, 180–240, 300–360
Swing angle type (θ)	four types (θ= 0, $\pi/4, \pi/2, 3/4\pi$ rad) (see Eq.(1))
Swing parameter (S)	three types (S=16, 32, 64 pel (see Eq.(1))
Gaze area method (am)	am1(at: All), am2(at: A1-1, A1-2), am3 (at: A2-0, A2-1, A2-2)
Comparison method	The conventional method: CM1, CM2 The proposed method: PM1, PM2 for each gtype1, gtype2
QH-parameters	D_n=4,8 and Q_n=1,2
Other parameters	(I_{sx}, I_{sy}) = (504, 280) pel, (R_x, R_y) = (32, 32) pel, (O_x, O_y) = (64, 64) pel, Th_{max1} = 10 %, Th_{frame} = 15 frame
Evaluation method	accuracy rate, Ev

(3,1,3), (4,2,4), (1,0,1), etc., and ptype2: (1,2,1), (1,3,1), (1,4,1), etc. Here, in this study, group transition state is simply evaluated by two changes, however the number of changes is a future issue. Additionally, only the change of maximum group ratio Th_{rate}=50% or greater is employed. The change from the current group to the other group caused by the screen shake can be analyzed, and consequently the shake directions can be obtained.

(h) To grasp the change of motion in the same motion group, the group ratio of maximum frequency is analyzed by the threshold value Th_{max1}. For the change ratio greater than Th_{max1}, a decision is made whether the ratio increases or decreases. It is possible to judge whether a group change has occurred due to a large change in group ratio, and consequently this step can find the direction in which it is shaking.

(i) To compare with the conventional method (i.e., only step (e)), the author considers four types of processing methods for screen shake analysis: dm1: the processing method of only step (g), dm2: that of only step (h), dm3: that of both steps (g) and (h), and dm4: the integrated processing from steps (f) to (h). In these group analysis methods, it is possible to select which method is easy to detect the screen shake, which cannot be detected by the conventional method.

(j) Finally, the caution frame with screen shake is determined in the same manner as step (f) in the conventional method. Thereafter, the processing of steps (e) to (j) is repeated for all input frames, and the final results are sequentially obtained.



(i) s1 (ii) s2 (iii) s3

Fig. 6 Examples of the test images

In particular, there is the major difference between the conventional method and the proposed method in the three following points: the utilization of gaze areas, group transition analysis, and group ratio analysis of maximum frequency. In the conventional method, it is necessary to set a strict threshold because the shaking movement is not directly captured, whereas, in the proposed method, the movement can be directly detected by the group analysis regardless of whether or not input images include general motion (e.g., a movement of an object and a background movement), if there is a shaking scene. In other words, a histogram correlation analysis is used as a kind of prefilter in this method. Here, the first method consisting of steps (a) to (i) is defined as PM1, and the second method consisting of all steps is defined as PM2 as two types of the proposed method.

4. Evaluation Experiments

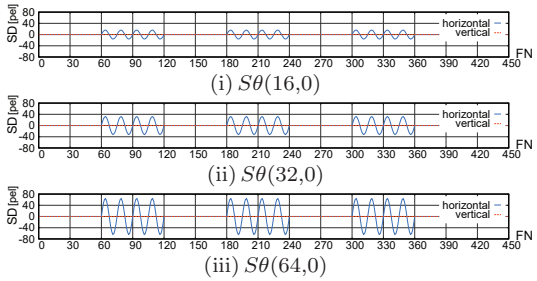
To evaluate the accuracy of detecting screen shake in the proposed method, the evaluation experiments are performed. Firstly, the experimental conditions of the conventional method and the proposed method in this study are shown. Next, the results of the experiments and the consideration are described.

4.1 Conditions of simulation experiments

The experimental conditions are shown in **Table 1**. As shown in Table 1, two sequence types: no-swing images and pseudo swing images in each sequence type are employed in the evaluation experiments. All these images are generated by downsizing and trimming the original images in ITE test sequences[27]. Note that these sequences have not only the movement of the object but also that of the camera. **Figure 6** depicts examples of the test images.

First, the pseudo swing images are generated under the conditions indicated by Table 1. The detailed conditions are shown as follows:

- A pseudo swing function is defined by using a sine function to make it easier to evaluate the accuracy of detecting the screen shake. When the original coordinates in the original image are given by (x_o, y_o) and the displacement coordinates in the pseudo image due



(i) $S\theta(16,0)$

(ii) $S\theta(32,0)$

(iii) $S\theta(64,0)$

Fig. 7 Examples of the pseudo swing function

to screen shake are given by (x_d, y_d), they are represented by Eq. (1).

$$\begin{cases} x_d &= x_o + S\sin(4\pi t) \times \cos\theta \\ y_d &= y_o + S\sin(4\pi t) \times \sin\theta \end{cases} \quad (1)$$

Where S denotes the amplitude parameter, $t = FN/30$ (FN: frame number), and the shaking direction is represented by $\theta\,(0 \leq \theta < 2\pi$ rad). Here, the frequency parameter is set to 2 under the condition that the number of pixels of the position gap at the extreme value is 1 or greater in Eq. (1).

- Twelve pseudo swing images based on Eq. (1) are generated by using three amplitude parameters (S) and four direction parameters (θ). The examples of the pseudo swing function are shown in **Fig. 7**, where SD denotes the amplitude and FN denotes the frame number. In this study, pseudo shake information is not added to the whole sequence, and it is partially added to three places at every 60 frames. Here, a combination of swing parameters is expressed as $S\theta(p, q)$, where $p = S$, $q = \theta$.

Next, three area types of gaze area methods are defined by am1–am3. All(am1), A2-0(am3) for Area E, A1-1(am2) and A2-1(am3) for Area B, and A1-2(am2) and A2-2(am3) for Area C are used as area types(at). Two types of gtype1 and gtype2 are defined as the group analysis method in the proposed method. Gtype1 denotes the group transition analysis in the maximum group, and gtype2 denotes that in the maximum group excluding GNMI. Note that, strictly speaking, the conventional method is used only based on am1 with $DQ(4, 1)$ and $DQ(8, 1)$. However, to evaluate the gaze area methods and QH-parameters, the results in $DQ(4, 2)$ and $DQ(8, 2)$ are included in the results of the conventional method, and the results of am2 and am3 are done. Other parameters and the threshold values are shown in Table 1.

4.2 Definition of evaluation value

Each accuracy rate is calculated by using no-swing and pseudo swing images in the conventional method and the proposed method. However, to evaluate the accuracy of screen shake detection by comparing the conventional method with the proposed method, it is necessary to consider whether the result of the obtained accuracy is correct. In brief, there is a trade-off relationship between each accuracy of no-swing images and pseudo swing images. Therefore, an evaluation index is required to simultaneously indicate whether each accuracy is high or not. Then, since the ideal goal is 100 % in accuracy rate, the Euclidean distance between 100 % and the obtained accuracy is also used as the performance evaluation. In particular, this paper focuses on 2D accuracy rate space between the accuracy of no-swing images and that of pseudo swing images, and the Euclidean distance between the 2D coordinates of (100,100) and the calculated 2D coordinates is used as the evaluation value. Let it be Ev, and it is calculated by Eq. (2), where the accuracy rate obtained in each no-swing image is given by Ar_o and that in each pseudo swing image is given by Ar_s.

$$Ev = \sqrt{(100 - Ar_o)^2 + (100 - Ar_s)^2} \quad (2)$$
$$(0 \leq Ar_o \leq 100,\ 0 \leq Ar_s \leq 100)$$

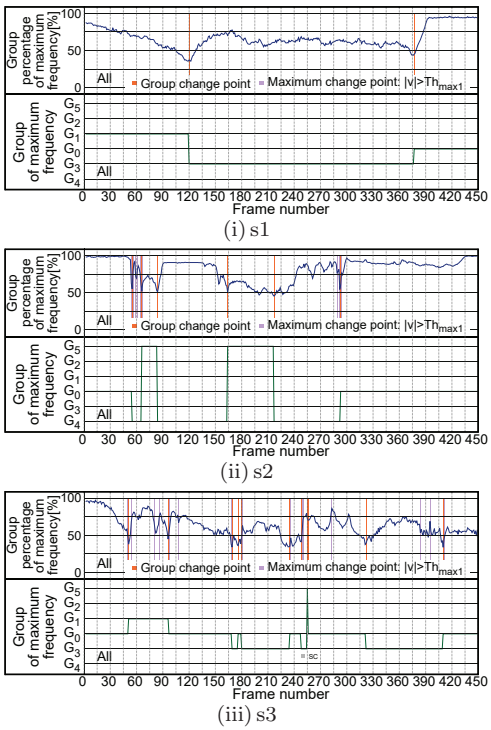
Therefore, it can be evaluated that the accuracy of detecting screen shake is improved if the value of Ev becomes smaller.

4.3 Simulation results and consideration

In the conventional method, there is a trade-off problem in the accuracy of detecting the screen shake. The movement of the object or the background is considered as general motion information. For example, when the screen shake is simultaneously detected for screen shake images and no screen shake images with and without the general movement, it is extremely difficult to find an optimum threshold in the determination method by using a BD threshold. That's because the threshold value should be set large in the conventional method not to detect the general movement except for screen shake. That is to say, the higher the sensitivity, the greater the false detection rate, whereas the lower the sensitivity, the lower the rate. In the conventional method, this problem is a trade-off problem, and it leads to the problem of the reduction of detection accuracy. However, the proposed method does not depend on the threshold setting for BD. Then, he evaluates whether the proposed method can improve the trade-off problem by the simulation experiments.

(a) Detection state by group transition and maximum rate analysis

The details of the detection state, i.e., the changes of the group percentage of maximum frequency and the

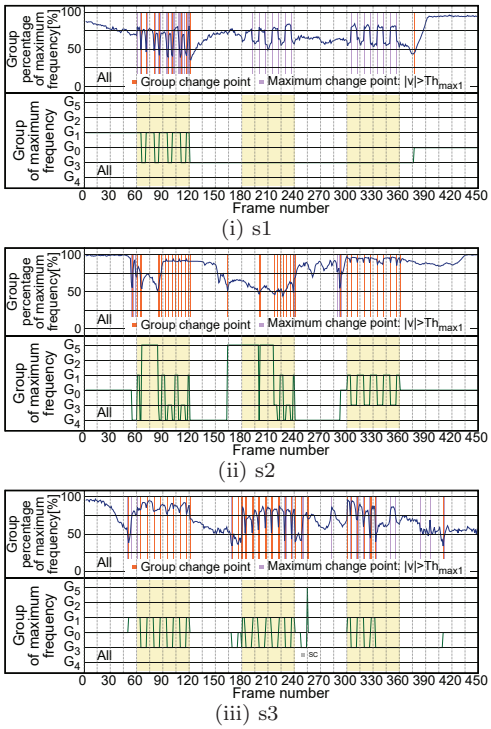


(i) s1

(ii) s2

(iii) s3

Fig. 8 Examples of each group change in no-swing images (am1 with $DQ(4,1)$)



(i) s1

(ii) s2

(iii) s3

Fig. 9 Examples of each group change in pseudo swing images ($S\theta(16,0)$ and am1 with $DQ(4,1)$)

group of maximum frequency, in no-swing and pseudo swing images are shown in **Fig. 8** and **Fig. 9**, respectively. Each result is obtained by the proposed method of am1 with $DQ(4,1)$. Each orange line indicates the frame of group transition (group change point), each violet line indicates the frame with the variation of the maximum value $\geq Th_{max1}$ (maximum change point), and each yellow area shows the true shaking images in Fig. 8 and Fig. 9. Note that s3 sequence has a scene change and it is represented by 'SC'. in Fig. 8 and Fig. 9. **Figure 10** and **Fig. 11** depict the examples of each group change in am1 with $DQ(8,2)$ and am3 with $DQ(4,2)$, respectively.

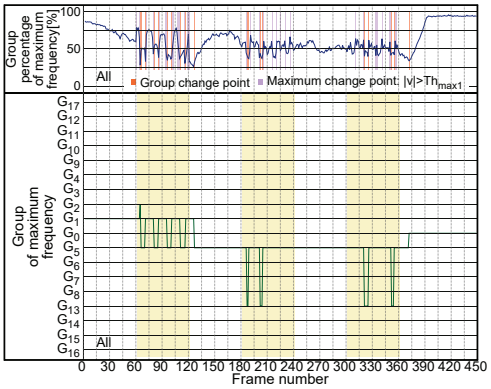


Fig. 10 An example of each group change in pseudo swing images (s1 in $S\theta(16,0)$ and am1 with $DQ(8,2)$)

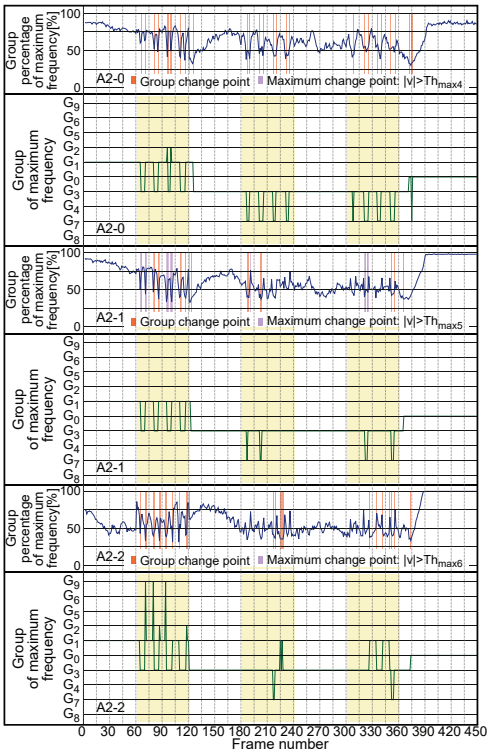


Fig. 11 An example of each group change in pseudo swing images (s1 in $S\theta(16,0)$ and am3 with $DQ(4,2)$)

As shown in Fig. 8, the object movement in frame or the movement of camera appears as group transition and maximum value change in the no-swing images for each sequence. On the other hand, Fig. 9 shows that the movement caused by shaking occurs in the pseudo swing images. In particular, it can be found that the number of group changes increases when screen shake occurs. However, as shown in Fig. 9 (i), when the amplitude of swing

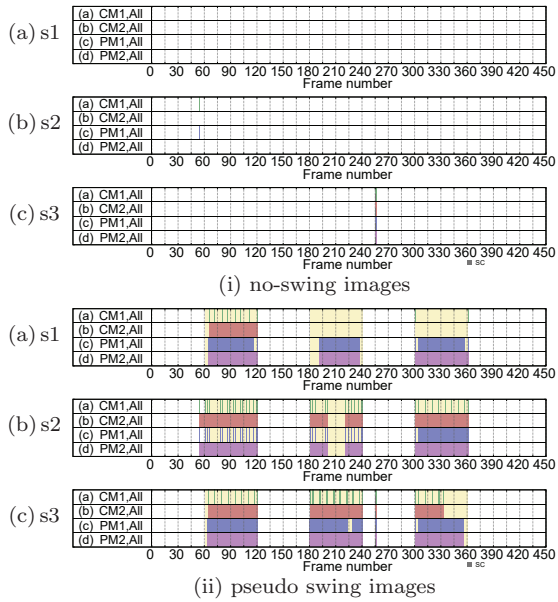


Fig. 12 Comparison of screen shake determination results (pseudo swing images in $S\theta(16,0)$ and am1 with $DQ(4,1)$)

is small, the group transition does not appear. Whereas, as shown in Fig. 10 and Fig. 11, even if the amplitude is small, the group transition appears in am1 with $DQ(8,2)$ and in am3 with $DQ(4,2)$. Next, it is required to extract screen shake in the same group when the group transition does not occur. Then, to solve the problem, the proposed method focuses on the change of the group ratio in the same group of maximum frequency. By Fig. 9–Fig. 11, the change can be found. However, the group transitions don't appear when the number of division areas is small such as $DQ(4,1)$, whereas, since the group transitions appear if the large number of division areas are used such as $DQ(4,2)$ and $DQ(8,2)$, a screen shake scene can be extracted. Am2 and am3 are more effective than am1 in s1 sequence as gaze area methods in this experiment. By the results, there is no problem of the frequency of QHs in each gaze area method.

In practice, the detection results are evaluated in the conventional method and the proposed method. **Figure 12** shows the comparison of the screen shake determination results in no-swing and pseudo swing images for each test sequence in each method. Note that the threshold value is set to 0.35 in a BD determination method, and the proposed method with gtype1 is employed. Since all no-swing images don't have screen shake, it should not be detected. However, it is falsely detected in CM1 and PM1 in s2 sequence, whereas the detection errors are improved in CM2 and PM2. As shown in Fig. 12, there are undetected frames of screen shake in pseudo swing images for the conventional method; however, the proposed method can detect some of them. In particular, the improvement is remarkable in s1 and s3 sequences, However, there are still some frames that cannot be detected in s2 sequence despite the screen shake frames even if the proposed method is used. One of the reasons is that an appropriate vector cannot be estimated for the movement of the background tree because the camera is moving diagonally.

(b) Evaluation of accuracy improvement in *Ev*

In the conventional method and the proposed method, each accuracy of detecting screen shake is analyzed in detail by using an Ev as an evaluation value.

Firstly, the comparison results of each Ev in each method using am1 for each threshold of BD are shown in **Table 2**. In Table 2, 'Ave' denotes the average value of Ev in each sequence, and this study discusses the effectiveness based on the average value in the conventional method and the proposed method. Note that each PM2 with gtype1 and gtype2 is employed; only the results of CM2 and PM2 are shown since they are basically more effective than CM1 and PM1, respectively; only the results of am1 are shown; the results of dm1–dm3 are omitted since dm4 is equal to or greater than them in the accuracy of all the simulation experiments.

As shown in Table 2, the highest accuracy value is 4.70 in CM2, and the highest accuracy values are 2.84 and 2.75 in PM2 with gtype1 and gtype2, respectively for $D_n = 4$. Consequently, it is improved by up to 1.95 in PM2. It is found that the quantization extension method is more effective in each method using CM2, and gtype2, which deals with the maximum group except for GNDI, is more effective than gtype1 in the proposed method using PM2. However, the average of Ev in D_n=4 is more superior than that in D_n=8. That is because if the threshold value is set large, the change in the histogram is less likely to appear in the determination method using only BD, and at the same time, the accuracy deteriorates due to the trade-off problem between no-swing images and pseudo swing images. It is necessary to consider the parameter setting of D_n and Q_n in the determination method using BD and motion analysis separately, and it is a future issue. On the other hand, although not listed in Table 2, by the results of Ev in A2-2 of am3, the accuracy in PM2 with gtype2 is 4.01, and it is 1.13 better than the accuracy of 5.14 in CM2. In particular, PM2 with gtype2 is more effective than that with gtype1. As shown in Table 2, the BD threshold value corresponding

Table 2 Comparison results of each Ev in the conventional method and the proposed method using am1 ($S\theta(16,\ 0)$)

Th	Q_n	$D_n=4$									$D_n=8$								
		CM2			PM2 with gtype1			PM2 with gtype2			CM2			PM2 with gtype1			PM2 with gtype2		
		s1	s2	s3	s1	s2	s3	s1	s2	s3	s1	s2	s3	s1	s2	s3	s1	s2	s3
0.20	1	3.12	4.60	11.22	1.56	6.16	9.19	1.56	6.16	9.19	3.12	22.59	11.52	0.89	23.58	17.17	0.89	23.73	17.17
		Ave: **6.32**			Ave: **5.64**			Ave: **5.64**			Ave: **12.41**			Ave: **13.88**			Ave: **13.93**		
	2	2.90	4.60	11.40	0.67	6.16	13.26	0.67	6.31	13.26	0.67	22.59	14.74	0.89	23.58	28.57	0.89	23.73	28.57
		Ave: **6.30**			Ave: **6.70**			Ave: **6.75**			Ave: **12.67**			Ave: **17.68**			Ave: **17.73**		
0.25	1	6.03	2.28	7.37	3.12	4.60	6.51	3.12	4.60	6.51	4.91	14.99	8.55	0.89	22.59	11.73	0.89	22.74	11.73
		Ave: **5.22**			Ave: **4.74**			Ave: **4.74**			Ave: **9.49**			Ave: **11.74**			Ave: **11.79**		
	2	5.80	2.28	6.03	2.90	4.60	8.05	2.90	4.81	8.05	0.67	14.99	10.59	0.89	22.59	15.85	0.89	22.74	15.85
		Ave: **4.70**			Ave: **5.18**			Ave: **5.25**			Ave: **8.75**			Ave: **13.11**			Ave: **13.16**		
0.30	1	27.45	2.28	7.37	4.02	2.28	2.23	4.02	2.28	2.23	26.12	14.99	7.37	0.89	14.99	8.55	0.89	15.15	8.55
		Ave: **12.37**			Ave: **2.84**			Ave: **2.84**			Ave: **16.16**			Ave: **8.15**			Ave: **8.20**		
	2	11.38	2.28	7.37	5.80	2.28	2.23	5.80	2.50	2.23	10.49	14.99	7.37	0.89	14.99	12.88	0.89	15.15	12.88
		Ave: **7.01**			Ave: **3.44**			Ave: **3.51**			Ave: **10.95**			Ave: **9.59**			Ave: **9.64**		
0.35	1	28.12	6.25	7.37	4.02	2.28	2.23	4.02	2.28	2.23	27.90	4.33	7.37	1.56	14.99	7.37	1.56	15.15	7.37
		Ave: **13.91**			Ave: **2.84**			Ave: **2.84**			Ave: **13.20**			Ave: **7.97**			Ave: **8.03**		
	2	28.12	2.01	7.37	5.80	2.28	2.23	5.80	2.50	2.23	27.90	4.33	7.37	5.80	14.99	13.48	5.80	15.15	13.48
		Ave: **12.50**			Ave: **3.44**			Ave: **3.51**			Ave: **13.20**			Ave: **11.43**			Ave: **11.48**		
0.40	1	36.61	9.60	7.37	4.46	6.25	2.23	4.46	2.01	2.23	27.90	3.04	7.37	1.56	10.98	7.37	1.56	11.16	7.37
		Ave: **17.86**			Ave: **4.32**			Ave: **2.90**			Ave: **12.77**			Ave: **6.64**			Ave: **6.70**		
	2	34.60	4.91	7.37	11.38	2.01	2.23	11.38	2.23	2.23	27.90	3.04	7.37	11.38	10.98	9.39	11.38	11.16	9.39
		Ave: **15.62**			Ave: **5.21**			Ave: **5.28**			Ave: **12.77**			Ave: **10.58**			Ave: **10.64**		
0.45	1	38.84	10.27	7.37	5.36	9.60	2.23	5.36	0.67	2.23	35.49	3.04	7.37	1.56	9.96	7.37	1.56	10.12	7.37
		Ave: **18.82**			Ave: **5.73**			Ave: **2.75**			Ave: **15.30**			Ave: **6.29**			Ave: **6.35**		
	2	38.84	9.60	7.37	12.05	5.13	2.23	12.05	5.36	2.23	35.49	3.04	7.37	11.38	9.96	9.39	11.38	10.12	9.39
		Ave: **18.60**			Ave: **6.47**			Ave: **6.55**			Ave: **15.30**			Ave: **10.24**			Ave: **10.30**		
0.50	1	40.18	17.19	8.71	5.80	10.27	2.23	5.80	0.67	2.23	40.18	7.59	7.37	2.01	9.96	7.37	2.01	10.12	7.37
		Ave: **22.02**			Ave: **6.10**			Ave: **2.90**			Ave: **18.38**			Ave: **6.44**			Ave: **6.50**		
	2	40.18	17.19	7.37	13.84	9.82	2.23	13.84	7.81	2.23	40.18	7.59	7.37	12.05	9.96	9.39	12.05	10.12	9.39
		Ave: **21.58**			Ave: **8.63**			Ave: **7.96**			Ave: **18.38**			Ave: **10.47**			Ave: **10.52**		

Table 3 Results of each Ev at the highest accuracy in each method and comparison results of an Ev gain between CM2 and each proposed method

Test Seq.	am	s1						s2						s3					
(S, θ)		CM1	CM2	PM1	PM2	Ar_1	Ar_2	CM1	CM2	PM1	PM2	Ar_1	Ar_2	CM1	CM2	PM1	PM2	Ar_1	Ar_2
(16,0)	1	21.93	0.67	2.01	0.67	-1.34	0.00	28.16	2.01	11.84	0.67	-9.83	1.34	33.04	6.03	3.80	2.01	2.23	4.02
(16,0)	2	22.70	1.56	15.19	1.34	-13.63	0.22	28.48	2.01	10.27	0.67	-8.26	1.34	32.59	5.80	7.59	5.58	-1.79	0.22
(16,0)	3	21.06	0.67	8.93	0.67	-8.26	0.00	28.70	2.01	6.92	0.67	-4.91	1.34	32.40	5.58	7.39	5.36	-1.81	0.22
(32,0)	1	17.04	0.67	8.26	0.67	-7.59	0.00	27.25	2.01	11.62	0.67	-9.61	1.34	33.93	1.56	2.90	1.34	-1.34	0.22
(32,0)	2	16.51	0.67	3.80	0.67	-3.13	0.00	26.61	1.56	6.47	0.89	-4.91	0.67	32.81	1.56	2.90	1.34	-1.34	0.22
(32,0)	3	18.18	0.67	2.68	0.67	-2.01	0.00	26.02	0.67	3.35	0.67	-2.68	0.00	32.37	0.67	2.90	0.67	-2.23	0.00

to the highest accuracy value in each PM2 is larger than the value in each CM2; therefore, the proposed method has the advantage that the accuracy is not easily affected by the threshold settings of BD, e.g., it is able to be used in Th of around 0.30–0.45 and D_n=4. Furthermore, the gaze area method in PM2 with $DQ(4,2)$ is more effective than CM1 with am1 and D_n=4. However, there is a problem of the parameter setting in both gaze area parameters and QH-parameters. In this experiment, the accuracy is higher when Q_n is set large than when D_n is set large. This evaluation is related to the accuracy of motion vectors for each sequence, and it is a future study.

Secondly, **Table 3** shows the results of each Ev at the highest accuracy in each method and the comparison results of an Ev gain between CM2 and each proposed method, where Ar_1 denotes the gain between CM2 and PM1, and Ar_2 denotes the gain between CM2 and PM2. Here, the results in only the cases of $S\theta$(16,0) and $S\theta$(32,0) are shown in Table 3. In the experimental results, the author pays attention to the case where the proposed method is superior to the conventional method. As shown in Table 3, in $S\theta$(16,0), Ar_2 is 0.22 in am2 in s1 sequence, Ar_2 is 1.34 in am1, am2, and am3 in s2 sequence, Ar_1 is 2.23 in am1 in s3 sequence, and Ar_2 is 4.02 in am1 and Ar_2 is 0.22 in am2 and am4 in s3 sequence. In $S\theta$(32,0), Ar_2 is 1.34 and 0.67 in am1 and am2 in s2 sequence, respectively, and Ar_2 is 0.22 in am1 and am2 in s3 sequence. Additionally, Ar_2 is 0.22 in am2 ($S\theta(32,\pi/2)$) and am3 ($S\theta(64,\pi/2)$). On the other hand, he considers the case where the proposed method is inferior. Ar_2 is -0.09 in am1 ($S\theta(16,\pi/4)$), Ar_2 is -1.34 in am2 ($S\theta(32,\pi/4)$), and Ar_2 is -0.89 in am1 ($S\theta(16,3\pi/4)$). In the other cases, Ar_2 is 0 and the accuracy in CM2 is the same as that in PM2. Of course, since PM1 does not include the time extension processing (see (j) in section 3), PM1 is inferior to CM2. However, it is greatly improved over CM1 in all cases. Consequently, if all gaze area methods can be adaptively employed, the proposed method is superior to the conventional method in all cases. In particular, it is revealed that the pro-

Table 4 Results of each Ev at the highest accuracy in the proposed method using dm3

S	area type					
	am1	am2		am3		
	All	A1-1	A1-2	A2-0	A2-1	A2-2
16	8.41	14.29	14.17	14.57	13.37	11.72
32	12.05	7.66	11.24	6.18	10.71	10.14
64	7.07	9.00	9.89	8.33	9.35	8.95

posed method is better than the conventional method if the screen shake is small, e.g., $S = 16$. However, there is a future issue in the improvement of accuracy reduction in the shaking direction (θ=$\pi/4$), it is necessary to study an adaptive gaze area method, and a theoretical analysis is required for the relationship between the shaking direction and the moving direction of the object to accurately grasp the performance limit of the proposed method.

Finally, he considers the results of motion detection of screen shake in the proposed method. In the conventional method, a determination method using a BD analysis could find whether the screen was shaking; however, it could not detect in which direction the screen was shaking. On the other hand, as shown in Fig. 8–11, since the proposed method uses the group transition and the maximum group ratio, it is a crucial point to be able to extract a scene of screen shake corresponding to simple shaking information defined as swing patterns. To evaluate the effectiveness of the group analysis in the proposed method, he focuses on dm3 corresponding to only the group analyses, and the Ev by using only dm3 for each gaze area method is calculated. The results of each Ev at the highest accuracy in the proposed method using dm3 are shown in **Table 4**, where $\theta = 0$. As shown in Table 4, the values of the highest accuracy are 8.41 in $S = 16$, 6.18 in $S = 32$, and 7.07 in $S = 64$. It is found that the screen shake can be detected only by the group analyses, and the motion direction of the screen shake can be obtained according to the accuracy. The accuracy can be further improved by using the gaze area method, e.g., am1 is effective in S=16 and am3 is effective in S=32. An adaptive gaze area method using the frequency distribution can be also considered. Hence, since the proposed method can detect the occurrdance of screen shake and its direction at the same time, it is more effective than the conventional method. Furthermore, it is important to accurately grasp the performance limit of the proposed method, and a theoretical evaluation is a future task.

As mentioned above, it is revealed that the proposed method by using QHs and a gaze area method is more effective than the conventional method. However, as a further study, there is an adaptive method in each gaze area parameter and each QH-parameter.

5. Conclusions

If videos have screen shake information, it is one of the important issues to prevent viewers from VIMS. A screen shake determination method, which had used the block matching method as a simple motion analysis, motion direction histograms, and this similarity, had been proposed. However, there was still the problem that the accuracy of detecting screen shake decreases, when the amount of screen shake was small. The conventional method could not extract the direction information of screen shake. To solve the problems, this paper proposed a novel screen shake determination method based on 2D motion histogram analyses. In particular, there were three features: the use of gaze areas, the analysis of group transition of maximum frequency, and the group ratio analysis of maximum frequency group in the proposed method. A new evaluation value Ev was defined in consideration of both the accuracy of no-swing images and pseudo swing images. Simulation experiments showed that the Ev in the proposed method was at most 4.02 smaller than that in the conventional method ($S\theta(16, 0)$). Furthermore, the direction of the screen shake could be extracted by the proposed method. Therefore, it was revealed that the proposed method not only improved the accuracy of detecting the small screen shake for the conventional method, but also it had the advantage of extracting the direction of the screen shake.

As further studies, an adaptive method for each gaze area parameter and each QH-parameter and a theoretical evaluation will be required.

References

1) S. Chiba: “Health Hazard by Video Image”, Journal of ITE, Vol. 56, No. 6, pp. 926–927 (2002). (in Japanese)
2) N. Sugita, M. Yoshizawa, A. Tanaka, K. Abe, T. Yambe, S. Nitta, S. Chiba: “Evaluation of the Effect of Visually-Induced Motion Sickness Based on Pulse Transmission Time and Heart Rate”, Proc. of SICE 2004 Annual Conference, Vol. 3, pp. 2473–2476 (2004).
3) I. Tsubaki, T. Morita, K. Aizawa, T. Saito: “The Analysis of Oscillatory Motion of Videos Affected by Camera Shaking for Visually Induced Motion Sickness”, IEICE A, Vol. 89, No. 3, pp. 262–267 (2006).
4) M. Abe, M. Yoshizawa, N. Sugita, A. Tanaka, S. Chiba, T. Yambe, S. Nitta: “Estimation of Effects of Visually-Induced Motion Sickness Using Independent Component Analysis”, Proc. of 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2007).
5) H. Watanabe, H. Ujike: “The Activity of ISO/Study Group on

"Image Safety" and Three Biological Effect", Proc. of Second International Symposium on Universal Communication (ISUC '08), pp. 210–214 (2008).

6) A. Tanaka, N. Sugita, M. Yoshizawa, M. Abe, T. Yambe: "Dynamic Characteristics between the Subjective Score of Motion Sickness Discomfort and Video Global Motion", Proc. of 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 1368–1369 (2010).
7) H. Ujike: "Developing an Evaluation System of Visually Induced Motion Sickness for Safe Usage of Moving Images", Synthesiology English Edition, Vol. 5, No. 3, pp. 139–149 (2012).
8) T. Kiryu, A. Iijima: "A Multi-Timescale Autonomic Regulation Model for Interpreting Visually Induced Motion Sickness", Proc. of 2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE), pp. 254–257 (2014).
9) N. Kobayashi, R. Iinuma, Y. Suzuki, T. Shimada, M. Ishikawa: "Using Bio-Signals to Evaluate Multi Discomfort in Image Viewing - Balancing Visually Induced Motion Sickness and Field of View -", Proc. of 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6198–6201 (2015).
10) P. Bala, D. Dionísi, V. Nisi, N. Nunes: "Visually Induced Motion Sickness in 360° Videos: Comparing and Combining Visual Optimization Techniques", Proc. of 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct) (2018).
11) S. Wibirama, K. Hamamoto: "Investigation of Visually Induced Motion Sickness in Dynamic 3D Contents Based on Subjective Judgment, Heart Rate Variability, and Depth Gaze Behavior", Proc. of 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4803–4806 (2014).
12) S. Wibirama, T. Wijayanto, H.A. Nugroho, M. Bahit, M.N. Winadi: "Quantifying Visual Attention and Visually Induced Motion Sickness During Day-Night Driving and Sleep Deprivation", Proc. of 2015 International Conference on Data and Software Engineering (ICoDSE) (2015).
13) A. Sugiura, K. Akachi, A. Yoshida, C. Ito, S. Kondo, K. Tanaka, H. Takada: "Experimental Study on Control of Visually Evoked Postural Responses by Galvanic Vestibular Stimulation", Proc. of 12th International Conference on Computer Science and Education (ICCSE) (2017).
14) Y. Onuki, I. Kumazawa: "Reorient the Gazed Scene Towards the Center: Novel Virtual Turning Using Head and Gaze Motions and Blink", Proc. of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (2019).
15) M. Kodama:"A Consideration on a Determination Method for Screen Shake with Histograms in Motion Vectors", Proc. of 5th IIEEJ International Workshop on Image Electronics and Visula Computing (IEVC2017), 1P-7 (2017).
16) M. Kodama:"A Screen Shake Determination Method Using Histograms of Motion Vectors in Video Scenes", IIEEJ Trans. on Image Electronics and Visual Computing, Vol.6, No.1, pp.1–12 (2018).
17) X.C. Chen, N. Chaddha: "Lossy Adaptive Coding of Motion Vectors with Global Camera Motion Compensation", Proc. of Conference Record of The Thirtieth Asilomar Conference on Signals, Systems and Computers (1996).
18) H. Jozawa, K. Kamikura, A. Sagata, H. Kotera, H. Watanabe: "Two-Stage Motion Compensation Using Adaptive Global MC and Local Affine MC", IEEE Trans. on Circuits and Systems for Video Technology, Vol.7, Issue 1, pp.75–85 (1997).
19) M. Kodama: "A Consideration of A High-Speed Video Identification Method Using Histogram Information of Coded Motion Vectors", The Journal of IIEEJ, Vol.43, No.4, pp.499–507 (2014). (In Japanese)
20) M. Kodama: "A Video Content Identification Method Using Moving Regions Information by Histogram of Motion Vectors", The Journal of IIEEJ, Vol.42, No.5, 625–632 (2013). (In Japanese)
21) M. Kodama: "A Consideration on Accuacy Improvement of Screen Shake Determination Methods Using Motion Histograms", Proc. of IIEEJ International Conference on Image Electronics and Visual Computing 2019 (IEVC2019), 1P-9 (2019).
22) A. Bhattacharyya: "On a Measure of Divergence between Two Statistical Populations Defined by their Probability Distributions", Bulletin of the Calcutta Mathematical Society, Vol.35, pp.99–109 (1943).
23) H.A. Sturges: "The Choice of a Class Interval", Journal of the American Statistical Association, Vol. 21, No. 153, pp.65–66 (1926).
24) D.W. Scott: "On Optimal and Data-Based Histograms", Biometrika, Vol. 66, Issue 3, pp.605–610 (1979).
25) D. Freedman, P. Diaconis: "On the Histogram as a Density Estimator: L_2 Theory", Zeitchrit für Wahrscheinlichkeitsheorie und Verwandte Gebiete, Vol.57, No. 4, pp.453–476 (1981).
26) H. Shimazaki, S. Shinomoto: "A Mthod for Selecting the Bin Size of a Time Histogram", Neural Computation, Vol.19, No. 6, pp.1503–1527 (2007).
27) ITE/ARIB HDTV Test Materials –Second Edition, ITE, https://www.ite.or.jp/contents/chart/ippan_en.html (2020).

(Received June 11, 2020)
(Revised October 12, 2020)

Mei KODAMA (*Fellow*)
He received the B.E, M.E, and Ph.D degrees, all in Engineering from Waseda University, Tokyo, Japan, in 1992, 1994 and 1997, respectively. He was the Research Associate of Waseda University from 1995 to 1998. He joined in Hiroshima University from 1998 to 2001 as the Associate Professor at Center for Technology Research and Development of Hiroshima University. He is currently working as the Associate Professor at Graduate School of Advanced Science and Engineering, School of Integrated Arts and Sciences, and Information Media Center of Hiroshima University. His research interests include audio-visual communication, especially, scalable video coding, image processing, video content distribution system and video communication services. He received Best Author Award of the Journal of ITE from ITE in 2013, Best Journal Paper Award from IIEEJ in 2018, and Excellent Journal Paper Award from IIEEJ in 2014 and 2020. He has been an editor in chief of Journal of IIEEJ as well as IIEEJ Trans. on Image Electronics and Visual Computing since 2016. He is a member of IIEEJ, IEICE, IPSJ, ITE, IEEE and ACM.

IIEEJ **Paper** *-- Special Issue on Extended Papers Presented in IEVC2019 Part II --*

Digital Contents for Creating and Watching 3DCG of Vehicles Based on Drawing their Pictures

Shinji MIZUNO†(*Member*)

†Faculty of Information Science, Aichi Institute of Technology

<**Summary**> In this paper, the author developed a method to generate 3DCG models of trains and cars by drawing their simple pictures with pens on papers. The author also developed a system that allows us to see the generated 3DCG models of trains and cars running in a 3DCG diorama three-dimensionally. The author created digital contents applying the developed method and system, and exhibited them at events hosted by a railway company and an automobile company. With this content, the users could run 3DCG models of train or car in the CG diorama and watch them immediately by just drawing pictures of vehicles on papers with pens. More than 300 children experienced the contents at each event, and enjoyed creating 3DCG vehicles by drawing and watching them. The author confirmed that the proposed method was useful to create interactive contents which attract many children at events.

Keywords: 3DCG, drawing, train and car, trick art, digital content

1. Introduction

Recently, the use of digital contents has greatly increased in various situations such as events, shops, restaurants, stages, and so on[1)–3)]. They have realized adding new values to familiar life and arts, creating unrealistic spaces, and have been attracting many people from children to the elderly.

In such events, digital contents that can generate 2D / 3DCG based on drawing or coloring are very popular and often used. In usual drawing and coloring, only papers and pens are necessary, children can easily experience them, and adults can also enjoy them. Therefore, various digital contents based on drawing pictures on papers with pens are developed[4)–6)]. These contents are particularly in high demand for events targeting families, and such contents with various target objects of drawing and expression methods are still expected.

In this research, the author developed two novel 3DCG contents: "I CAN DRAW" and "Toyota City Creation" for children based on drawing objects on papers with pens. The target objects of drawing are trains in "I CAN DRAW", and cars in "Toyota City Creation". The author developed methods to generate 3DCG models of trains and cars from hand-drawn simple pictures of them for each content. The author also developed a system that allows us to see the trains and cars generated from hand drawn pictures running in a 3DCG diorama three-dimensionally.

Both trains and cars are very popular as entertainment and education contents. Miniature trains and cars are very popular among children's toys[7),8)]. There are many animations on trains and cars[9)–11)], and museums[12)–14)]. Furthermore, since railway / automobile companies have many opportunities to hold events for children and families, the demand for digital contents on trains and cars is considered to be high.

A content for creating 3DCG models of trains and cars based on hand drawn pictures has been also developed. "Paper Train" and "Paper Racer" generate 3DCG models of trains and cars by drawing their side views in the same way as the propsed method in this paper[15)]. However, only the side of trains or cars are textured, and the front and top of the vehicles have inappropriate colors. "SKETCH RACING" also generates 3DCG models of trains and cars by drawing their side views[16)]. However, the created 3DCG cars are based on 3D models prepared beforehand, and the shapes of the generated models are not always as shown in their pictures. Also, the generated trains and cars are only placed in a general 3DCG scene in both contents. On the other hand, textures are appropriately applied not only to the side but also to the front and top of the vehicles, and the shapes of the generated models are exactly the same as the pictures of their

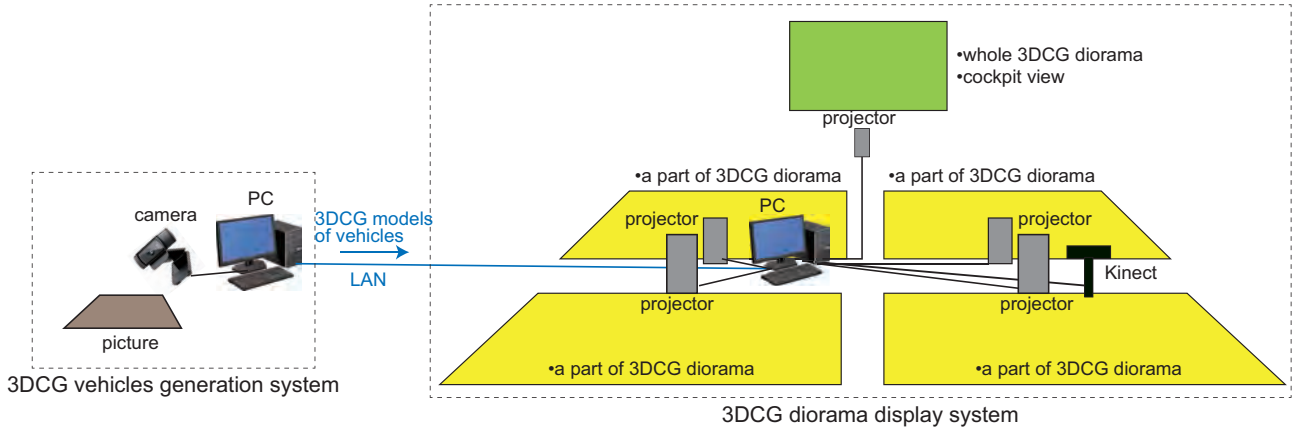


Fig. 1 Overview of our content

side views in our method. And the generated vehicles can be observed three-dimensionally as if they were actually running in a 3DCG diorama in front of us. In "Toyota City Creation", it is also possible to interact with the 3DCG diorama.

2. Overview of the Contents

The contents developed in this research are "I CAN DRAW" for drawing and generating trains, and "Toyota City Creation" for drawing and generating cars. Both contents consist of a system that generates 3DCG models of trains and cars from hand drawn pictures, and a system that displays and interacts with a 3DCG diorama in which trains and cars generated from pictures are running. **Figure 1** shows the overview of our contents.

The system for generating 3DCG models of vehicles consists of a PC and a camera. This system takes an image of the hand-drawn vehicle in the paper using the camera, analyzes the image, and generates a 3DCG model of the vehicle from the image in real time.

The system for displaying 3DCG diorama consists of a PC and 5 projectors. This system receives 3DCG vehicles' data via the network from the system for generating 3DCG models. In addition, the system builds a city diorama by 3DCG, and runs the received 3DCG vehicles in the diorama. The 3DCG image of the whole scene is divided and projected by four projectors on a floor.

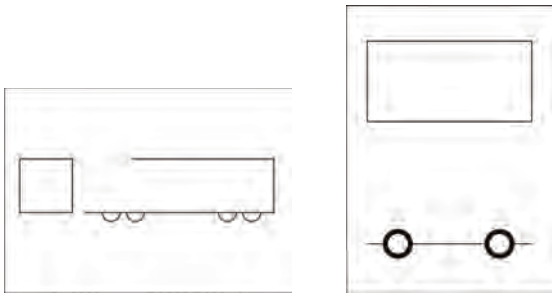
3. Generating 3DCG Models of Vehicles from Hand-Drawn Pictures of Vehicles

3.1 Policy of generating 3DCG models in the proposed contents

Many methods have been proposed to generate 3D models from hand-drawn sketches. Igarashi et al. proposed a method for automatically generating a 3D shape from a hand-drawn outside contour[17]. Li et al. proposed a method for generating a more detailed 3D shape by drawing contours inside in addition to the outside contour[18]. Zeleznik et al. proposed a method for generating a 3D shape by analyzing the contour lines of a bird's eye view of an object[19]. Kondo et al. proposed a method for generating a three-dimensional shape by analyzing the shadow and contour of an object[20]. Eitz et al. proposed a sketch-based 3D shape modeling by retriving from a large amount of shape data[21].

In the content of this paper, the user creates a 3DCG model of the vehicle by drawing. The users of this content are mainly children, and it is necessary to be able to easily generate 3DCG models of cars while enjoying drawing on papers with pens. From the perspective of the contents in this paper, the methods of the references[17] and[18]are suitable for modeling curved surfaces and are not suitable for modeling vehicles. The methods of the references[19]and[20]require drawing while imagining a 3D shape, which is too difficult for children. The method of the reference[21]would be unrealistic because it requires a huge amount of data for the child's freely drawn shapes. As mentioned in the former section, the methods of the references[15]and[16]generate a 3DCG of a vehicle by drawing, but the texture and shape are incomplete. Also, drawing only the side view of a car vehicle may not give enough satisfaction.

In this paper, the author considered the balance between the enjoyment of drawing and the ability for children to easily generate a 3DCG of vehicle. That is, the shape and color of the side view of a vehicle can be freely drawn, and in addition, the front or top surfaces of the vehicle can be drawn. Then, it is possible to generate 3DCG of various vehicles by only using the pictures.



(a) train (b) car

Fig. 2 The template sheets for drawing vehicles

3.2 Drawing pictures of vehicles

In the contents of this paper, the user draws a picture of a train or a car with color pens on a paper. The users are assumed to be children including infants. Thus, the author has prepared template sheets on which the minimum elements necessary for drawing a train or a car are drawn beforehand (**Fig. 2**). It is not necessary to use the template as long as the positional relationship between the side view and the front view in drawing a train or the side view and the top view in drawing a car is matched. Details are described in the next section.

In the case of a train, the user draws a train seen from the side on the right and seen from the front on the left of the template sheet. In the side view area of the template sheet, the rough shape of the train and the wheels are drawn beforehand as a guide. In the front view area, a rectangle is drawn beforehand. Users who are not good at drawing can draw the appropriate train by drawing the outer shape of the train as well as the guide. Users who are good at drawing can draw the side view of the train in any shape. The outer shape should be drawn with dark lines such as black or blue. This is to reliably extract the side view region, the front view region, and the top view region of a vehicle from the picture taken with a web camera, which is the first step of generating a 3DCG model described in the next section. The user can draw freely using various colored pens inside the train (**Fig. 3**).

In the case of a car, a user draws a picture of the car seen from the side and the top. The area for the side view and the area for the top view of a car are prepared in the template sheet. In the side view area of the template sheet, the bottom of the car and the wheels are drawn beforehand as a guide. In the top view area, a rectangle is drawn beforehand. The shape of the car is generally more varied than the train, so the guide of the template sheet is minimal. The user can draw various shapes of cars on the template sheet. The user can draw freely

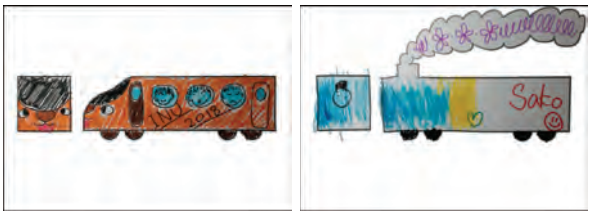


Fig. 3 Examples of pictures of train

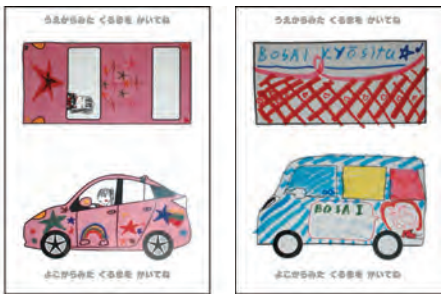


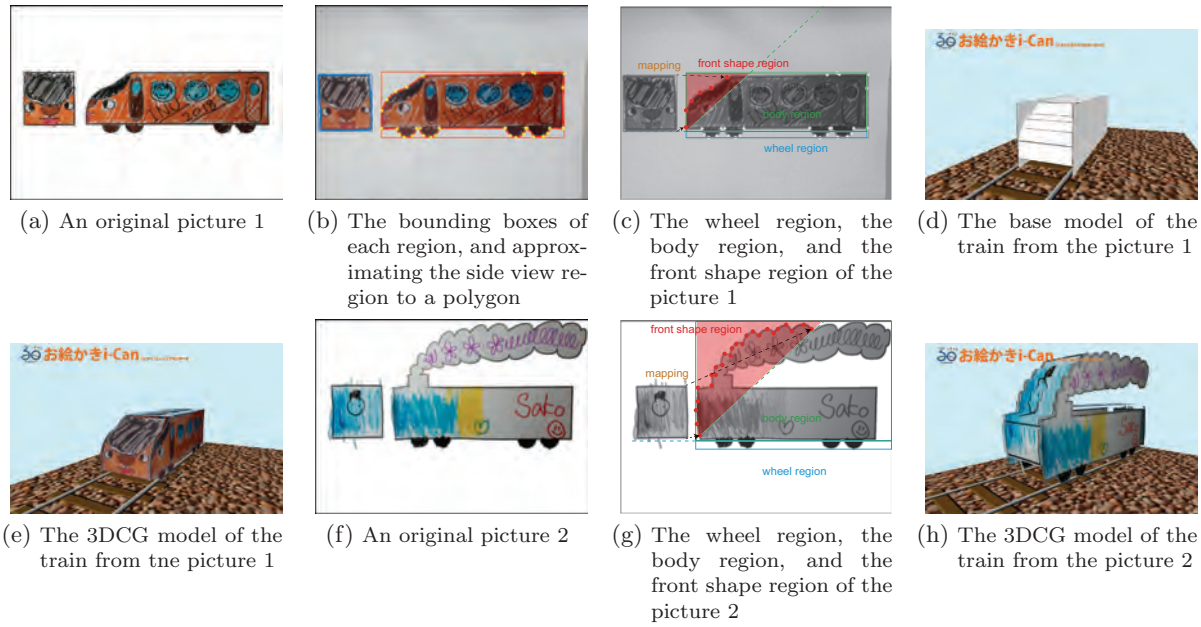
Fig. 4 Examples of pictures of car

using various colored pens inside the car (**Fig. 4**).

3.3 Generating 3DCG models of vehicles from pictures

After drawing a vehicle, the pitcture is taken with a web camera. The paper is placed horizontally when a train is drawn, and it is placed vertically when a car is drawn. The system applies adaptive thresholding to the picture to extract two closed regions from the picture with the area sizes larger than a threshold. Then the positional relationship between the two regions is examined by using the center of gravity of each region. If the regions are at the left and right of the picture, it is determined to be a train, and if they are above and below the picture, it is determined to be a car.

Figure 5 shows the process of generating a 3DCG model of a train from a picture. The region on the right is the side view of a train and the region on the left is the front view. First, the system finds the bounding boxes of each region (Fig. 5(b)). The region of the side view is also approximated to a polygon. Then, the side view is divided into the wheel region and the body region based on the lower point of the front view. In the side view, the front shape region is determined based on the straight line extending from the lower left corner of the body region (Fig. 5(c)). The slope of the straight line is 1.0, which was determined experimentally. The 3D shape of the front surface of the train is modeled by using vertices of the polygon inside the front shape region. Then, the bounding boxes of the side view are put on both sides of the front surface, and a rectangle for a roof is also put. As a result, a 3DCG base model of a train is generated



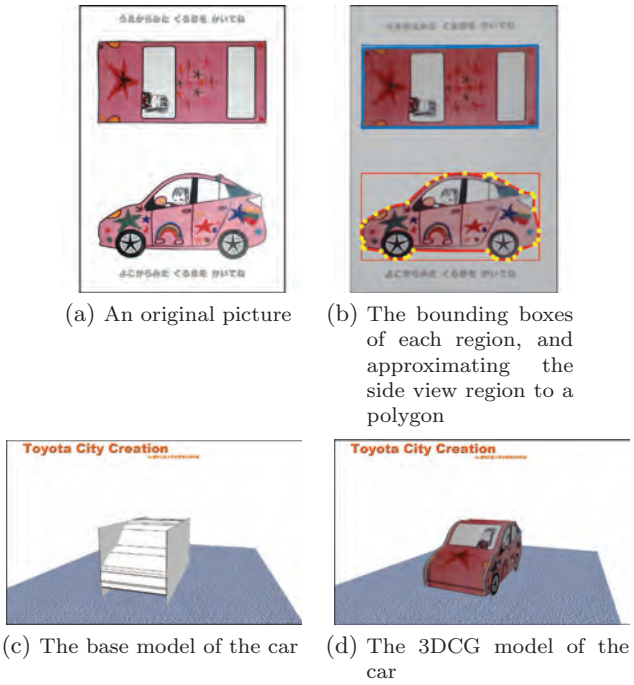
(a) An original picture 1
(b) The bounding boxes of each region, and approximating the side view region to a polygon
(c) The wheel region, the body region, and the front shape region of the picture 1
(d) The base model of the train from the picture 1
(e) The 3DCG model of the train from tne picture 1
(f) An original picture 2
(g) The wheel region, the body region, and the front shape region of the picture 2
(h) The 3DCG model of the train from the picture 2

Fig. 5 The process of generating a 3DCG model of a train from the picture

(Fig. 5(d)). The installation height of the roof is determined by the height of the inclusion rectangle in the front view. The width of the roof is determined by the width of the bounding box of the front view, and the length of the roof is determined by the length of the bounding boxes of the side view and the edge of the front shape region. Finally, the side view image is mapped to the rectangles of both sides of the base model as textures with masks. The 3DCG base model is cut out in the shape of the side view image of the train. The front view image is also scaled to the height of the front surface and mapped to the surface. As a result, a 3DCG model of the train is completed (Fig. 5(e)). Figure 5(f)–(h) is another example. Although the front shape region in Fig. 5(g) is larger than in the case of Fig. 5(b), the process is the same. The 3D shape of the front surface of the train is modeled by using vertices of the polygon inside the front shape region, and the front view image is scaled and mapped to the front surface. In this example, the front view image is elongated in the vertical direction. The upper and lower parts of the smoke have only sides.

In the case of a car, a 3DCG model is generated in almost the same process as for a train (**Fig. 6**). There is no setting of the front shape region like a train, and the shapes of the front, top, and rear surfaces of the model are generated based on the polygon that approximates the side view. Then the top view image is scaled and mapped onto the front-top-back model by parallel projection.

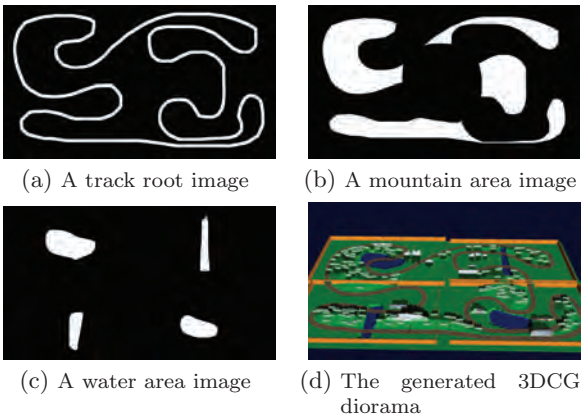
In the method proposed in this paper for generating a 3D model of a vehicle from a drawing, it is assumed that



(a) An original picture
(b) The bounding boxes of each region, and approximating the side view region to a polygon
(c) The base model of the car
(d) The 3DCG model of the car

Fig. 6 The process of generating a 3DCG model of a car from the picture

the side view and the front view (the top view) of a vehicle are arranged horizontally when the picture is taken by a web camera. In the case of a train, the body region and the wheel region are divided based on the positional relationship between the side view and the front view. Therefore, if the picture is tilted greatly when taking with a web camera, an appropriate 3DCG model of vehicle may not be generated.



(a) A track root image (b) A mountain area image

(c) A water area image (d) The generated 3DCG diorama

Fig. 7 Generating the 3DCG diorama

4. Displaying a 3DCG Diorama and Vehicles Running in It

4.1 Generating a 3DCG diorama

The 3DCG display system of our content has a 3DCG diorama for running 3DCG models of vehicles. The terrain of the 3DCG diorama is composed of a two-dimensional grid (128×72), and the geographical feature of the 3DCG diorama and the color are automatically generated by giving binary images which are showing the track, the mountain area, and the water area, respectively (**Fig. 7**). The altitude of the mountain area is determined based on the Euclidean distance transformation of the mountain area image. CG trees are automatically placed randomly in the mountain area. CG buildings are placed manually. The track is approximated as equilateral polygons for applying track textures and guiding vehicles.

The 3DCG display system receives 3DCG model data of each vehicle generated by the 3DCG model generation system via network. In this implementation, the author used the standard file sharing function of OSX to receive 3DCG model data via network. Each data is composed of the polygon of outer shape of a vehicle's side view and texture images (the side view, the front view, the top view).

Each time the data is received, a 3DCG model of a vehicle is reproduced and placed at a point on the track of the 3DCG diorama, and starts moving. The speed of the vehicle is determined by the direction of the side of the polygon approximating the track at that point. Thus the vehicle moves along the track. In the case of the train, a two-car train is realized by arranging two 3DCG train models in different directions (**Fig. 8**).

Multiple vehicles can run on the track at the same time. In this implementation, up to 50 vehicles can run at the same time. Each car is composed of about 200 triangles, and the diorama is composed of about 36,000 triangles. An ordinary PC (MacPro / MacBookPro) can render a 3DCG scene composed of these data in real time (30fps).

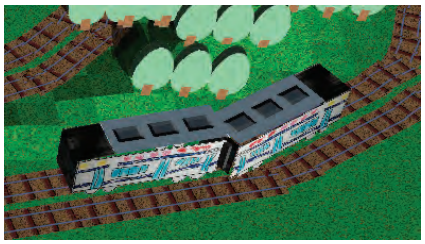


Fig. 8 A two-car train

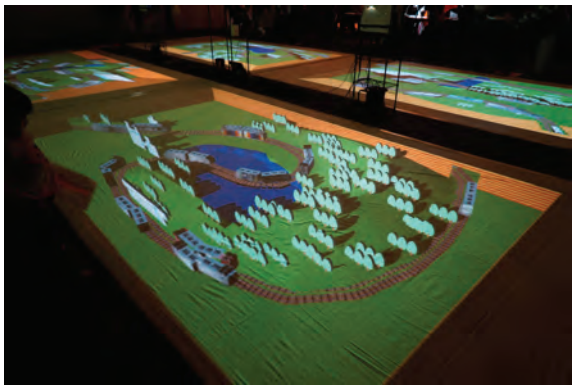


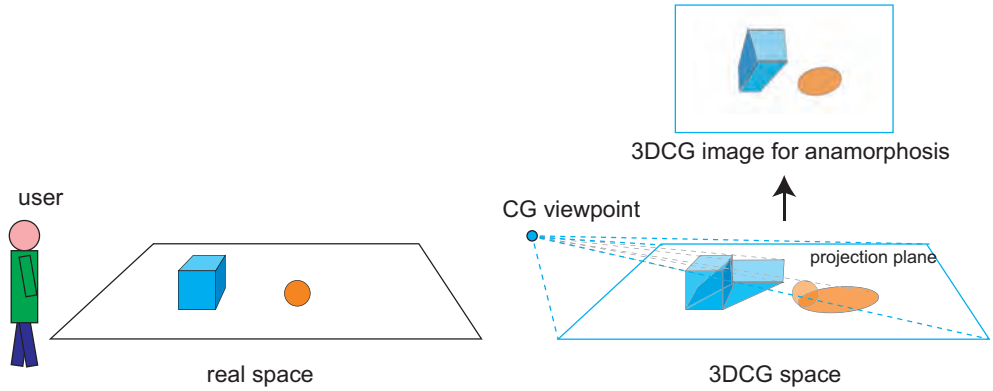
Fig. 9 The whole 3DCG diorama divided into four areas and projected by four projectors

4.2 Three-dimensional displaying of the 3DCG diorama

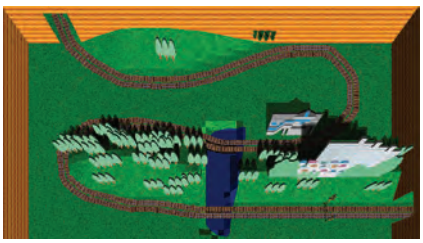
The image of the 3DCD diorama is projected by five projectors connected to the display system. Four projectors are used to project 3DCG images on a floor. The whole 3DCG diorama is divided into four areas, each of which is individually projected by the four projectors (**Fig. 9**). The size of the image by one projector is about 460×260 (cm).

At this time, an anamorphosis method: one of methods for trick art is applied to the 3DCG diorama images projected on the floor. A picture drawn based on the anamorphosis method looks distorted when viewed from the front, but it can be properly observed when viewed from a specific place and it can give the observer a three-dimensional effect. The system of this paper realizes the three-dimensional effect on the 3DCG diorama by applying the anamorphosis method.

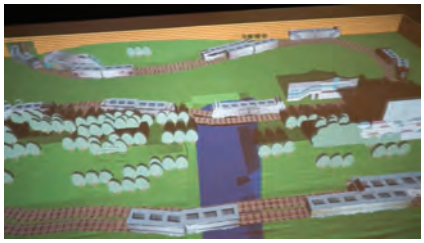
Figure 10 (a) shows the process of generating a 3DCG image with a principle of anamorphosis. To generate an image with anamorphosis from a 3DCG model, the viewpoint for observation and the screen in the real space are determined in advance, and the CG viewpoint and projection plane that reproduce them are set in a 3DCG space.



(a) The process of generating a 3DCG image with a principle of anamorphosis



(b) An image for projection



(c) Seeing the projected 3DCG diorama from the specific place

Fig. 10 Seeing the 3DCG diorama three-dimensionally based on the principle of the trick art

Fig. 11 Scene from a cockpit of a car

Then, a 3D model is placed in the 3DCG space and a CG image is generated. The generated image is distorted, but when displayed on a predetermined screen in the real space and observed from a predetermined viewpoint, the image is correctly observed. As a result, the anamorphosis effect is obtained, and the CG image is observed stereoscopically.

In this implementation, four viewpoints in the real space to observe each floor image are determined in advance. The system reproduces a projection plane and a CG viewpoint in the 3DCG space as well as the relationship between each projection area and each viewpoint in the real space, and generates 3DCG diorama images for projection (Fig. 10(b)). When a user sees the 3DCG diorama images projected on the floor from the predetermined viewpoints for each image, the images are observed as if the 3DCG diorama were built three-dimensionally on the floor by the anamorphosis method (Fig. 10(c)). The trains and cars are also observed as if they were actually running in the 3DCG diorama. Note that, the four viewpoints are set to a height of 110 (cm) on the assumption that children will observe them. This was decided based on the average height of the 3rd grade elementary school students is about 128(cm)[22], and the eye position when looking at the floor is 15–20 cm below the head top in our experiment.

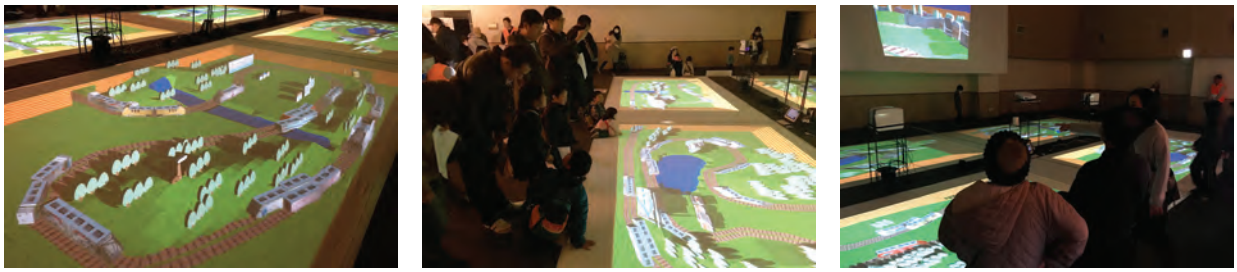
Another projector is used to project the whole 3DCG diorama image on a wall. It is also used to project images seen from the cockpit of the train or the car running in the 3DCG diorama (**Fig. 11**). With these images, the user can enjoy the atmosphere running in the 3DCG diorama as a driver.

4.3 Interaction

The 3DCG display system can interact with the user's action. A Kinect sensor is installed at the center of the four images projected on the floor, and the system acquires the positions and hand movements of users within a certain area around the projection images.

When user shakes the hand, objects are generated and thrown from the corresponding position in the 3DCG diorama. In this implementation, users can fly petals interactively to make flowers bloom in the 3DCG diorama because the theme of the event is for recovery from disaster. As a result, the user can feel as if flying objects to the 3D diorama.

(a) Children drawing pictures of trains (b) Taking a hand-drawn picture of a train with a camera and generating a 3DCG train (c) A 3DCG train appearing in the 3DCG diorama from above



(d) 3DCG trains running in the 3DCG diorama (e) Visitors observing the 3DCG diorama (f) Visitors watching the driver's view image

Fig. 12 Exhibiting "I CAN DRAW" at an event hosted by a railway company

5. Exhibiting Contents in Events

The author implemented the 3DCG modeling system and the 3DCG diorama display system for two contents: "I CAN DRAW" and "Toyota City Creation". The 3DCG modeling system was implemented in C++ on Mac Book Air (2GHz Core i7) using OpenGL and OpenCV. The resolution of the web camera is 1920×1080 (pixels).The 3DCG diorama display system was implemented in C++ on MacPro (3.5GHz 6-core Xeon E5) using OpenGL and OpenCV. The reason of using MacPro was to connect five projectors.

The author exhibited the content "I CAN DRAW" for drawing and generating trains at an event hosted by a railway company for the company's 30th anniversary held in Toyota, Japan on February 3, 2018. **Figure 12** shows the event, and **Fig. 13** shows examples of 3DCG models of trains generated by children. About 350 children created 3DCG models of trains by drawing pictures of train. The children who experienced it seemed to be very happy while being surprised that 3DCG models of trains were generated from pictures they drew. Then, while watching the 3DCG diorama, the children looked for the trains generated from their pictures with their family, and they were delighted and chased when they found their trains. Some children walked around the diorama, following the trains they had generated.

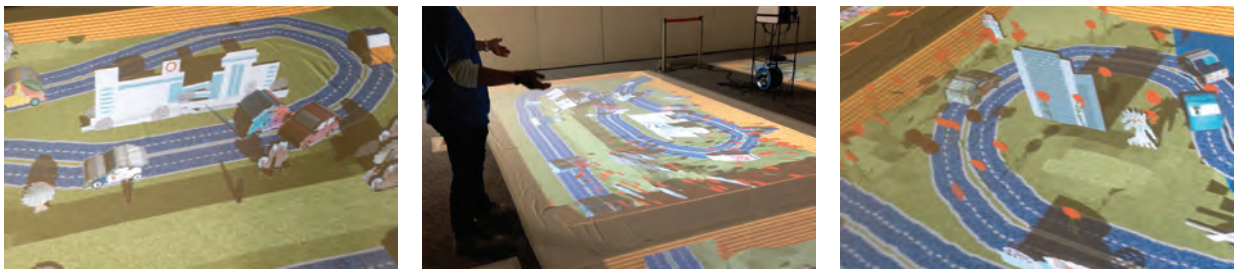
The author received a report from CBC Creation, Inc.: a management of this event that the railway company was pleased that the event was a great success with a large number of children participating.

Fig. 13 3DCG models of trains generated by children

The author exhibited the content "Toyota City Creation" for drawing and generating cars at an event hosted by a an automobile company in Toyota, Japan on May 13, 2018 (**Fig. 14**). About 330 children created 3DCG models of cars by drawing pictures of car. The children drew a wide variety of cars, and the 3DCG models of cars generated from the pictures were also very unique (**Fig. 15**). Since a lecture on fish was held during the same event, many 3DCG cars with fish motifs were created. In the interaction of flying petals to make the 3DCG diorama bloom, not only children but also many adults enjoyed this interaction.

The author received a report from CBC Creation, Inc. that the automobile company was pleased that many employees of the company participated in the event with

(a) Children drawing pictures of cars (b) Taking a hand-drawn picture of a car with a camera and generating a 3DCG car (c) The 3DCG diorama projected with five projectors on a floor and on a wall



(d) 3DCG cars running in the 3DCG diorama (e) A user interacting with the 3DCG diorama (f) Flowers that bloomed in 3DCG diorama due to interaction with users

Fig. 14 Exhibiting "Toyota City Creation" at an event hosted by an automobile company

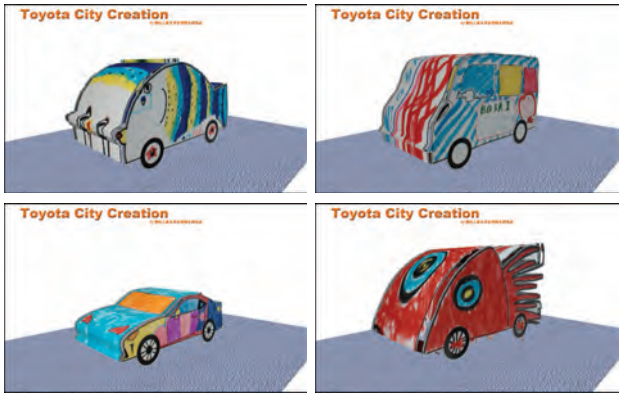


Fig. 15 3DCG models of cars generated by children

Fig. 16 "Rakugaki Cruise"

their children and enjoyed drawing.

The vehicle 3DCG generation system was also used in another content "Rakugaki Cruise" at a commercial event held at a shopping mall in Okinawa from February 28 to April 12, 2020[23]. The content itself is an extension of the two contents introduced above, such as being able to create a 3DCG city by drawing, and being able to immerse the 3DCG scene, but the part of generating a 3DCG vehicle is the same (**Fig. 16**). The method of generating a 3DCG model of a vehicle from a drawing proposed in this paper has worked successfully for over a month at the commercial event, over 500 people generated vehicles from their drawing, and the content has been very well received by visitors and event organizers. Therefore, it can be said that the proposed method is useful sufficiently.

6. Conclusion

In this paper, the author developed a method to generate 3DCG models of trains and cars by drawing their simple pictures with pens on papers. The author also developed a system to see the generated 3DCG models of trains and cars running in a 3DCG diorama three-dimensionally. The author created digital contents by applying the proposed methods, and exhibited them at events hosted by railway companies and automobile companies. Many children enjoyed the contents.

In the contents of this paper, participants can draw pictures of cars and run them in the 3DCG diorama, but they cannot create the 3DCG diorama itself. As mentioned at the end of section 5, "Rakugaki Cruise" has realized that participants can draw and create buildings

in the 3DCG diorama, and the author would like to develop a drawing interface that allows participants to draw and generate terrain of the 3DCG diorama.

To realize more complex 3DCG models such as airplanes, buildings and characters, and more complex interactions are future works.

Acknowledgements

I would like to thank CBC Creation, Inc. for supporting the exhibition of the two contents developed by us at the events.

References

1) Ikebana Sogetsu flow × FLOWERS BY NAKED Live Performance, Old Imabashi ward, https://flowers.naked.works/event/4285/ (2019).
2) MOMENT FACTORY: TABEGAMI SAMA, https://momentfactory.com/work/all/all/tabegami-sama (2017).
3) SIRO-A: TECHNO CIRCUS, https://www.technocircus.tokyo/home-en (2019).
4) teamLab: “Sketch Aquarium”, https://www.teamlab.art/w/aquarium/ (2013).
5) S. Mizuno, M. Isoda, R. Ito, M. Okamoto, S. Sugiura, M. Kondo, Y. Nakatani, M. Hirose: “Sketch Dance Stage”, Proc. of SIGGRAPH 2015, Posters (2015).
6) X. Cao: “Wonder Painter: Turn Anything into Animation”, Proc. of SIGGRAPH 2018, Real-Time Live! (2018).
7) TAKARA TOMY, PLA RAIL, https://www.takaratomy.co.jp/english/products/plarail/what/index.html (2020).
8) TAKARA TOMY, TOMICA, https://www.takaratomy.co.jp/english/products/tomica/index.html (2020).
9) Ludorum plc, CHUGGINGTON, https://www.chuggington.com/ (2020).
10) Gullane Limited, THOMAS & FRIENDS, https://www.thomasandfriends.com/en-us/ (2020).
11) Disney, Cars, https://cars.disney.com/ (2020).
12) East Japan Railway Culture Foundation, The Railway Museum, http://www.railway-museum.jp/ (2020).
13) Transport Heritage NSW Limited, NSW Rail Museum, https://www.nswrailmuseum.com.au/ (2020).
14) Toyota Motor Corporation, Toyota Automobile Museum, https://www.toyota.co.jp/Museum/english/ (2020).
15) RICOH, Paper Train & Paper Racer, http://www.ricoh.co.jp/rental/paper_app/ (2017).
16) Little Planet: SKETCH RACING, http://litpla.com/attraction/sketch_racing/ (2019).
17) T. Igarashi, S. Matsuoka, H. Tanaka: “Teddy: A Sketching Interface for 3D Freeform Design”, Proc. of SIGGRAPH ’99, pp. 409–416 (1999).
18) C. Li, H Pan, Y. Liu, X. Tong, A. Sheffer, W. Wang: “Robust Flow-Guided Neural Prediction for Sketch-Based Freeform Surface Modeling”, ACM Transactions on Graphics, Vol. 37, No. 6, Article 238 (2018).
19) R. C. Zeleznik, K. P. Herndon. J. F. Hughes: “SKETCH: An Interface for Sketching 3D Scenes”, Proc. of SIGGRAPH ’96, pp. 136–170 (1996).
20) K. Kondo, H. Shizuka, W. Liu, K. Matsuda: “A Sketch Interpreter System with Shading and Cross Section Lines”, Journal for Geometry and Graphics, Vol. 9, No. 2, pp. 177–189 (2005).
21) M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, M. Alexa: “Sketch-Based Shape Retrieval”, ACM Transactions on Graphics, Vol. 31, No. 4, Article 31 (2012).
22) Ministry of Education, Culture, Sports, Science and Technology: “Annual Report of School Health Statistics Research 2019”, https://warp.ndl.go.jp/info:ndljp/pid/11293659/www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2019/03/25/1411703_03.pdf (2019).
23) S. Mizuno: “Proposal and Development of a Content to Immerse and Move in a Hand-drawn 3DCG space”, IPSJ SIG Technical Report, Vol. 2020-DCC-25, No. 2 (2020).

(Received March 9, 2020)
(Revised August 9, 2020)

Shinji MIZUNO (*Member*)
He received the Ph.D degree from Nagoya University in 1999. He was a Research Associate at Toyohashi University of Technology from 2000 to 2009. He is currently a professor at Faculty of Information Science, Aichi Institute of Technology.His research interests include computer graphics, image processing, virtual reality, and interactive arts. He is a member of IIEEJ, IPSJ, the Society of Art and Science, VRSJ, and ACM SIGGRAPH.

IIEEJ System Development Paper -- Special Issue on Extended Papers Presented in IEVC2019 Part II--

Cooperative E-learning Applications Based on HTML-5 Canvas for Japanese Classical Literature Education

Eri YOKOYAMA†(*Member*), Hiroshi SUNAGA†, Makoto J. HIRAYAMA†(*Fellow*)

† Osaka Institute of Technology

<**Summary**> This paper proposes two e-learning applications specially designed for classical Japanese literature classes. The first one is a groupware allowing users to put comments on the handscroll images where a part of the handscroll is shown as one scene, and users can put memo cards on it and the scene with them moves when scrolled. The devised point of the application is that message cards can be placed in any point of the handscroll, and can be moved and modified, and the processed data can be stored in the database. The other is a jigsaw puzzle game using classical literature images. The shape of each piece is same and rectangular, but users must look at the detail of each piece to accomplish the puzzle game, and it helps the students watch the literature images seriously. Students who have actually used these applications say they have been helpful in learning literature and it can be said that they effectively work to help get unmotivated students interested in literature classes. It was also acknowledged that they have enhanced their interest in programming techniques as well, through playing these games.

Keywords: HTML-5, e-learning, literature application, handscrolls and paintings on a folding screen

1. Introduction

In the faculty of Information Science and Technology in Osaka Institute of Technology, as the social sciences and humanities field in the general education course, there are two literature classes, namely "Elementary Literature" and "Literature with Visual Representation". Their syllabi describe that these literature courses are linked tightly with the Information Technology as the courses of this faculty and aim at letting students feel literature familiar with them.

By introducing Web/smartphone applications into the classes, it is expected to enhance students' concentration on historical content and to further understand the meaning and backgrounds of literature works through discussion and mutual communications by group working. The content includes illustrated handscrolls and pictures drawn on a folding screen. They are included into applications and manipulated by students.

The main purpose of this research is applying methods for increasing students' interests and motivations to learn Japanese classical literatures. Since most of students in science and engineering departments are not familiar with studies in humanities and have little interests and motivations to learn. We are trying to increase students' interests and motivations to learn Japanese classical literature by developing and using cooperative e-learning applications related to edutainment technologies.

This paper proposes e-learning applications to activate the classes in literature. We also expect a synergistic effect, where students get close to programming or software development for their forthcoming graduation researches through the manipulation of these applications. The applications are based on the computing services framework[1], where Cloud/Web services, existing applications, and software components are combined and necessary parts are programmed to satisfy the requirements[2)–4)]. This paper proposes such e-learning applications, and shows the effects of introducing them to the literature classes. The rest of this paper is structured as follows. Section 2 discusses requirements and application images. Section 3 describes implementation techniques and section 4 shows the effects of introducing these applications into literature classes. Finally, section 5 concludes this paper.

2. Requirements

2.1 Features of literature classes

"Elementary Literature" deals with both proses and verses by using image content, video content, and printed images as well as explanation by voice. Here, by overlook-

ing at works from the ancient/medieval times to the Edo-era, "The Tale of Genji (Genji Monogatari)" and other masterpieces such as "Ogura Anthology of One Hundred Tanka by One Hundred Poets (Hyakunin Isshu)", students are led to understanding of the backgrounds or situations of the story. "Literature with Visual Representation" uses image content from a literature database named "The Largest Digital Reference on Pre-Modern Japanese Works" of "National Institute of Japanese Literature"[5] or the web-site of "National Diet Library Digital Collections"[6]. In particular, it is important for students working on a printed image, e.g., an illustrated handscroll, to add text or information on it.

We have changed these analogue ways to digital environments. Image content has been included in Web/smartphone applications to be manipulated and given new values by students. Also students very carefully study image content in the form of group working or game. Hopefully, they may learn basic knowledge and understand what ancient people think of through visual content and interesting way of learning avoiding excessive explanation only on original texts. Among various e-learning applications we have developed, this paper deals with two applications that are excellent in terms of questionnaire results as shown in section 4, namely a group-work application using handscrolls and a puzzle game using folding screens.

2.2 Current situations of classical literature educations

The reason why we have developed and introduced such e-learning applications and settled the syllabi is that we would like to propose a solution to improve the current miserable situations of Japanese classical literature educations. According to investigations performed by the National Institute for Educational Policy Research[7], only 29.3% of the junior high school students surveyed positively answered to the question "Do you like classical literature?". This was one of the questions in the questionnaire performed in 2013 on academic performance and learning situations. Also, to the question "Do you like classical literature and Chinese classics?", performed for high school students in 2005[8], 71.2% and 72.1% of them negatively answered respectively.

Then, according to this report[8], the main reason was the excessive teaching on basic and elementary knowledge and techniques, thus leading to boredom and dislike of learning. The solutions were as follows. Teachers should educate so that their students can touch how ancient people see, feel, think of things at that time by not only teaching original texts but also devising educational tools. Then, teachers should enhance and deepen this teaching method to augment the interest and motivation of the students.

Fig. 1 Group-work application on illustrated handscrolls

Considering these reports, we started to try the active usage of IT materials to the classes so that students may eventually acquire elementary knowledge and techniques for classical literatures with proactive manner.

2.3 Overview of applications

2.3.1 Group-work application using handscrolls

This application uses illustrated handscroll content, named "Shigisan Engi Emaki", offered in the "National Diet Library Digital Collections" site. "Shigisan Engi" means the history of Shigisan Chogosonshi-ji Temple in the Mt. Shigi. This content consists of three volumes, each of which has height of about 30cm and width of 800–1,400 cm. The site offers them as 26–39 jpg files.

As shown in **Fig. 1**, the application has the display area, a slider bar, text area and several buttons. In this implementation, the display size is 600×1,500 px. In case of Volume 1, 26 jpg files are combined by hand to one jpg file of a long length. The currently displayed image is changed in accordance with the slider value.

In the conventional work, handscrolls are ordinarily displayed by using buttons[9]and some pictures with long-length can be scrolled by slider or mouse operation[10]. In this paper, we devised to use message cards which can be placed in any point of the handscroll, and they can be moved and modified, and processed data can be stored in the database. Such functions are our originality.

If a user wants to write some text on the currently

displayed image, the user inputs text to the text area. By clicking a point on the display, a box area i.e., a message card is created at that point. The characters are vertically written. Even if the display area is changed by the slider, the relative position of the card is kept. The color of the card and text can be later modified by using buttons.

Each user can select a color, thus each card can be identified so that teacher can previously give basic information and students can add what they feel, think of, and want to show. This application can be extended so as to paste photo files.

To share information with users including the teacher, a special formatted record data are devised. Records are managed by a Web database server which provides services for recording, updating and retrieving. By these functions users can submit their own opinions or information, receive those of the others, and build up a consensus. Technically, the store/restore function and database connection function are provided. Also, a currently displayed image can be converted to a png file.

2.3.2 Puzzle game using folding screens

Pictures drawn on a folding screen for Genji Monogatari are transformed to a Jigsaw puzzle like game. **Fig. 2** shows the area for setting a question, and **Fig. 3** for the answering area. Note that jigsaw-cut is not used for each piece, but rectangle-cut is used for the easy implementation and image content itself is of the largest importance for education. An original image is separated into small pieces each of which is randomly disposed. There are many jigsaw puzzle games found in programming sites[11),12)], in which one picture file is divided into pieces and randomly disposed. In our application, as we attach importance to teaching and learning, the shape of each piece is set same (simply, square or rectangle) and let the users look into the details of image pieces to learn the historical and literal backgrounds of the era. A user touches a piece in the questioning area and sets it at a position in the answering area. If the user wants to change the position in the answering area, a touched piece can be set at another position. When completed, all positions are judged from the viewpoint of whether or not they are correct. In case of an incorrect answer, one piece can be put back.

Fig. 2 Jigsaw puzzle of Genji Monogatari pictures drawn on a folding screen for questioning

Fig. 3 Puzzle of Genji Monogatari Emaki for answering

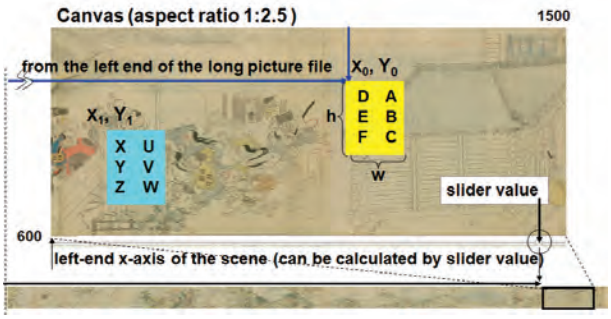


Fig. 4 Setting a part of file to canvas area

3. Techniques to Construct Applications

3.1 Group-work application using handscrolls

3.1.1 Display and slider control

The display area is constructed by using the HTML-5 Canvas element[13)], which is managed by JavaScript[14)] programs. A Canvas area is created by the canvas tag, whose height and width attributes are set as 600×1,500 px, respectively (**Fig. 4**) in this implementation. The original file size is 34,311×1,240 px, which has been made by adding separated jpg files, and a part of this file is cut and shown as a scene in accordance with the slider value.

The slider bar is created by an input tag with the "range" type. By moving the slider, a range value can be extracted, where the max and min values correspond to the left and right ends, respectively. By dividing the absolute value of the slider by the length between min and max values, a relative position can be determined.

In the figure, the aspect ratio of the display is 1:2.5, while that of the original file is 1:27.6. To show the piece of the picture file designated by the slider, the parameters of the "drawImage function" are set as follows. The y-axis value of the original file to be cut is 0. That of x-axis is the position calculated by the slider relative value. In the same way, the height is 1,240 px, and the width is 27.6 px. Each time a user moves the slider, the picture file is redrawn.

3.1.2 Message card setting

Figures and text are overwritten on the canvas, i.e., the currently displayed piece of the handscroll file. If a user clicks a point on the Canvas, x- and y-axes values are obtained. A rectangle is drawn by the "fillRect function" of JavaScript by setting the top-left coordinate of the card, and its width and height values as shown in **Fig. 4**. Text is drawn by the "fillText function". As this function only permit horizontal writing, we make a subroutine by which the given characters are divided to a single character and each is disposed vertically in the card area by calculating the number of characters and the card size.

The values of each card are kept as an object and then pushed into a list (an array of a variable size). In case of JavaScript, an object is defined as a function containing only variables without methods as shown in **Fig. 5**. This corresponds to a composite data type representing the properties of one message card.

3.1.3 Modification of message card

When the slider is moved and a new scene is drawn, all the records are extracted from the list, and all the message cards are redrawn in accordance with data in each record. The record contains the slider and x-axis value. The original handscroll file is cut to show the background scene designated by those values. Also, the record contains relative x-axis positions on this scene. By using them, all the message cards are overwritten on the scene.

If a user wants to move an existing card or modify current message or color, first the designated card must be identified. When a user clicks a positon in the display area, x- and y-axes coordinate values are obtained by the event listener function programmed for collision detection processing. This function successively compares the stored x/y coordinate values in each record with the clicked coordinate values. If the clicked coordinate values are within the range of "$x + width$" and "$y + height$", this message card is identified, that is, the detection is successful. If the user wants to modify the text or color, these data in the record are replaced from the older ones.

```
function record {
  the sequence number;
  the current display position data;//relative slider position
  the x/y coordinate in this display area;
  the width/height of the card;
  text message;//character sequence
  colour code//RGB vlue
 }
```

Fig. 5 JavaScript class for the record

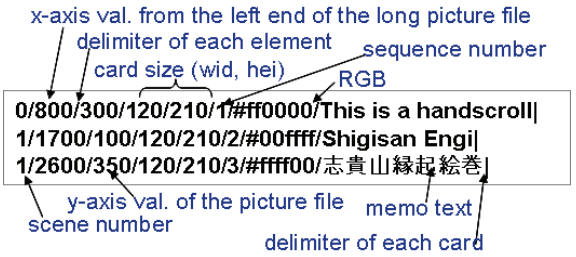


Fig. 6 Example of the special data format

```
{"data":[{"tablename":"shigisan_vol1","name":"guest1",
"pass":"123","key":"135","memo":"test","record"
:"1/7.5/600/400/50/100/Buddha statue/#ffff00|2/6.0/700/500
/80/100/local aristocracy/#00ff00|"}, {…}]}
```

Fig. 7 Example of the Json data

If the user wants to delete this card, this record is removed from the list. Also, the card position can be moved.

3.1.4 Application programming interface

As this application is driven by JavaScript programs on the browser, all the records will disappear if the browser is closed. To keep them, we provide a database function in a Web server and define Web APIs. If the user push the save button, the program sends the records to the database server by means of the Ajax[15] communication. The request is of the form of a URL, its parameters must be a character sequence.

Therefore we introduce a special data format for the Web access. All the records are extracted from the list, and every data element in them is separately read and concatenated to a long character sequence as shown in **Fig. 6**. This sequence is added to the URL with the argument name (record).

The example of Fig. 6 includes three records. If we explain the first one, it has values "0, 800, 300, 120, 210, 1, 'ff0000', 'This is a handscroll' ", which mean the scene number, the x-axis value from the lest end of the file, the y-axis value, the width of the card, the height of the card, the sequence number, the RGB value, and the message, respectively.

When this sequence is transferred to the server, it is kept in the data base as one record value. When it is

returned to the browser from the server, the handling program decomposes it by '|', and further decomposes each separated element by ',' to data values, RGB and a message. These are used to replicate the scene on the canvas.

The server provides functions to insert, update, delete and read the designated table. This server provides REST[16] interface, and uses jersey[17] as the library. As for the database, sqlite[18] is used because it is light-weighted. To the read request, the response is formed in the Json[19] format. An example is shown in **Fig. 7**.

3.1.5 Character and scene database

To let students understand scenes more deeply, we introduce a database explaining major characters, animals, objects and landscapes, as shown in **Fig. 8**. Each picture is cut from the handscroll and visually processed so as to become clearer by using Photoshop. Also, each is attached metadata explaining the name of the person or object, or the scene with historical/literary meanings. The keyword category search function is provided as well as a keyword search function.

As shown in **Fig. 9**, the picture files can be displayed. To show the contents in accordance with the user device, the Responsive Web Design (RWD) approach is introduced. The layout employs a fluid and proportion-based grid structure, which is changed by judging the viewing environment through CSS3 media queries. Pictures form up in four columns when viewed in a PC display, three columns for a tablet, and two for a smartphone.

Moreover, voice input, voice recognition and text-to-speech functions are implemented by using Amazon Alexa[20] and Amazon Echo[21]. These functions improve usability and aspects of entertainment.

3.2 Puzzle game application

3.2.1 Picture file handling

A jigsaw-like puzzle is consisting of many pieces each of which corresponds to the answer picture. From the programming viewpoint, it is very easy if separated files for each part are previously prepared, while the separation work is time-consuming and boresome. We propose a method using only the original picture file.

As shown in the lower part of **Fig. 10**, a picture file is managed in a grid form (6×3 in this example), each piece of which is given an index (i, j) and identified by using this index. From this original picture, each piece is dynamically copied to the answer area as shown in the right part of Fig. 10. As the HTML Canvas capabilities include the partial copy function as discussed in section 3.1.1, at the time of game initialization, all the pieces of the original file are copied and positioned randomly on the question grid area.

More specifically, each pieces of the original file is given a sequence number, $0 \cdots 17$ in this case. This sequence number corresponds to grid coordinate $(0,0) \cdots (i,j) \cdots$ of the original file. The numbers, $0 \cdots 17$, are randomly sorted as an array and kept until the end of the game. This array is used as the conversion array, and copied

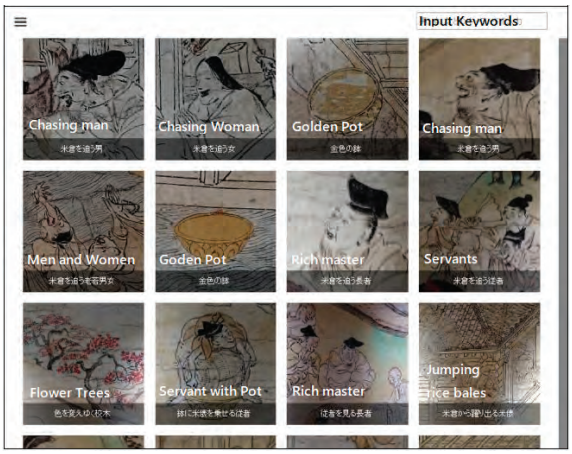


Fig. 8 Display of major element in handscroll

Fig. 9 Smartphone interface

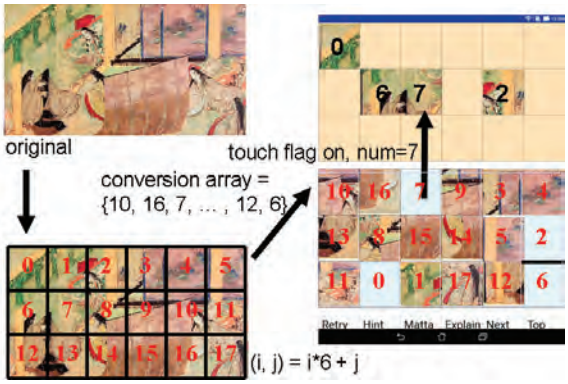


Fig. 10 Puzzle handling

as the question array. In the question area, each partial file of the piece is drawn in accordance with this order. This game is implemented for both the Web browser and Android. The drawing method for the former is basically same as described in section 3.1.1.

In case of Android, the Bitmap class is prepared for the file handling. The Bitmap instantiated from a file can be cut partially by designating the pixel coordinate of the left-top, *width* (px), and *height* (px). This partial bitmap can be placed to any position by designating these parameters. For example, assume that grid coordinate (0, 0) is given sequence number 11, which is randomly generated. Number 11 corresponds to grid coordinate (1, 5), when the image is composed of 6×3 grids. The original position is designated by the left-top coordinate, i.e., $x = 0$ (px) and $y = 0$ (px), the grid width, and the grid height, and thus a partial bitmap is created. The new position is designated by the coordinate of the new position, $x = 5 \times width$ (px), $y = 1 \times height$ (px), the grid width, and the grid height. In this way, each piece is set randomly to the question area from a single original picture file.

3.2.2 Game processing

The answer area is managed in the same way. Also, the answer area is given grid coordinate $(0, 0) \cdots (i, j) \cdots$ from top-left to bottom-right. The initial values of the question area are the same as the conversion array. So the randomly positioned pieces (image) are drawn in the question area. On the other hand, all the initial value of the answer area are set to -999 (out-of-scope value), which means that no image is drawn in the area.

If a user touches a position on the question area, the event listener informs its handler program of the touched coordinate values. The grid coordinate values of i and j will be calculated by using *width* and *height* of the piece. At this time, the piece number as well as the touched status are memorized, and the sequence number of this piece is given -999. By redrawing the question area based on the question array, the touched position is not shown.

Then the user touches a position in the answer are, and the answer array is given the piece number at this position. By redrawing the answer area based on the answer array, the selected image is shown.

When all the pieces are set on the answer area, the answer array is compared with the conversion array. If all the numbers match correctly, the game is over. If there is more than one mismatch, the user can reconsider. The last piece is brought back to the question area. As each piece in the answer area can be moved by a touch processing event handler, the user can rearrange the position to come closer to the right answer. In addition, explanation pages are prepared. These pages as well as the game itself help the students to more deeply understand the scenes, total story, and historical backgrounds. The game variations can easily be increased for teacher by utilizing the original image files and their explanation texts.

4. Evaluation of Applications

4.1 Effects of the applications

4.1.1 Group-work application

We introduced applications to “Elementary Literature” and “Literature with Visual Representation”. This application helps students to understand picture content deeply and express their findings about what the content means, what the characters are doing, what situations are described. This is a kind of active learning, where students try to find something important or even trivial in a spontaneous way and learn a research and opinion delivery process.

A student first searches for a character or a place he/she is interested in by scrolling the handscroll by using the slider or jump buttons. Then, he/she writes a comment about it, selects color, and attaches the message card on the display area. If he/she finds a new fact or has new opinion by investigating references or group discussion, he/she may use new color, which can record the transition of understanding. The teacher can show a specific scene and easily move to another scene. Also, the teacher can prepare what he/she want to tell in the form of message card. Formerly it was difficult to hand a paper handscroll of a long length or many sheets of cutouts to all the students but every student can manipulate the total part of the handscroll in a large class. In addition, it promotes remote studies by accessing to the server from outside of the class room.

We performed questionnaires survey on the handscroll application in January 2018 towards 219 students in the “Elementary Literature” class, and 42 students in the “Literature with Visual Representation” class. The students first used the application before receiving explanation about the handscroll, and then we performed the first survey. At the next class, we lectured the handscroll as an ordinary class, and after it we performed the second survey.

The first survey included two questions. The first one is whether or not the impression of each student about

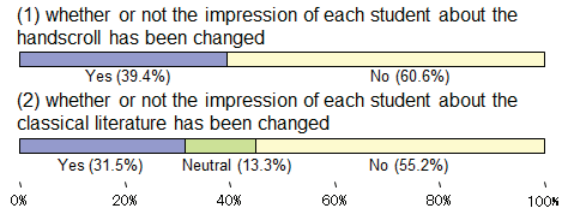


Fig. 11 Survey on effects of handscroll application

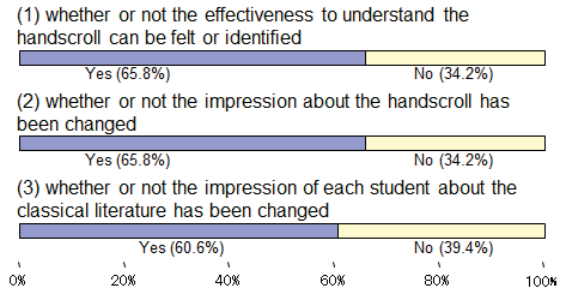


Fig. 12 Survey on effects of handscroll application in conjunction with lecture

the handscroll has been changed. As the second one, the same question to the classical literature was posed. As show in **Fig. 11**, these results mean that only the usage of the application did not lead to augmentation in the interests of the students.

The second survey examines the effects of the application in conjunction with the lecture. As show in **Fig. 12**, three questions were posed. The first one was about the effectiveness to understand the handscroll. The next one is whether or not the impression about the handscroll has been changed. The last one is whether or not the impression of each student about the classical literature has been changed. In comparison with the previous questions posed before the lecture, the results imply that the use of the application in conjunction with the lecture is effective to augment the interests and understandability of the students.

4.1.2 Folding screen puzzle application

We observed students in the two classes as described in section 4.1.1. The folding screen picture, "Genji Monogatari Emaki", is difficult to understand without knowing what the scene implies. By transforming the original picture to may pieces of puzzle, each student must concentrate on details of the scene, particularly the four borders, and come naturally to understand what the scene describes. As the Genji pictures have a characteristic, i.e., interior description, this application leads the students to understand deeply and effectively.

We also performed questionnaires on the jigsaw puzzle application in April 2019 towards 45 students in the

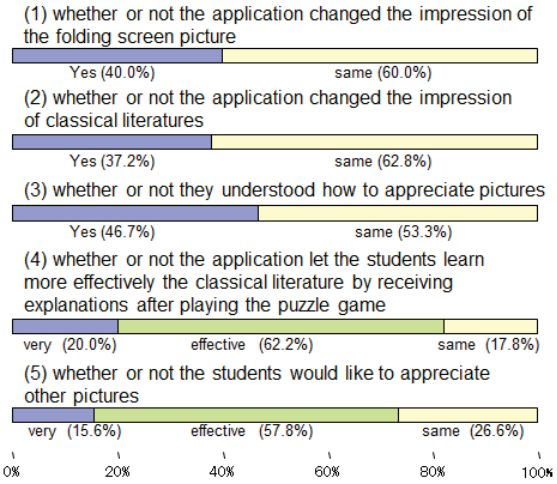


Fig. 13 Survey on effects of folding screen puzzle application

"Literature with Visual Representation" class, as shown in **Fig. 13**. The first question was whether or not the application changed the impression of the folding screen picture, and 40.0% answered positively that they became much more interested than before. The others were neutral, and there was no negative answer. The second question was whether or not the application changed the impression of classical literatures, and 37.8% answered positively and the others answered neutrally. The third question was whether or not they understood how to appreciate pictures, and 46.7% answered positively, while the others answered neutrally. The fourth question was whether or not the application let the students more effectively learn the classical literature by receiving explanations after playing the puzzle game, and 20.0% answered "very effective", 62.2% "effective", and the others answered the same. The fifth, final, question was whether or not the students would like to appreciate other pictures, and 15.6% answered very positively, 57.8% positively, and the others answered the same. By performing chi-square tests to these data, significances were validated for all data at 5% level.

Considering the answer of the fourth question, which obtained about 80% positive answers, to find neighboring pieces, the shape, tincture, or continuity of lines must be checked carefully and the situation must be understood. The students proceeded with this work with special concentrations and memorized pieces. This action very much helped the understanding of the teacher's explanations after the game. Some students added a comment that they could understand the structure of the old Japanese building and interior arts through this game. Also, the result of the fifth question shows more than 70% would

like to enthusiastically study other classical pictures. In the other questions, it seems that the use of this game had a level of effects while the attitude of about 60% of the students answered was the same.

With respect to free comments, there were answers to enhance the above mentioned results. Student answered that they would like to try more complicated puzzle, or that the game made students more familiar with classical literatures. These comments show that the game playing contributes to nourish the attitude of the students to appreciate the depth of the scene the picture implies.

4.1.3 Overall effects

As described in section 2.2, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) expects teachers to help their students better perceive how ancient people saw, felt, and thought in the past. Considering the above mentioned evaluation results, our applications would fill such requirements. There are three points we would like to emphasize.

(1) Lectures using the applications can shorten the distance between classical literature and students. By touching image content through tablet or Web applications, it is expected to motivate students who are reluctant to study the subject. In a lesson about traditional Japanese folding screens, for example, the teacher can focus on the part he/she attaches importance to without having to go into details about techniques to appreciate the works.

(2) The Genji puzzle application creates a traditional Japanese cultural space. It gives the students an opportunity to experience how ancient people might have seen, felt, and thought at the time, which the MEXT prioritizes for literature education.

(3) Each student can adaptively manipulate the applications in accordance at his/her own pace. In particular, the Shigisan group work application can show and deliver his/her opinions or awareness to the work by the message card function. As study progresses, such opinions can be revised or enhanced. Also, by the group work function, data can be shared in the class, while cooperation spirit is likely to be fostered in the process. Thus, our applications are expected to help students deepen their study in multiple ways.

4.2 Further study items

It is expected that the introduction of IT applications to literature classes will help the below-mentioned planning of desirable lectures for classical literature classes and lead to a future image of education fields in the digitalized environments. First, it can transform the teaching style from a text-reading-centered one to a kind of project-based one aimed at learning through studying on one 's own initiative. Secondly, from the IT literacy viewpoint, it helps students to acquire capabilities of extracting information from content and expressing what they understand and analyze. Finally, it cultivates their perceptiveness to connect rich heritages from classical literatures with present lives and information technologies.

5. Conclusion and Future Work

We have developed e-learning applications for classical literature classes and introduced into lectures. These help students to understand classical handscrolls and pictures on a folding screen more deeply and enjoyably. After all, students acquire precise knowledge, and discuss actively with each other, through the acquisition of the manipulation of the applications.

Technically, by using the HTML-5 and Android functions new types of teaching applications or self-learning applications have been created. The manipulation to the display image attracts attention to precise scenes, character actions, objects surrounding or possessed by them, and backgrounds. The dynamic behavior of our application motivates students to learn more deeply, think of details of the situation, and investigate the teaching materials without teacher's excessive inclusion, compared with the conventional applications dealing only statistic texts and images.

Effects to the classroom are firstly improvement in attitude to the learning of unmotivated students by the combination of ordinary explanation and the applications. In particular, as the handscroll application provides a try-and-error approach for both individual and group work, where students can freely think of comments, select color and dispose the message card to the most impressive position, they proceed with their own work or group work effectively.

We have developed and incorporated other applications. All of them have dynamic properties that lead to the synergistic effect of visual and auditory perception. For example, although the learning of Hyakunin Isshu (100 poems) is ordinarily apt to only use visual images and text, in our application images move in the screen and the poem is read aloud. Quest games are also implemented and used for further precise knowledge acquisition. We will also report effects of a variety of ap-

plications except for the two applications picked up this time.

References

1) H. Sunaga, A. Morikawa:"Memory Usage Reduction Method for Bitmap Drawing in Android Programming - Evacuation Guidance Application -," IEICE Technical Report SC2018-18(2018).
2) IEEE, IEEE Technical Committee on Services Computing, http://tab.computer.org/tcsvc/index.html (2020).
3) H. Sunaga, Y. Yamato, H. Ohnishi, M. Kaneko, M. Iio, M. Hirano: "Service Delivery Platform Architecture for the Next-Generation Network", Proc of ICIN 2008 International Conference on Intelligent Networks, Session 9-A (2008).
4) H. Sunaga, K. Nakajima:"Data Management for 3D Floor Plan Drawn on HTML-5 Canvas - Floor Plan Data Management," IEICE Technical Report SC2017-15 (2017).
5) National Institute of Japanese Literature, Database of Pre-Modern Japanese Works, https://kotenseki.nijl.ac.jp/ (2020).
6) Center for Open Data in the Humanities, Dataset of Japanese Classics, http://codh.rois.ac.jp/pmjt/ (2020).
7) Ministry of Education, Culture, Sports and Technology (MEXT) National Institute for Educational Research (NIER), National Assessment of Academic Ability 2013, https://www.nier.go.jp/13chousakekkahoukoku/data/research-report/13-questionnaire.pdf (2013)
8) Ministry of Education, Culture, Sports and Technology (MEXT), Survey on Operational Situation of Curriculums, Analysis and Improvement per Subject (Japanese and Integrated Japanese) in 2005 (2005).
9) National Diet Library, National Diet Library Digital Collections Legend of Mount Shigi, http://dl.ndl.go.jp/info:ndljp/pid/2574276,2574277,2574278 (2020).
10) web-emaki.com, Urashima Taro, http://web-emaki.com/ (2020).
11) Ponksoft, jQuery based Jigsaw Puzzle, http://ponk.jp/jquery/basic/jigsaw (2020).
12) ao-system, Online Jigsaw Puzzle Making, https://ao-system.net/jigsawpuzzle/ (2020).
13) WHATWG and W3C, HTML-5, http://www.HTML-5.jp/ (2020).
14) ECMA, JavaScript (ECMAScript), ECMA-262, Edtion 10 (2019).
15) W3C, Asynchronous JavaScript + XML, Ajax Tutorial," http://www.w3schools.com/ajax/default.asp (2020).
16) Java Community Process, The JavaTM API for RESTful Web Services, JSR-000311 JAX-RS (2020).
17) Eclipse Foundation, Jersey, http://jersey.java.net/ (2020).
18) Hipp, Wyrick & Company, Inc.(Hwaci), SQLite, http://www.sqlite.org (2020).
19) IETF, The JavaScript Object Notation (JSON) Data Interchange Format, RFC7159 (2014).
20) Amazon, Amazon Alexa, http://alexa.amazon.com/spa/index.html (2020).
21) Amazon, AmazonEcho, https://www.amazon.com/ (2020).

(Received September 12, 2019)
(Revised July 31, 2020)

Eri YOKOYAMA (*Member*)
She received Ph.D in Literature, from Nara Womens's University in 2016. Currently, she is an assitant professor at Faculty of Information and Science and Technology, Osaka Institute of Technology. Her research interest includes classical literature especially for Emaki (illustrated hand scrolls).

Hiroshi SUNAGA
He received the Ph.D. degree in information sciences fromTohoku University in 2002. Currently, he is a professor of Osaka Institute of Technology, where he has contributed research and development of Web and smartphone applications since 2009. Before then, he worked on developmental research of switching node systems, IP telephony technologies, and Peer-to-Peer communications in NTT. He is a member of IEICE and IEEE.

Makoto J. HIRAYAMA (*Fellow*)
He received B.E. and D.E. from Waseda University in 1985 and 1996 respectively, M.B.A. from Leicester University in 2004, and B.L.A. from the Open University of Japan in 2008. He worked at Hewlett-Packard Japan 1985–1994, also at ATR Auditory and Visual Perception Research Laboratories 1989–1992 and ATR Human Information Processing Laboratories 1992–1994, at Hewlett-Parckard Laboratories Japan 1994–1999, and at Kanazawa Institute of Technology 1999–2013. Since 2013, he is a professor at Osaka Institute of Technology. His research interest includes speech production and multimedia applications.

Upon the Special Issue on CG & Image Processing Technologies for Automation, Labor Saving and Empowerment

Editor: Masanori SEKINO
(Fuji Xerox Co., Ltd.)

A declining birthrate and an aging population are a common problem in developed countries. In order to solve the labor shortage, it is necessary not only to expand the workforce but also to improve labor productivity. Therefore, improvement in productivity is expected through automation and labor saving technologies using evolving image recognition technology, increasing information from IoT devices, and cheaper robots.

At the call for submission, we received 12 papers, 8 in Japanese and 4 in English. Concerning English papers, one of them is adopted and included in this issue, and other three papers are still under review. The fields covered in the submitted papers are diverse and interesting, such as defect detection at production sites, long-term care support, sports scene recognition, CG image conversion, and pathological image segmentation. Despite the variety of covered area, many of them discuss robustness against environmental changes and disturbances, and the stability of machine learning, and ensuring robustness. Therefore, stability seems to be a common issue in social implementation of such technology.

In our journal, we are soliciting “system development papers” and “practical papers”, which are the categories of the papers that make it easy to present the outcome of research focusing on the practical and applied aspects of technology. As the technology recruited in this special issue will well match with these paper categories, we hope your further submission of the papers related to this issue especially in these categories.

Finally, I would like to appreciate the reviewers and editors for careful review and all the efforts to improve the quality of papers. Also, I would to thank to the editorial committee members of IIEEJ and the staff at IIEEJ office for various kinds of support.

IIEEJ Paper -- Special Issue on CG & Image Processing Technologies for Automation, Labor Saving and Empowerment --

Bidirectional Mapping Augmentation Algorithm for Synthetic Images Based on Generative Adversarial Network

Haoqi GAO† (*Student Member*) , Koichi OGAWARA†

† The Faculty of System Engineering, Wakayama University

<**Summary**> In training deep neural networks for supervised learning tasks, we often use data augmentation methods to increase training dataset sizes. Furthermore, this technique is particularly useful when the size of the training datasets is small, such as when the content of the training datasets includes privacy issues that cannot be made public, or the categories in the train datasets are unbalanced. During the training process, small training datasets will lead to model overfitting. Nowadays, data augmentation methods using Generative Adversarial Network (GAN) and Neural Style showed to provide performance improvements for the task of supervised learning. However, the traditional GAN is easy to cause the network to collapse, which makes the generation process free and uncontrollable. Hence, it may cause the network model to fail to produce deterministic results, which makes the application limited. In this paper, we propose an improved GAN-based data augmentation method for image classification tasks. We compare our model with the latest GAN model, and the results show that our algorithm is effective. On generated images, when applying synthetic images to facial expression attribute classification task, our method achieves 72.5% accuracy rate on the FER2013 PrivateTest datasets and 71.2% accuracy rate on the FER2013 PublicTest datasets.

Keywords: data augmentation, generative adversarial network, neural style, image classification

1. Introduction

Nowadays, training models using neural networks usually require a large number of labeled datasets. However, the process of labeling data is definitely time-consuming and expensive, and manually labeling the data can easily lead to many errors. In contrast, synthetic datasets do not require manually annotated data, which reduces the burden of both the data collection and labeling process. Theoretically, we can synthesize an unlimited number of training datasets. In addition, the synthetic samples allow us to precisely control the rendering process of the images, such that the datasets have different characteristics. Nevertheless, how to narrow the "domain gap" between synthetic datasets and real-world datasets is still a major challenge.

Among them, face analysis has always been a research hotspot in the field of computer vision and neural networks due to its theoretical significance and great practical application value. While the technology of face research is mature, it is still definitely challenging for faces captured by a camera in an unconstrained real-world environment. Real-world face datasets usually face the following challenges: (1) The number of specific face datasets (such as face data with large offset angles or complex expressions) is not sufficient for the training of the model. (2) Due to privacy reasons, there are only a few public datasets available at present. (3) Preparation and labeling of training datasets usually require a great deal of time and expensive investment. In this way, according to the main challenges in the given research field, this paper introduces the topic of synthetic face datasets.

With the rapid development and advancement of graphic imaging technology, researchers have been using some 3D models[1] to generate synthetic datasets. For many tasks, synthetic datasets[2),3)] were used to minimize the cost and risk of training the network models. Nonetheless, it is still technically challenging to construct high-quality synthetic-to-real datasets. Though the model is well-trained, it is not well-suited for real-world test datasets. Briefly, if synthetic datasets are significantly different from real datasets, it will affect the decision made based on the model. Therefore, if the synthetic datasets are similar to the real datasets, the model trained with the synthetic datasets can be used more effectively on the real test datasets.

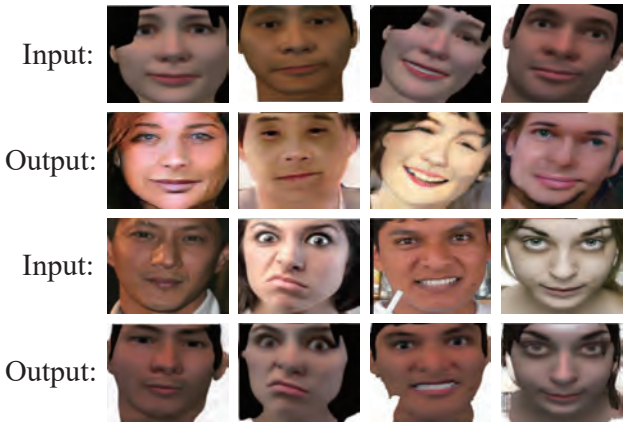


Fig. 1 Our experiment results

Our research plan considers the privacy of the face datasets. Through FaceGen[4], we can obtain large-scale datasets from various angles and various facial expresses. Nevertheless, even when FaceGen synthesized datasets achieve high accuracy in network training, it is not sufficient to validate on real-world test datasets, so we consider adding feature information of real datasets to synthetic datasets. This paper, however, is different from CycleGAN[5] which transfers from mode A to mode B. We hope that the generated face needs to maintain the characteristics of the synthetic face and the real face at the same time.

In our translation model, there will be two mapping functions. The top of **Fig.1** represents one mapping: the input is synthetic images, our network attempts to generate an image as real as possible. The bottom shows another mapping: with real-world images as input, our model generates a synthetic image that incorporates feature information of the input.

Our works and contributions will be summarized as follows:

1. Our network improves the sense of the reality of synthetic data and narrows the "domain gap" between synthetic datasets and real datasets.
2. We used FaceGen-based synthetic facial datasets with different facial appearances (like smiles, anger, fear, and neutral) as our input images. Our model is trained with these facial expression datasets separately. At test time, if you input an image, the output is a synthetic image with different facial expressions.

2. Related Work

Since we all know, the algorithms based on deep learning are more effective when providing more training datasets. Past studies have shown the effectiveness of data augmentation by performing minor modifications on training datasets (such as image cropping, rotation, and flipping)[6]. Up to now, Generative Adversarial Network (GAN)[7] and Neural-style transfer[8] techniques have been proposed as a powerful technique to do the data augmentation. In this section, we briefly review some representative works.

2.1 Traditional transformations

Traditional transformation is based on a series of well-known affine transformations to process training datasets, such as rotation, flipping, scaling, and distortion, as well as some simple image processing methods, such as lighting color transformation, contrast transformation, and adding noise[6]. We use these methods to generate new datasets from a limited number of datasets. This can alleviate the problem of overfitting of the neural network to a certain extent, and improve the generalization ability of the model. Nevertheless, compared to the original datasets, the increased datasets have not been fundamentally addressed by the problem of under-fitting of the datasets. Simultaneously, this augmentation method requires manually adjusting the transfer function and corresponding parameters, which are usually dependent on empirical knowledge. It is hard to achieve the best performance with those data augmentation methods.

2.2 Background on GANs and style transfer

Consider the limited problem of traditional transformation methods. Image synthesis is a relatively new research topic at present[9),10]. Generative Adversarial Nets (GANs) has been a powerful technique to generate new images for training[6]. GANs have been applied to a variety of tasks and have achieved remarkable results in all aspects of image generation[11),12].

As GANs have been proposed in 2014, there have been issues such as the difficulty of training, the difficulty of convergence, and the lack of diversity in the generated samples. Since then, many researchers[13)–16)] have tried to solve those problems and proposed improvements. Like W-GAN[13] uses the Earth-Mover instead of the Jensen-Shannon divergence to measure the distance between the real sample and the generated sample distribution. Conditional GAN (CGAN)[14] proposes to add additional information, which can be labels or other ancillary information. SimGAN[15] complements the adversarial loss with a self-regularization L1 loss that penalizes large changes between the synthetic and refined images to make the synthetic datasets more realistic and can be used to enrich

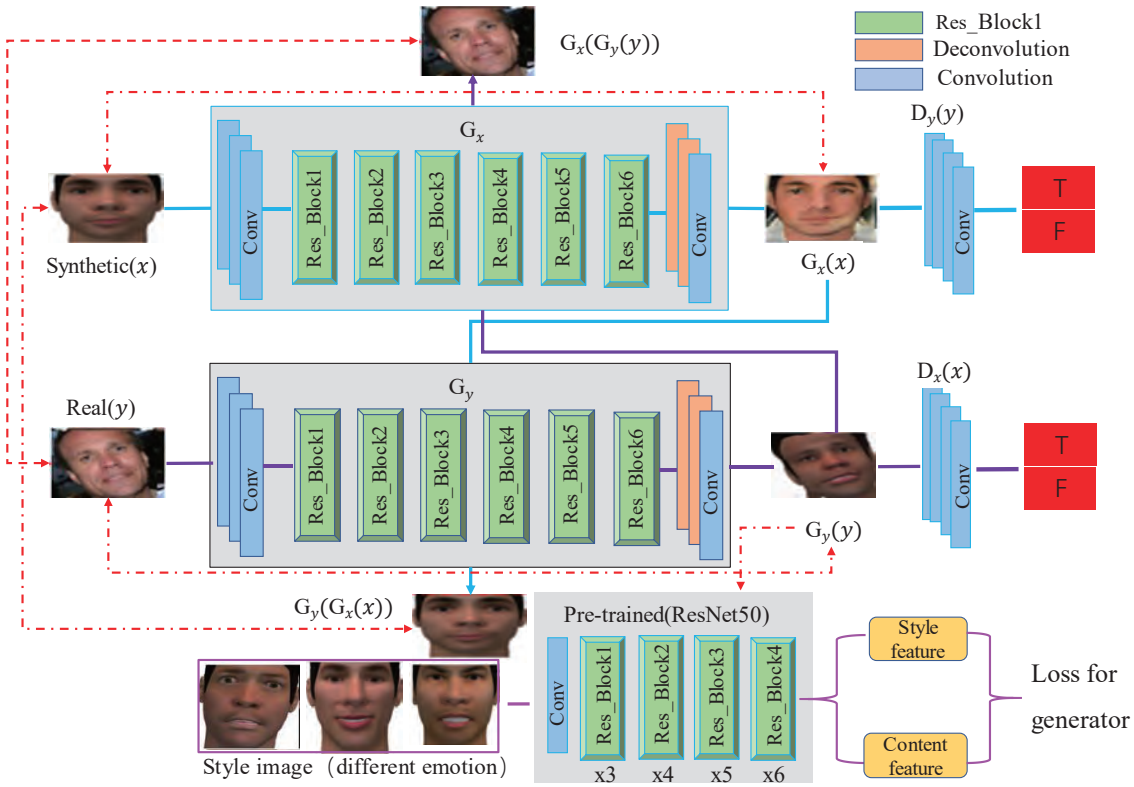


Fig. 2 The architecture of our proposed GAN

unlabeled real datasets. The pix2pix framework[16] proposes a generator to learn the mapping function between two paired images. CycleGAN[5] introduces a cycle generator confrontation network and puts forward a transfer between different image domains without the need for specific image pairs. Another hot and creative research area is that of style transfer[17], which combines style and content of an image skillfully through a neural network to form an interesting image. These depend on features extracted from a pre-trained neural network.

Furthermore, there has been a great deal of work in the area of generating synthetic face datasets by using GANs[18)–20]. These methods can address problems such as privacy of data, class imbalance, and denoise of images. For instance, Wan et al.[21] proposes FM-GAN for face generation of fine-grained multi-attribute. STGAN[22] treats the face synthesis problems as transferring styles of one face to another. PGGAN[23] is a breakthrough for synthesizing high-quality realistic face images. Paired-CycleGAN[24] relies on a cycle-consistent generative adversarial network method for editing a portrait photo. FaceID-GAN[25] used expression features to modify facial expression.

3. Method Overview

Figure 2 shows the overall architecture of our proposed network with both synthetic images and real images as input. Consider our datasets: synthetic images of $x \in X$ and real images of $y \in Y$ as well. In this work, we use two generators: G_x, G_y and two discriminators: D_x, D_y. Branches of both the synthetic and real image generation processes are specified by the blue and purple channels, respectively.

We adopt CycleGAN[5] network architecture and Wasserstein adversarial loss[13] to learn the mapping for the generator network. The generator network is composed of four convolutional layers, several residual blocks[26], and transpose convolutional layers. The number of residual blocks is determined by the size of the input image, for example, we use six blocks for 128×128 images and nine blocks for 256×256 and high-resolution training images. The discriminator network consists of four convolutional layers with leaky rectified linear units(ReLU). Researchers[5),16)] proposed a path-level discriminator architecture which has fewer parameters than a full image discriminator. And it can work on arbitrarily sized images in a fully convolutional fashion. For discriminator networks, we use PatchGANs[5),16),27),28)], which aim to classify whether 70×70 overlapping image patches are

real or fake. Besides that, we also use the instance normalization scheme[5),29)] for all layers of the generator and discriminator networks, with the exception of the input and output layer.

3.1 GAN review

The objective function of GAN is a minimax game of a two-player: the generator and the discriminator. We define P_r to be the data distribution, P_g to be the model distribution defined by $\tilde{x}_g = G(z)$, where z is the input to the generator. The discriminator network D adjusts its weights to reliably distinguish real data samples $x_r \sim P_r$ from fake data samples $\tilde{x}$, randomly sampled from some distribution P_g, via the generator network. The generator network G adjusts its weights to fool D. The two networks are trained iteratively using a loss function given by:

$$L_{GAN} = \min_G \max_D E_{x_r \sim P_r}[log(D(x_r))] + E_{\tilde{x}_g \sim P_g}[log(1 - D(\tilde{x}_g))]. \quad (1)$$

When the discriminator is trained to its optimal state before each generator parameter is updated. Minimization of the function is equivalent to minimization of the Jensen Shannon divergence between P_r and P_g. However, doing so often leads to vanishing gradients as the discriminator saturates. To address GAN's problems, Arjovsky[13)] illustrated the GAN problem mathematically caused by the JS divergence approximation and proposed to use the Earth-Mover Distance. The WGAN value function constructed using Kantorovich-Rubinstein duality[30)], can be simplified to calculate:

$$L_{WGAN} = \min_G \max_{D \in Z} E_{x_r \sim P_r}[D(x_r)] + E_{\tilde{x}_g \sim P_g}[1 - D(\tilde{x}_g)], \quad (2)$$

where Z is the set of 1-Lipschitz functions. To enforce constraints, WGAN applies a simple clipping to restrict the maximum value of the weight. The weights of the discriminator must be within a specific range controlled by the given hyperparameters. However, they found that weight clipping in WGAN leads to optimization difficulties due to the interactions between the weight constraint and the cost function, and even in the case where the optimization succeeds, the resulting critic may have a pathological value surface. Gulrajani[13)] proposed WGAN-GP uses a gradient penalty instead of weight clipping to enforce the Lipschitz constraint. The differentiable function is 1- Lipschitz if and only if it has gradients with the norm at most one everywhere, so they consider directly constraining the gradient norm of the critic's output concerning its input. To circumvent the tractability issues, they also enforced a version of the constraint with a penalty on the gradient norm for random samples $\tilde{x}_g \sim P_g$.

We express the objective as:

$$L_{W-GP} = L_{WGAN} + \lambda E_{\tilde{x}_g \sim P_g}(||\nabla_{\tilde{x}_g} D(\tilde{x}_g)||_2 - 1)^2. \quad (3)$$

In this paper, we set our parameter $\lambda = 10$ in the equation3.

3.2 Our loss functions

In comparison to the CycleGAN loss, we use the Wasserstein GAN (WGAN-GP) rather than the sigmoid cross-entropy loss that is used in the CycleGAN model. Results show the superior performance to the original generator architectures in terms of both the convergence of the generator and the sample quality.

The goal of our task is to learn a mapping function that the distribution of images from G_x is indistinguishable from the distribution of Y by using an adversarial loss. Two generators mapping G_x: $X \rightarrow Y$ and $G_y : Y \rightarrow X$. Two adversarial discriminators D_x and D_y proposed to distinguish whether images translated from another domain. So, we also use the cycle consistency loss defined by CycleGAN.

$$L_{CYC}(G_x, G_y) = E_{y \sim P_y}[||G_x(G_y(y)) - y||_1] + E_{x \sim P_x}[||G_y(G_x(x)) - x||_1] \quad (4)$$

For encouraging the transfer, we preserve the consistency of the facial feature information between the input and the output. We added style loss and content loss from Johnson et al.[17)] have shown impressive results for neural style transfer and super-resolution.

Applying the MSE loss function will cause the output image to be loose, such as details or high-frequency parts that are lost in the image. So appropriately selecting the features of a layer output as the input of the perceptual loss function can enhance the details. Our goal is to minimize the content loss between the content image and the generated one, while minimizing the style loss between the style image and the generated one. So in this paper, we use the ResNet50 network model[26)]. Let $\phi_j(x)$ be the activations at the $j - th$ layer of the ResNet50 for the input x , which is a feature map of shape (C_j, H_j, W_j).

The formula of the Gram matrix can be expressed as:

$$G_j^{\phi} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'}. \quad (5)$$

The correlations between the different filters applied at a given layer. These feature correlations are captured the texture patterns at a spatial scale, which correspond to the physical appearance of textures, colors, patterns at this scale. We pass the filter over the image to create the feature maps at our layer, and then we compute the inner product between feature maps of a given layer. This inner product is called the Gram matrix.

The Style loss is the squared norm of the difference between the Gram matrices of the output and target image. For optimization, style reconstruction is based on a set of layers of size J rather than a single layer. Incorporating feature correlations of multiple layers, we can obtain stationary, multi-scale representation of the input image. We define L_{style} to be the sum of losses for each layer $j \in J$. And it can be expressed as:

$$L_{style} = \sum_j (||G_j^{\phi}(x) - G_j^{\phi}(G_x(x))||^2 + ||G_j^{\phi}(y) - G_j^{\phi}(G_y(y))||^2). \quad (6)$$

However, even though minimizing the style reconstruction loss preserves stylistic features from the target image it does not preserve its spatial structure. The content loss is the Euclidean distance between the output image and the target image. Moreover, earlier layers in our network are associated with more local information, whereas, in the higher layers, activations will contain more global information. Since the content is defined by the macrostructure of the image, the topmost layers in our network will capture the content of the image[17]. It enforces the generated image retaining the prescribed structure, which is defined as：

$$L_{content} = \sum_j \frac{1}{C_j H_j W_j} (||\phi_j(x) - \phi_j(G_x(x))||^2 + ||\phi_j(y) - \phi_j(G_y(y))||^2). \quad (7)$$

So, our total objective is:

$$L_{total} = L_{W-GP} + \beta L_{CYC} + L_{style} + L_{content}. \quad (8)$$

For training, we minimize the above objective function, which consists of adversarial loss, cycle consistency loss, style loss, and content loss. β is non-zero weights for balancing loss functions.

3.3 Training details

We use Adam solver with a fixed learning rate of 0.0002 and beta1=0.5. We set $\beta = 10$ in Eq.(8). The

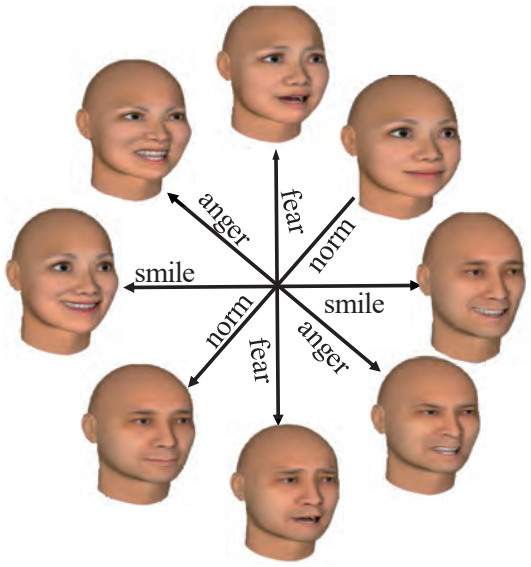


Fig. 3 FaceGen allows limited parametric control of facial expressions

Fig. 4 Samples of our training datasets

ResNet50 network is used to produce a feature descriptor for style transfer algorithms. The output of this network is the style and content features. We set content layer: 'res5b_branch2a', style layers: 'res3a_branch2a', 'res4a_branch2a', 'res5a_branch2a'.

4. Experiment Result

4.1 Datasets

Synthetic Images: Synthetic face images based on FaceGen[4] model reconstruction which is shown in **Fig.3**.

The FaceGen system uses a "parameterized" approach to define the attributes that compose a face, and using a set of fixed parameters, a face model can be made to morph and modify it independently. Once the texture of the face is attached, the geometric details will be more abundant in the face. You can see more information about synthetic face images from the website[31].

Real Images: We use the 300-W challenge[32] real images in training. It includes AFLW[33], AFW[34], Helen[35], IBUG[32], LFPW[36]. As shown in **Fig.4**, the upper part represents real face images from the 300-W challenge. The below part represents synthetic images from Face-

Fig. 5 Sample of images from FER2013 training database

Fig. 6 One mapping results of our translation model

Gen 3D model reconstruction (from left to right, representing the facial expression attributes of neutral, anger, happy, and fear respectively).

FER2013 Datasets: FER-2013[37] is from a Kaggle Facial Expression Recognition Challenge. It was collected by Pierre and Aaron through the Google Image Search Application Programming Interface (API). It is currently the largest publicly available facial expression database in the wild, making it possible for many researchers to train machine learning methods where a large amount of data[38] is needed.

But this dataset is challenging: (1) Facial expressions come from the natural environment, so low image resolution and different angles will increase the difficulty, as shown in the first row in **Fig.5**. (2) There are various watermark images(marked with a purple circle in Fig.5), cartoon images(marked with a green circle), non-face images(marked with a blue circle) in the training datasets. Especially, we need to focus on some facial expressions (marked with a red circle). The labeled category of this picture is fear, but even for the human eye, it is difficult to determine whether it is neutral or fear without considering the actual environment.

4.2 Qualitative evaluation

In our result, we can get two models: $S \rightarrow R$ and $R \rightarrow S$(S: synthetic images R: real images). We can convert the synthetic face images generated by the FaceGen 3D model into more realistic face images. Besides, this process is reversible. Depending on different tasks and goals, you can choose from synthetic images to real images or from real images to synthetic images. We give our experiment results in **Fig.6** and **Fig.7**.

Figure 6 shows female and male results of the model $S \rightarrow R$. We can convert the synthetic face image into a real face with multiple styles of expression. The advantage is that the generated face image has the high perceptual quality and can be used to improve the performance of the facial expression classifier.

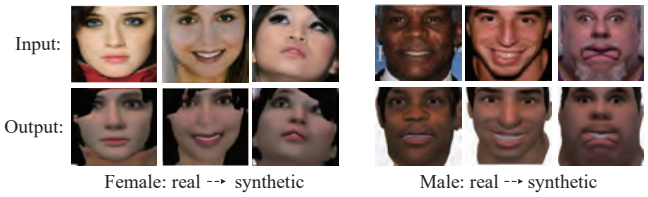


Fig. 7 Other mapping results of our translation model

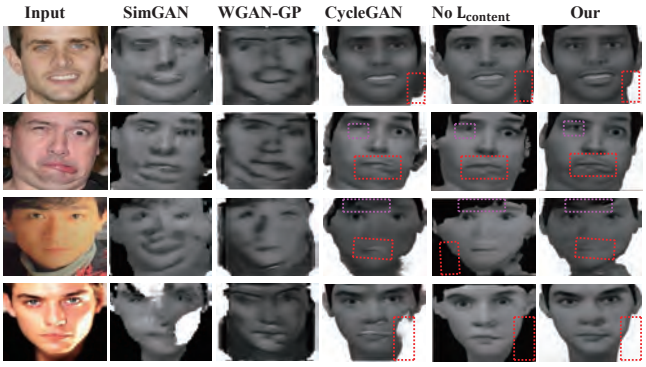


Fig. 8 Different models result for translation from real image to synthetic image

Figure 7 shows female and male results of the model $R \rightarrow S$. We can see that our experiments perform well with large-scale expression poses. And the generated faces maintain the characteristics of synthetic and real images at the same time.

We use the same evaluation datasets and metrics and compared our algorithm against several following latest works: SimGAN[15], WGAN-GP[13], CycleGAN[5]. Among the four input images in **Fig.8** is the normal image, an image with exaggerated expressions, a dark picture, and an image under intense light. From left to right: input, SimGAN[15], WGAN-GP[13], CycleGAN[5], Ours without $L_{content}$, Ours. Nearly all images work well by our method despite the diversity of their colors, expression gesture, and brightness. The image generated by SimGAN is affected by the background of the image, causing the face connected to the background. WGAN-GP lacks the details of the face. For the image under strong light and dark light, some positions that were affected by light or darkness are missing from CycleGAN results. And if we remove $L_{content}$ from our method, it cannot learn the background of synthetic data. The dashed box on the picture also shows the comparison between CycleGAN and our model.

We use an illustrative -Toy Experiment[39),40)] as our

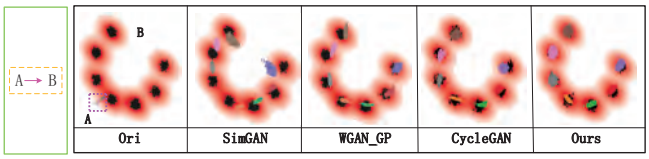


Fig. 9 Toy domain experiment result of different GAN models

evaluation. It based on synthetic data in domains A and B. Both samples are drawn from Gaussian mixture models. The task is to find relations between domain A and domain B and translate from domain A into domain B, which has seven modes spread around the arc of a circle. In this experiment, we provide 100,000 datasets and randomly divide the datasets into a training set and a validation set at a ratio of 0.33, and test on 1000 samples. The training performed for 50,000 iterations, the resulting translated samples in **Fig.9** show very different behavior depending on the model used.

In Fig.9, the target domain mark with black points. Colored points on the domain B planes represent samples from domain A that mapped to domain B, and each color denotes samples from each mode in domain A. The left-most figure demonstrates the initial state of the toy experiment, where all modes of domain A mapped to almost a single location because of the initialization of the generator. Results show many translated points of different colors localize around the same mode in domain B. However, both SimGAN and WGAN-GP models fail to cover all of the modes in domain B, because the mapping from domain A to domain B is injective. For example, some translated points of different colors (likes pink and gray) locate around the same B domain mode. CycleGAN and Our (with a reconstruction loss) results prevent mode-collapse by translating into distinct well-bounded regions that do not overlap. But some colored points from domain A in CycleGAN mapped to fewer areas of the domain B than our results. Our discriminator for domain B perfectly fools by translated samples from domain A around the modes of domain B.

4.3 Quantitative evaluation

Our method can alleviate the class imbalance in the training datasets by generating synthetic images. Thus, we tried to use the augmented synthetic datasets as augmented data for training a classifier to distinguish the facial expressions from each other. The overall procedure is described in **Algorithm1**.

(a) Metric

In **Table 1**, we summarize the calculation method of

Algorithm 1: Classifier training using GAN data augmentation

Data: Labelled FaceGen dataset $D_{facegen}$, real dataset D_{real}, Number of samples to append each iteration N_{iter};

Train our GAN using$\{D_{facegen}, D_{real}\}$;
Generate sample set D_{gen} from GAN transfer;
Form augumented dataset $D_{augmented}$ $\Leftarrow\{D_{gen}, D_{real}\}$;
Set $D_{training}\Leftarrow\{\}$;
Initialize CNN parameters;
for *Number of CNN training iterations* **do**
  for $x_n \in D_{augmented}$ **do**
    Append the samples x_n with the N_{iter} to $D_{training}$;
    train CNN using $D_{training}$;
    Evaluate CNN balanced accuracy using testsets;
  end
end

Table 1 Definitions of the evaluation metrics

Performance Measure	Mathematical Formula
Precision	TP/(FP+TP)
F1	2TP/(2TP+FP+FN)
Recall	TP/(FN+TP)
Accuracy	(TP+TN)/(TP+TN+FN+FP)
TP:true positive, FP:false positive, TN:true negtive, FN:false negative	

F1-score, Accuracy, Precision, and Recall rate as our base evaluation metric. We use this metric to compare the difference between the classifier's results and the ground truth. The precision rate(the number of correct targets divided by the number of all returned samples) shows the positive predictive value. The recall (the number of right targets divided by the number of targets that should be returned) can reflect the sensitivity and real positive rate of the model. The higher the F1-score(the harmonic mean of precision and recall), the better the performance of the model.

(b) FER2013

The datasets for training the model contains 35,887 samples of 48x48 pixel grayscale images of faces. It includes 28708 training images, 3589 public test images, and 3589 private test images. These images marked with one of the 7 emotion categories: 4953 images express anger, 547 images express disgust, 5121 images express

fear, 8989 images express happiness, 6077 images express sadness, 4002 images express surprise and 6198 images are emotionally neutral. As shown in **Fig.10**.

We build a neural network classifier that trained on: (I) only real samples taken from FER2013, (II) only original FaceGen synthetic samples. (III) the combination of FER2013 and original FaceGen synthetic samples. (IV) only our synthetic datasets(Using one of our mapping models $S \rightarrow R$ to generate the synthesized image with different facial expression styles). (V) a combination of FER2013 and our synthetic samples. (I),(III) and (V) consisting of the same FER2013 images but with different synthetic images. Please note that the test datasets and validation datasets of our model are still the original test datasets and validation datasets of FER2013.

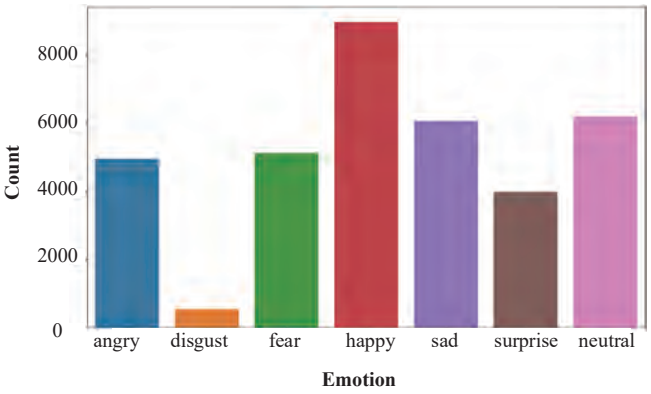


Fig. 10 An overview of FER2013

In **Table 2**, II, and III denote the model used different synthetic training datasets respectively. The results show that the model that used our synthetic datasets as training outperforms the model that used the original synthetic FaceGen dataset. Comparing to I, we can also see that there are improvements in terms of Precision, Recall and F1-score with V. In the experiments, the model that trained using our generated images and real images achieved the highest scores.

In **Table 3**, 'R' denotes 'Real images', 'S' denotes 'Synthetic images','SR' denotes 'Real images and Synthetic images'. Our method based on the backbone of the VGG network[41),44)]. As a result, we achieve an accuracy rate of 72.5 on the Private Test dataset. Observe that this performance is a consequence of our adaptive data generation algorithm and the use of bidirectional mapping.

We presented the results on the Public Test datasets in

Table 2 Comparison of the evaluation metrics of the classification model under different training datasets

Evaluate / Methods	Top1-acc	Top5-acc	Pre	Rec	F1
I	64.70	95.59	60.31	64.27	63.61
II	27.92	77.91	26.84	27.95	27.24
III	65.16	96.13	61.56	64.65	64.58
IV	38.28	81.76	32.37	37.91	32.52
V	**66.54**	**96.29**	**63.02**	**66.14**	**66.07**

Table 3 Comparison of the different classification models on FER2013 datasets

Method	Trained On	Tested with	Accuracy(%)
Human[37)]	-	R	65.0± 5
GoogleNet[41)]	R	R	65.2
Xception[42)]	R	R	66.0
VGG[43)]	R	R	70.8
Ours	SR	R	**72.5**

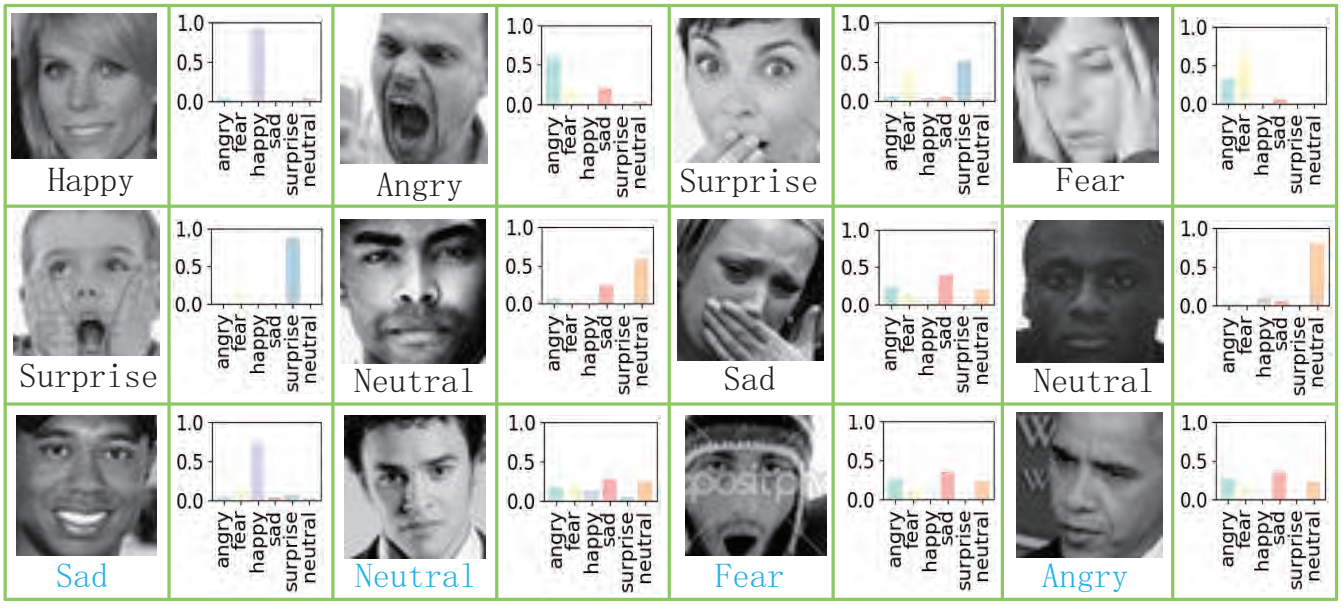


Fig. 11 The predicted labels and corresponding probabilities of the sample face randomly selected in the test datasets

Fig.11. In the first two rows of Fig.11, we have shown the ground-truth labels and the corresponding predicted probabilities of our model. In the last rows of Fig.11, we show a few examples with light blue labels to indicate false predictions. These examples need to be considered in a real-world environment and typically have a second possible emotion in them. For example, for the first sample of the last rows, the correct label is blue, the maximum probability predicted by our model is happy. Under the natural condition (neutral expressions), the corners of eyes and mouth corners will drop slightly, which is also the main feature of sad, so these two expressions will also be confused during classification. Fear is most easily confused with sadness or anger. Because both facial expressions have the same characteristics of tightening eyebrows, straining the forehead, and opening lips.

Figure 12 and **Fig.13** are the confusion matrix for the model predictions on the test dataset. The matrix gives the counts of emotion predictions and some insights into the performance of the multi-class classification model. From the confusion matrix, we can see that the fear recognition is the most difficult one, and happy and surprise

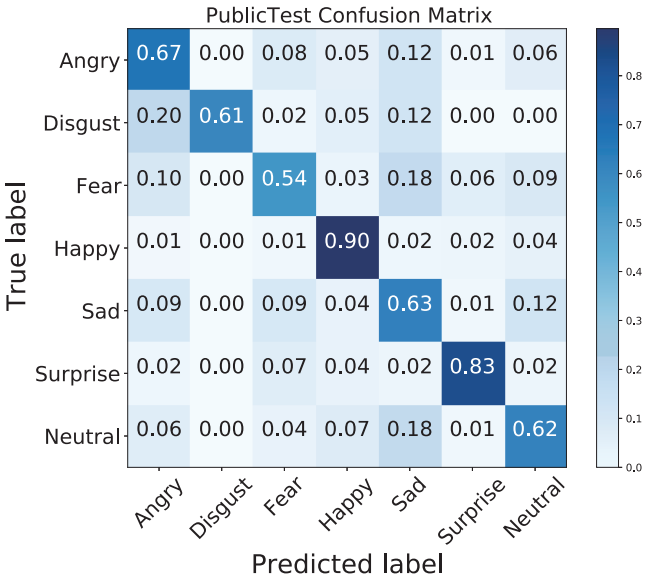


Fig. 12 Normalized confusion matrix of our method on FER2013 PublicTest datasets

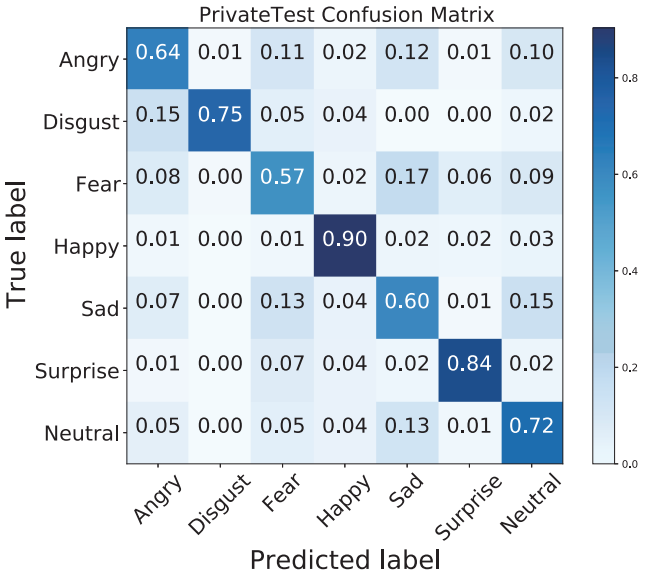


Fig. 13 Normalized confusion matrix of our method on FER2013 PrivateTest datasets

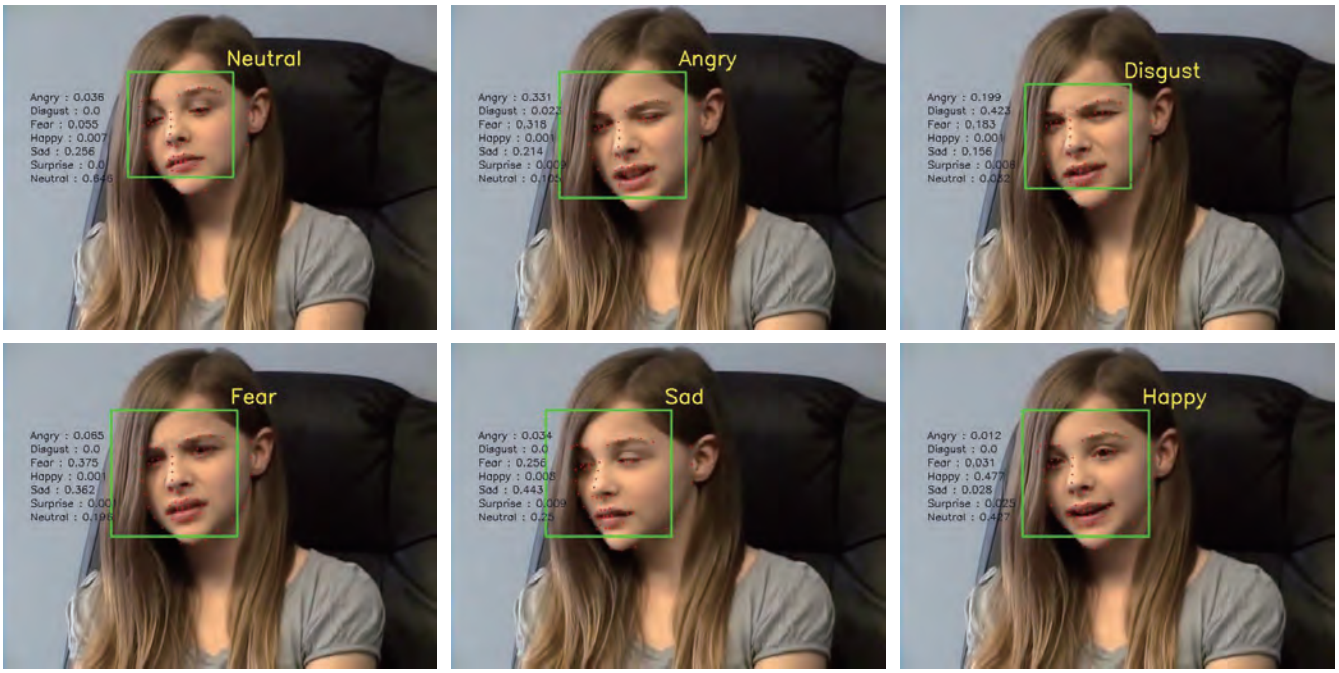


Fig. 14 The results of real-time facial application on 300-VW datasets[45)]

are easier to recognize than the other facial expressions. The majority of the happy expressions have distinctive features, such as the slightly open mouth, the lifted corners of the mouth. In contrast, disgust is easily confused with anger. Because the two expressions have the same properties of the drooping eyebrow and the squinting.

An example of our complete application show in **Fig.14**. Once the model train to its optimal state, we can apply it to practical applications. Our Real-time facial applications include face detection, facial expression classification, and facial marker detection.

5. Conclusion

We propose an Improved CycleGAN for bidirectional mappings translation. Our network not only generates a synthetic image that incorporates real facial characteristics information but also real-world images with synthetic style. The generated images demonstrate the superiority of our proposed method. Next, we make use of the facial expression classification model to assess the validity of the generated face images. On our datasets, the recognition rate of classification is improved. We achieve an accuracy rate of 72.5% on FER2013 PrivateTest and an accuracy rate of 71.2% on FER2013 PublicTest on the task of facial expression attribute classification. This performance improvement attribute to our adaptive data generation algorithm. In this experiment, we didn't consider occlusion when creating synthetic datasets, such as sunglasses, masks, and so on. But we know that real datasets may have various possibilities. Therefore, further improvement to cope with such cases will be needed.

References

1) A. Todorov, N. N. Oosterhof: "Modeling Social Perception of Faces [Social Sciences]", IEEE Signal Processing Magazine, Vol. 28, No. 2, pp.117–122 (2011).
2) X. Zhang, Y. Fu, S. Jiang, X. Xue, Y. G. Jiang, G. Agam: "Stacked Multichannel Autoencoder–an Efficient Way of Learning from Synthetic Data", Multimedia Tools and Applications, Vol. 77, No. 20, pp. 26563–26580 (2018).
3) J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, D. Ramanan: "Depth-Based Hand Pose Estimation: Data, Methods, and Challenges", Proc. of the IEEE International Conference on Computer Vision (ICCV), pp.1868–1876 (2015).
4) FaceGen, http://www.facegen.com (2020).)
5) J. Y. Zhu, T. Park, P. Isola, A. A. Efros: "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks", Proc. of the IEEE International Conference on Computer Vision (ICCV). pp.2223–2232 (2017).
6) L. Perez, J. Wang: "The Effectiveness of Data Augmentation in Image Classification Using Deep Learning", Convolutional Neural Networks Vis. Recognit, Vol.11 (2017).
7) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio: "Generative Adversarial Nets", Advances in Neural Information Processing Systems, pp.2672–2680 (2014).
8) L. A. Gatys, A. S. Ecker, M. Bethge: "Image Style Transfer Using Convolutional Neural Networks", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423 (2016).
9) E. L. Denton, S. Chintala, R. Fergus: "Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks", Advances in Neural Information Processing Systems, pp.1486–1494 (2015).
10) D. P. Kingma, M. Welling: "Auto-Encoding Variational Bayes", Proc. of Second International Conference on Learning Representations (ICLR) (2014).
11) T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen: "Improved Techniques for Training GANs", Advances in Neural Information Processing Systems, pp.2234–2242 (2016).
12) M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, Y. LeCun: "Disentangling Factors of Variation in Deep Representation Using Adversarial Training", Advances in Neural Information Processing Systems, pp.5040–5048 (2016).
13) I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville: "Improved Training of Wasserstein GANs", Advances in Neural Information Processing Systems, Vol.30, pp.5767–5777 (2017).
14) J. Gauthier: "Conditional Generative Adversarial Nets for Convolutional Face Generation", Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter Semester, Vol.2014, No.5, pp. 2 (2014).
15) A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb: "Learning from Simulated and Unsupervised Images through Adversarial Training", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (ICPR), pp.2242–2251 (2017).
16) P. Isola, J. Y. Zhu, T. Zhou, A. A. Efros: "Image-to-Image Translation with Conditional Adversarial Networks", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (ICPR), pp.5967–5976 (2017).
17) J. Johnson, A. Alahi, L. Fei-Fei: "Perceptual Losses for Real-Time Style Transfer and Super-Resolution", Proc. of European Conference on Computer Vision (ECCV), pp.694–711 (2016).
18) H. Q. Gao, K. Ogawara: "Generative Adversarial Network for Bidirectional Mappings between Synthetic and Real Facial Image", Proc. of Twelfth International Conference on Digital Image Processing, Vol. 11519, p. 115190J (2020).
19) Y. Wu, F. Yang, Y. Xu, Y. Ling: "Privacy-Protective-GAN for Privacy Preserving Face De-Identification", Journal of Computer Science and Technology, Vol.34, No.1, pp.47–60 (2019).
20) C. Tian, Y. Xu, L. Fei, J. Wang, J. Wen, N. Luo: "Enhanced CNN for Image Denoising", CAAI Trans. on Intelligence Technology, Vol.4, No.1, pp.17–23 (2019).
21) L. Wan, J. Wan, Y. Jin, Z. Tan, S. Z. Li: "Fine-Grained Multi-Attribute Adversarial Learning for Face Generation of Age, Gender and Ethnicity", Proc. of International Conference on Biometrics, pp.98–103 (2018).
22) T. Karras, S. Laine, T. Aila: "A Style-Based Generator Architecture for Generative Adversarial Networks", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4401–4410 (2019).
23) T. Karras, T. Aila, S. Laine, J. Lehtinen: "Progressive Growing of GANs for Improved Quality, Stability, and Variation", arXiv:1710.10196 (2017).

24) H. Chang, J. Lu, F. Yu, A. Finkelstein: "Pairedcyclegan: Asymmetric Style Transfer for Applying and Removing Makeup", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.40–48 (2018).
25) Y. Shen, P. Luo, J. Yan, X. Wang, X. Tang: "FaceidGAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.821–830 (2018).
26) K. He, X. Zhang, S. Ren, J. Sun: "Deep Residual Learning for Image Recognition", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778 (2016).
27) C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi: "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4681–4690 (2017).
28) C. Li, M. Wand: "Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks", Proc. of European Conference on Computer Vision (ECCV), pp.702–716 (2016).
29) D. Ulyanov, A. Vedaldi, V. Lempitsky: "Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6924–6932 (2017).
30) L. Hörmander, The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis, Springer (2015).
31) qiqi7788, https://github.com/qiqi7788/dataset-and-result (2020).
32) C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic: "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge", Proc. of the IEEE International Conference on Computer Vision Workshops (ICCV), pp.3–18 (2016).
33) M. Koestinger, P. Wohlhart, P. M. Roth, H. Bischof: "Annotated Facial Landmarks in the Wild: A Large-Scale, Real-World Database for Facial Landmark Localization", Proc. of 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp.2144–2151 (2011).
34) X. Zhu, D. Ramanan: "Face Detection, Pose Estimation, and Landmark Localization in the Wild", Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2879–2886 (2012).
35) V. Le, J. Brandt, Z. Lin, L. Bourdev, T. S. Huang: "Interactive Facial Feature Localization", Proc. of European Conference on Computer Vision (ECCV), pp.679–692 (2012).
36) P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman: "Localizing Parts of Faces Using a Consensus of Exemplars", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.35, No.12, pp.2930–2940 (2013).
37) I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. Lee, Y. Zhou, C. Ramaiah, F. X. Feng, R. F. Li, X. J. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. J. Xie, L. Romaszko, B. Xu, Z. Chuang, Y. Bengio: "Challenges in Representation Learning: A Report on Three Machine Learning Contests", Proc. of International Conference on Neural Information Processing, pp.117–124 (2013).
38) A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, M. H. Mahoor: "Facial Expression Recognition from World Wild Web", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), pp.58–65 (2016).
39) T. Kim, M. Cha, H. Kim, J. K. Lee, J. Kim: "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks", Proc. of the 34th International Conference on Machine Learning, Vol.70, No.18, pp.57–1865 (2017).
40) Z. Yi, H. Zhang, P. Tan, M. Gong: "Dualgan: Unsupervised Dual Learning for Image-to-Image Translation", Proc. of the IEEE International Conference on Computer Vision (ICCV), pp.2849–2857 (2017).
41) P. Giannopoulos, I. Perikos, I. Hatzilygeroudis: "Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013", Advances in Hybridization of Intelligent Methods, pp. 1–16 (2017).
42) O. Arriaga, M. Valdenegro-Toro, P. Plöger: "Real-Time Convolutional Neural Networks for Emotion and Gender Classification", Proc. of 27th European Symposium on Artificial Neural Networks, pp. 24–26 (2019).
43) K. Simonyan, A. Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition", Proc. of 3rd International Conference on Learning Representations (ICLR) (2015).
44) S. Minaee, A. Abdolrashidi: "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network", Computing Research Repository, Vol.abs/1902.01019 (2019).
45) J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, M. Pantic: "The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results", Proc. of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 50–58 (2015).

(Received May 29, 2020)
(Revised October 7, 2020)

Haoqi GAO (*Student Member*)
She received the M.E. degree from Beihang University of Software in 2016. From 2019, she has been staying at Ph.D. course in System Engineering at Wakayama University. Her research interests include face detection, 3D face reconstruction, medical image processing, and deep learning.

Koichi OGAWARA
He received his Ph.D. degree in Information and Communication Engineering from the University of Tokyo, Japan, in 2002. From 2006 to 2011, he was a guest Associate Professor at Kyusyu University. Currently, he is an Associate Professor at Wakayama University. His research interests include robotics, computer vision, and machine learning. He won the Best Vision Paper Award at the IEEE International Conference on Robotics and Automation 2007.

IIEEJ **Paper**

Robust Sphere Detection in Unorganized 3D Point Clouds Using an Efficient Hough Voting Scheme Based on Sliding Voxels

Jaime SANDOVAL†(*Student Member*), Kazuma UENISHI†(*Member*),
Munetoshi IWAKIRI††(*Member*), Kiyoshi TANAKA†(*Fellow*)

†Shinshu University, ††National Defense Academy of Japan

<**Summary**> Sphere detection in point clouds is an important task in 3D computer vision with various applications such as reverse engineering, medical imaging, Terrestrial Laser Scans (TLS) alignment, and so on. So far, several approaches have been proposed to detect spheres in point clouds. However, conventional methods are inefficient and inaccurate because they depend on random sampling, point-wise voting or normal vectors estimation to generate hypothetical spheres. To overcome these drawbacks, we propose a novel algorithm that employs sliding voxels and Hough voting to robustly and efficiently detect spheres in unorganized point clouds. The proposed method can analyze all the points contained in point clouds without deteriorating its efficiency and accuracy in contrast to conventional methods. Through experiments, we found that the proposed method can drastically reduce the processing time and achieve more accurate and robust performance in severer conditions than conventional methods.

Keywords: sphere detection, unorganized point clouds, sliding voxel, Hough voting, efficient RANSAC

1. Introduction

Point clouds are sets of points in an $\mathbb{R}^3$ space that resemble the surface of objects. They can be obtained by a wide variety of laser-based sensing techniques and photogrammetry. However, because of these diverse ways of obtaining them, we face several problems such as huge variations in points density, sensing patterns, sensor artifacts, and noise.

Point clouds can be categorized as organized and unorganized. Organized point clouds can be arranged into an image-like 2D matrix, in which each pixel is associated with a 3D point. These point clouds come from range/depth sensors or stereovision. On the other hand, unorganized point clouds have no specified order and are just a list of 3D points, which are a more general form than the organized counterparts. These come from sensors that change their coordinate system while scanning, like rotating-head 3D LIDAR sensors or dense point clouds from photogrammetry. However, because unorganized point clouds have no order, basic techniques such as nearest neighbor searching become increasingly difficult and time-consuming in unorganized point clouds. Therefore, in this paper, we focus on developing algorithms for unorganized point clouds, which are in high demand because of their numerous applications.

Man-made objects can be approximated by geometric primitives such as spheres, planes, and cylinders[1),2)]. These geometric primitives can act as proxy entities for other real-world objects, for instance, the trunk of a tree and human arms and legs can be approximated by cylinders, also, the head of animals and humans can be approximated as spheres.

Sphere detection is an important technique in 3D computer vision with applications in broad areas such as materials engineering[3)], measuring[4)], medicine[5)] among others. As opposed to other parametric shapes such as planes or other quadric surfaces, spheres have a viewpoint-independent geometry. This property allows us to calculate its coefficients with partial views such as those obtained from 3D sensors. Therefore, they are preferred as targets in point clouds registration[6)–8)] where it is crucial to detect their coefficients with the highest accuracy.

However, the conventional sphere detection methods[1),9)–12)] fail to detect spheres when the inliers ratio becomes too small due to noise, and the range of the point cloud.

Moreover, the processing time can increase exponentially as some algorithms depend on point-wise normal vectors, and their accuracy is highly dependent on their

estimation. Also, normal vectors are calculated via Principal Component Analysis (PCA), a method known to be very susceptible to noise and outliers.

Therefore, in this paper, we propose a novel, highly accurate, robust, and drastically more efficient (high speed) method for sphere detection based on sliding voxels. Instead of random sampling, it uses an efficient octree subdivision to detect robustly hypothetical spheres deterministically. Then, the most prominent spheres are detected via Hough voting[13)–15)]. Lastly, its coefficients are refined and pruned by their completeness. Experiments with synthetic and real point cloud data from Terrestrial Laser Scanning (TLS) confirm the superior performance of the proposed method.

This paper is organized as follows. In section 2, we briefly give a literature overview and explain the conventional method. Section 3 describes in detail the proposed method. Section 4 is focused on describing the datasets, the evaluation method, used metrics, and showing experiments results. In section 5, we summarized the results and future works.

2. Previous Work

2.1 Literature overview

Several approaches to detect spheres using spatial information have been developed in recent years. A survey work[16)] summarizes the conventional methods for geometric primitive detection from 3D data. Region growing clusters similar regions of the point cloud at random locations iteratively. Although computationally expensive, it can be used to later fit geometric primitives in segmented regions.

Hough Transform (HT)[13),14)] based algorithms for parametric shape detection have been proposed in the past. As spheres become perfect circles when projected into a plane, the circular HT was applied to detect spheres as circles in images[17)]. Unfortunately, unorganized point clouds do not have a defined projection to an image array. Moreover, when using images to detect spheres, the radius of circles vary depending on the distance from the sensor.

The parameter space for the sphere has four dimensions. Therefore, applying the Standard Hough Transform (SHT)[14)] is unfeasible both in computational complexity and in memory usage of the accumulator array used to store votes and detect shapes. Ogundana et al.[9)] proposed a fast HT detector by fixing the radius, casting a single vote for each point, and detecting spheres using a sparse accumulator. Nonetheless, they require computationally expensive point-wise normals estimation. Also, fixing the radius limits its applications in real-world scenarios where it is usually unknown.

Abuzaina et al.[10)] also proposed a HT for sphere detection using a sparse accumulator. Their approach uses a polar representation of the parameter space and a more exhaustive per-point voting. To overcome this, they fixed the radius and reduced the number of input points based on the points' density of a Kinect sensor limited at a certain range.

In the past, strategies based on random sampling were proposed to reduce the computational complexity of the voting phase in Hough transform methods. The Probabilistic Hough Transform (PHT)[18)] reduces the number of evaluated points by selecting a random subset. The Combined Multi-Point Hough Transform (CMPHT)[11)] follows the PHT approach to reduce computational complexity, and evaluates several Hough transforms for sphere detection. A single-point HT with coarse accumulator quantization serves as an efficient coarse approximation that identifies regions-of-interest (ROI) where spheres are more likely to be found. Then, a 4-point Hough transform is chosen as a detection refinement over the ROI. Experiments with real point clouds with a Kinect sensor showed that the computational complexity and the success rate of CMPHT are highly affected by the inliers ratio

$$\phi_r = \frac{N_{\text{in}}}{N}, \tag{1}$$

where N_{in} is the number of inliers of the point cloud and N is the number of total points, which means that lower ϕ_r contains more outliers induced by noise or other non-spherical surfaces.

Reducing the dimensions of the parameter space was an unexplored approach for primitive detection. A recent work[12)] shows a study of a multi-shape and multi-model detector based on Point-Pair Features (PPF). However, it expects an input cloud with normal vectors, and the PPF is highly dependent on the correct estimation of its normal orientation. Although it outperforms state-of-the-art primitive detection in their experiments, the dataset they used were CAD models and point clouds with removed background and geometric primitives present in the foreground where points are denser and less noisy. Theoretically, if we remove the computational complexity of normal vectors estimation, its processing time is still heavily dependent on the total number of points since it needs

to compute PPFs for every unique combination of point-pairs with normals. Therefore, the authors restricted the number of points to compute to 2048 random points, which is unacceptable for long-range (long distance between the sensor and its farthest point) point clouds such as those used in TLS or large-scale photogrammetry.

2.2 RANSAC-based methods

RANSAC tries to fit a model into a point cloud by random sampling iteratively. The number of trials k is defined by

$$k = \frac{\log(1-z)}{\log(1-b)}, \tag{2}$$

where z is the probability that at least one of the data points is error-free, and b is the probability that any set of selected data points is within the error tolerance to the model. Finally, it selects the model that has above a defined number of inliers within a threshold distance. As b can be calculated from the sampling, z is a parameter defined by the user.

RANSAC and its variations, such as MSAC[26] are implemented alongside geometric models of planes, spheres, and cylinders. However, the PCL implementation follows the original RANSAC[20], which is a single-instance and single-model fitting algorithm. When detecting more than one sphere using this approach, we would need to iteratively detect and remove inliers each time we find a good model. Hence, decreasing ϕ_r in each successful detection making it increasingly difficult to detect more shapes, and to decide finishing conditions.

Wang et al. combine a RANSAC-like sampling strategy with energy minimization to detect spheres in Kinect point clouds[21]. It starts by drawing a small set of hypothetical sphere models from random samples and then use energy minimization to label spherical points. Although it is not dependent on distance thresholds, its accuracy depends on the initial sphere models and weighing terms of the energy minimization function, which have to be guessed by the user depending on the outliers rate.

Spheres are a subtype of quadric surfaces (quadrics); therefore, they have nine Degrees-of-Freedom (DoF) parameters. A study[22] provides a quadric detector algorithm that overcomes current limitations by using RANSAC to search for sets of points (basis) that are likely to be on a quadric surface. These bases provide a coarse identification of a quadric surface to be further refined. Nonetheless, they proposed to fit linearly, a non-linear (quadric) surface. Hence, their fitting results are biased, and the effect of this bias in the final model coefficients of their method is unknown since it was out of the scope of their study[22]. Therefore, this method is not suitable for applications where the accuracy of the resulting sphere coefficients is crucial[6)–8)]. Furthermore, selecting an appropriate basis out of a long-range point cloud becomes complex task and highly depends on the inliers ratio of the point cloud (ϕ_r) since it inherits RANSAC disadvantages.

2.3 Efficient RANSAC

Efficient RANSAC (EFRANSAC)[1] uses octrees for a more efficient localized sampling. For spheres, it uses two random sampled points with their normal vectors to generate hypothetical spheres. Then, iteratively executes RANSAC over disjoint random subsets of the point cloud to validate the generated model.

Each sphere is refined by thresholding the expected curvature at each point N_{th}. Also, they map sphere inliers within a threshold ϵ to a low-distortion bitmap that resembles the surface of the sphere. This bitmap of bin size C_ϵ allows EFRANSAC to select the biggest connected component as the final inliers of each hypothetical sphere. The final coefficients are refined using non-linear least squares[27].

EFRANSAC finishing condition takes into account the octree level of the samples and is parameterized similarly to RANSAC. At a lower z, these algorithms will increase their determinism at the cost of increasing their iterations. According to the EFRANSAC implementation in the CGAL library[25], z is thresholded against

$$\mathrm{stop}_p = \left(1 - \frac{|L_c|}{4|P|O_{\mathrm{depth}}}\right)^{|C|}, \tag{3}$$

where L_c is the largest candidate size (in number of points), and $|P|$ is the number of available points that are not part of the selected shape candidates. O_{depth} is the depth of the octree, and $|C|$ is the number of candidates drawn so far.

2.4 Drawbacks of conventional methods

Table 1 lists the conventional methods, their approach for sphere detection, efficiency strategies, and their drawbacks. Conventional methods are inefficient because they depend on point-wise voting or normal vectors estimation to generate hypothetical spheres.

Methods based on the Hough transform (and Hough voting), due to point-wise exhaustive search, they resort to fix the radius of the spheres they can detect[9],

Table 1 Summary of conventional methods

Method	Detection approach	Efficiency strategy	Drawbacks
Fast HT[9]	Single-vote HT	Fixing radius, sparse accumulator	Fixed radius
Kinect HT[10]	SHT[14]	Background points removal	Point-wise search
CMPHT[11]	PHT[18]	Coarse-to-fine voting, random subsampling	Weak to low ϕ_r
PPF Hough voting[12]	PPF[19]	Random subsampling (up to 2048 points)	High combinatorial complexity, weak to low ϕ_r
Wang et al.[21]	RANSAC[20], energy minimization	Random sampling	Weak to low ϕ_r
Birdal et al.[22]	RANSAC[20], quadric voting	Random sampling	Weak to low ϕ_r, biased
EFRANSAC[1]	RANSAC[20], energy minimization	Random sampling, octree	Weak to low ϕ_r, biased

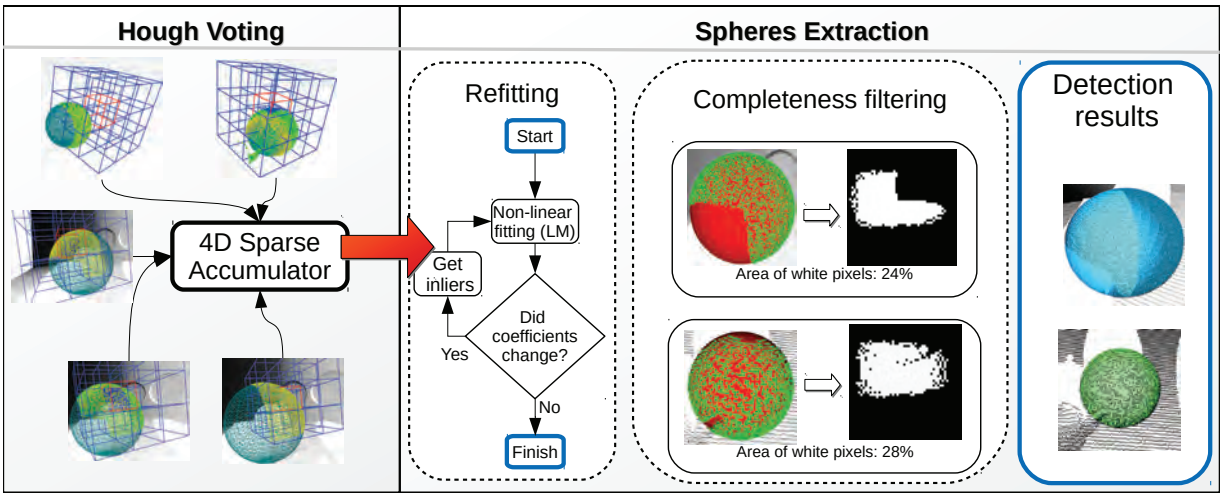


Fig. 1 Overview of the proposed method

or to limit the points they process[10)–12)]; thus, diminishing their accuracy and narrowing their applicability. To avoid doing exhaustive search, RANSAC-based methods[1),21),22)] resort to random sampling; making them weak to the inliers ratio ϕ_r of point clouds. As noise or outliers in point clouds increase, these methods fail to obtain valid hypothetical spheres; leading to misdetections. Moreover, their accuracy highly depends on the correct estimation of point-wise normal vectors.

Furthermore, 3D LIDAR point clouds possess a long-range, and the detection of relatively small and non-invasive sphere targets[6)–8)] is an extremely difficult task for all the conventional methods proposed so far.

In addition, most of the conventional methods lack of public implementations, only EFRANSAC[1)] is implemented in the CGAL Library[25)]. Therefore, in this work, we provide comparative assessments with EFRANSAC and the proposed method.

3. Proposed Method

3.1 Main features and superiority

In order to solve the aforementioned drawbacks of the conventional methods, in this paper, we propose an efficient and deterministic method to detect spheres in unorganized point clouds. The proposed method has two main features; (i) it employs a 3D space subdivision called sliding voxels that generates hypothetical spheres for Hough voting without discarding any point. In other words, the proposed method is capable of analyzing the whole point cloud without resorting to naive random sampling for hypothesis generation. Therefore, the sliding voxel technique contributes to achieving superior accuracy and robustness in sphere detection even in point clouds with low ϕ_r. (ii) Also, the proposed method transforms voxelized regions of the point cloud into local planes, which efficiently reduces the number of computations for Hough voting. That is, as opposed to conventional subsampling strategies prone to noise and outliers, the proposed method can achieve highly efficient Hough voting by em-

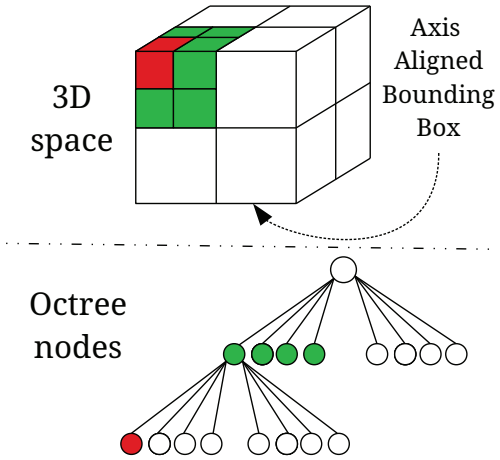


Fig. 2 3D space subdivision using an octree

ploying sliding voxels, which contributes to reducing the entire processing time drastically without deteriorating its accuracy. Moreover, these superiorities allow us to extend the applicability of the proposed method to the case of processing a huge amount of point clouds captured by TLS in real-world situations.

3.2 Process flow

Figure 1 shows a graphical overview of the proposed method. We proposed an efficient octree-based point cloud subdivision to robustly estimate hypothetical spheres with our novel sphere fitting algorithm. To globally detect those spheres, we performed Hough voting with a memory-efficient accumulator based on nested tree structures. Finally, the spheres are pruned by a completeness score and refitted using the connected components of the projecting bitmap.

The proposed method starts by dividing the point cloud space in a 3D grid using an octree. **Figure 2** shows how the 3D space of the point cloud is enclosed by an Axis Aligned Bounding Box (AABB) and subdivision occurs recursively until it complies with a voxel size V_s. The bounding box is expanded accordingly such that the leaves of the octree match the desired voxel size.

In each node of the octree, we save information about the local geometry of its points: centroid and normal vector. The normal vector was computed using PCA and corresponds to the eigenvector associated with the smallest eigenvalue. Therefore, the number of points to process becomes the number of occupied voxels with three or more points.

For each leaf, we are going to select their 26-neighbors leaves, such that contiguous cells overlap and share geometrical properties, i.e., a 3D $(3 \times 3 \times 3)$ sliding window with a stride of 1. We call this structure a sliding voxel, and it helps us to robustly and efficiently identify spherical-like regions in the point cloud. Then we take advantage of this structure to generate a hypothetical sphere for every sliding voxel in a point cloud.

3.3 Hypothetical spheres generation

Several local sphere fitting methods exists, among them, algebraic fitting uses linear least squares to fit a sphere in a point set. It works by rearranging the sphere equation

$$(x - C_x)^2 + (y - C_y)^2 + (z - C_z)^2 = r^2 \tag{4}$$

to the following

$$2xC_x + 2yC_y + 2zC_z + \alpha = x^2 + y^2 + z^2, \tag{5}$$

where $\{C_x, C_y, C_z, r\}$ are the parameters of the sphere and $\alpha = r^2 - C_x^2 - C_y^2 - C_z^2$.

Then, we can obtain the sphere parameters using the least squares normal equation of its matrix representation

$$A^T Ax = A^T b. \tag{6}$$

As $A^T A$ is symmetric and positive-definite, it can be solved with Cholesky factorization to get the sphere parameters from x.

A second alternative is to use a non-linear least-squares approach. The Levenberg-Marquardt(LM)[28] method uses gradient-based optimization to find the sphere coefficients that minimize the error between a sphere model and a point set. In the proposed method we used the error function

$$\arg\min_{C,r} \sum_{i=1}^{n} (P_i - C)^T (P_i - C) - r^2 \tag{7}$$

where P_i is a point of a set of n number of points, C is the center of the sphere and r is the radius. The Jacobian J of Eq. (8) given a point P_i is

$$J = [-2(P_{ix} - C_x), -2(P_{iy} - C_y), -2(P_{iz} - C_z), -2r]^T \tag{8}$$

where $\{P_{ix}, P_{iy}, P_{iz}\}$ are the point coordinates.

Since its correct convergence depends on an initial hypothesis, it is not clear which value is best for each case. When a sliding voxel falls in a planar surface we would expect to get a sphere with an arbitrarily far center and a big radius. However, both linear and non-linear least-squares approaches produce unpredictable results on planar point sets and are greatly affected by their number of outliers.

Therefore, we introduce a novel local fitting algorithm that provides a robust estimate of the best fit sphere of

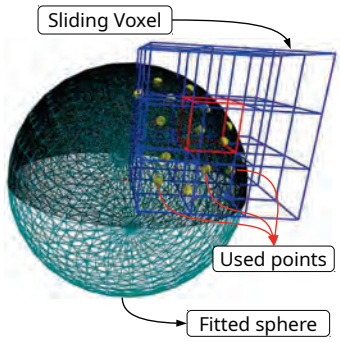


Fig. 3 Sphere fitting with sliding voxels

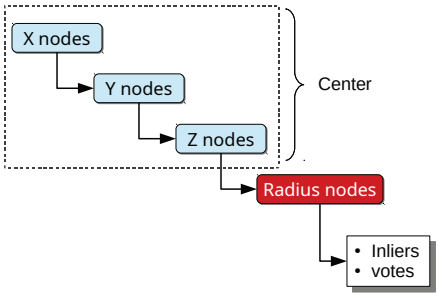


Fig. 4 Spheres accumulator structure

a point set. Given a set of N points with normal vectors Ψ of every sliding voxel, for all combinations of its items $\binom{N}{2}$, we estimate a sphere using the model generation method described by Schnabel, et al.[1] which uses two points and their normal vectors to estimate the parameters of one hypothetical sphere at a time. However, they do not indicate the specific method they used to estimate the sphere center. Therefore, we refer to the appendix section of this work for details about the estimation of a sphere using two oriented points.

By doing this for every pair of points and normals inside each sliding voxel, we are obtaining a set of all possible spheres $S \in \mathbb{R}^4$. As the sliding voxel has a maximum of 27 occupied voxels, the maximum cardinality of S is 351. Ideally, if the surface is perfectly spherical, all the spheres estimated will be the same, and S would have 0-variance in all its four dimensions. If we face with noise or outliers, its variance will fluctuate and so its mean and median.

To robustly get the most probable sphere, we select the median sphere $S_m \in S$ such that

$$\arg\min_i \|S_i - S_\mu\|_2, \ i \in [1, |S|] \tag{9}$$

where $S_i \in S$ and $|S|$ is the cardinality of the set S, $\|\cdot\|_2$ denotes the Euclidean norm, and S_μ is the mean sphere in S.

Given the mean sphere S_μ and S_m, we also calculate the spherical likelihood

$$\delta\mu = \|S_m - S_\mu\|_2 \tag{10}$$

If the sliding voxel falls into a perfect sphere, $\delta\mu$ will tend to 0 and will represent the deviation of the median with respect to the mean of the set of all possible spheres. Therefore, $\delta\mu$ is thresholded with the parameter $Th_{\delta\mu}$ to avoid regions that do not possess a spherical geometry.

Figure 3 shows how a portion of the point cloud in the sliding voxel can define the underlying spherical geometry by using our fitting method.

3.4 Hypothesis verification

Once all our hypothetical spheres are estimated, we make them converge into a Hough accumulator, since the spheres accumulator has four dimensions, we chose a memory-efficient nested tree structure shown in **Fig. 4**. It is common in the literature to find the accumulator discretization defined by the number of bins of each dimension. However, since all sphere dimensions are expressed using the same metric, we can parameterize it using the accumulator bin size Acc_{res} instead of the bins number.

After the voting finished, the most prominent spheres are extracted by accumulating the votes and performing peak detection in a $3 \times 3 \times 3 \times 3$ sliding window over the accumulator. A sphere will only be extracted if it has a minimum number of votes V_{min}.

As a result of the discretization, noise and outliers, the extracted spheres are not the final spheres but an approximation of what they should be. Therefore, we have to refit and prune them, as shown in Fig. 1. Voting took place with sliding voxel centroids, but refitting will take place with the actual points. Therefore, we start to search for inliers and refit iteratively until the sphere coefficients changes are negligible.

To avoid the negative influence of outliers, we map the inlier points of the refitting into a plane

$$\begin{aligned} x &= \frac{1}{\pi}\arccos\left(\frac{P_z}{\|P - C\|_2}\right) \\ y &= \frac{1}{2\pi}\text{atan2}(P_y, P_x) + \frac{1}{2} \end{aligned} \tag{11}$$

such that it results in a squared points distribution. Then a bitmap is generated by deciding the number of the discretization bins B_{bins}. From this bitmap, we perform two things:

- To select the biggest cluster using 8-neighbors clustering.
- To estimate a completeness measure.

After getting the biggest cluster, completeness K is

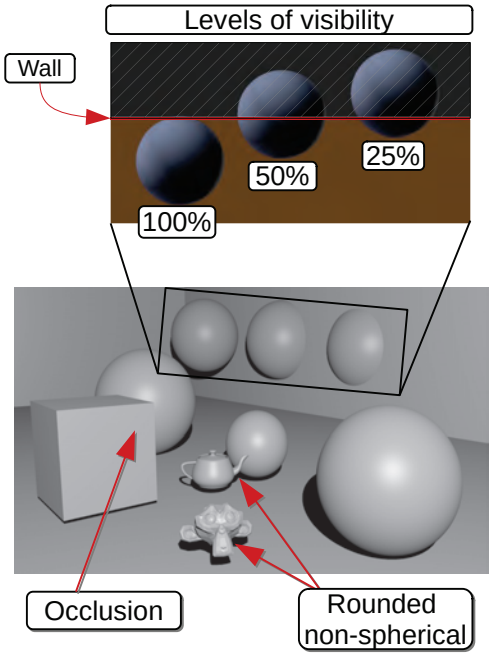


Fig. 5 Synthetic dataset $\mathcal{M}$

measured by the ratio

$$K = \frac{S_O}{S_T}, \tag{12}$$

where S_O is the number of occupied pixels, and S_T is the total number of pixels. Although there is no mapping between a sphere and a plane without deformations, we find this ratio to be a close approximation of the sphere completeness. Particularly in the visible quadrants, deformations are negligible.

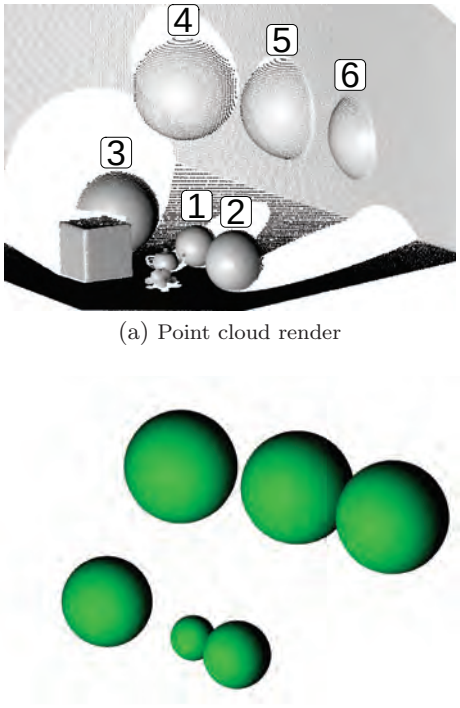
This process of refitting and pruning is applied for every sphere extracted from the accumulator. After all the extracted spheres are processed, the algorithm finishes.

4. Experiments and Discussion

4.1 Datasets

We evaluated EFRANSAC and the proposed method with experiments against synthetic and real data.

The synthetic dataset $\mathcal{M}$, shown in **Fig. 5**, was generated using sensor simulation in Blensor 1.0.18-RC10 [29] and a model with six spheres, among other objects recreating realistic scenarios of occlusion and loss of the spheres' surface. Although the simulation provides organized point clouds, we treat them as unorganized. In the lower part, we can observe a render of the model used to simulate a realistic time-of-flight (TOF) sensor point clouds. The three spheres on the back have the same radius, but we varied their levels of potential visibility from the sensor: 100%, 50% and 25%. Furthermore, as we are aiming to simulate a real sensor as close as possible, we



(a) Point cloud render

(b) Ground truth spheres

Fig. 6 A point cloud example from the $\mathcal{M}$ dataset

set the parameters of the sensor simulation software to match those of the KinectV2.

The simulations occurred at the same sensor position but with variations in the level of noise, with a 0-mean Gaussian noise, and its variance σ set to $[0.004, 0.008, 0.012, \ldots, 0.04]$. These variations are used to modify the coordinates of the sensed points. Since it is a KinectV2 simulation its number of points is 217,088, its Bounding Box Diagonal Distance (BBDD) is around 9.00[m] and its resolution is about 0.014[m]. Where BBDD is the vector length of the extreme points of the bounding box.

Figure 6 shows an example of a point cloud projected from $\mathcal{M}$ that corresponds to the lowest level of Gaussian noise ($\sigma = 0.004$). Note that the size of objects appear to be different from Fig. 5 because their projection method and camera position are different. Their index is shown in the numbers displayed above the spheres points in Fig. 6(a), and the ground truth spheres are shown separately in green as 3D models in Fig. 6(b).

The NDAJ dataset $\mathcal{N}$ consists of 34 point clouds obtained from a FARO® 3D LiDAR scanner, they are highly dense and cover a broad area.

The scanning target is a building of the Engineering campus of the National Defense Academy of Japan (NDAJ). **Figure 7** shows a render of the registration of all the point clouds of the dataset. This render was gen-

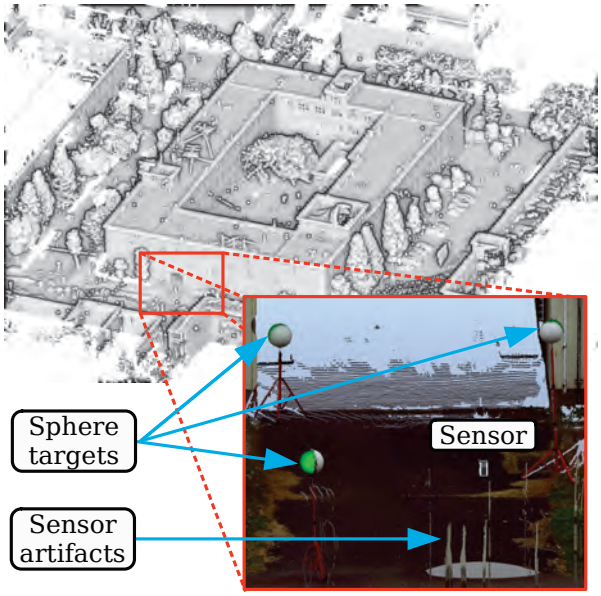


Fig. 7 Rendering of all subsets of the $\mathcal{N}$ dataset

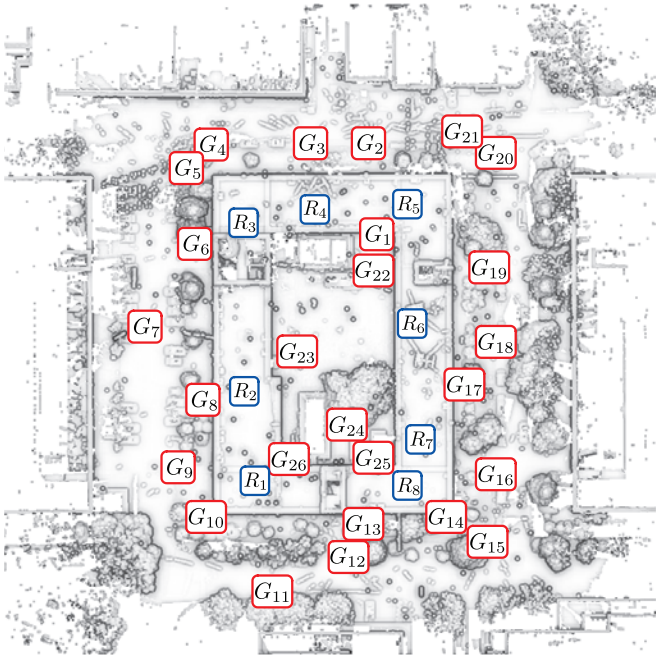
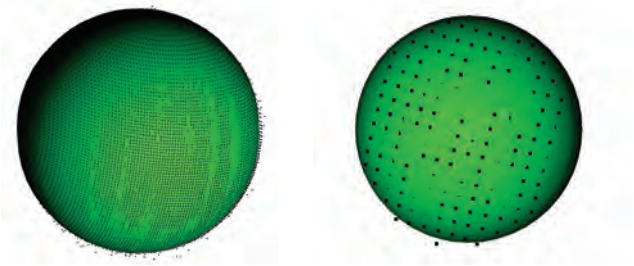


Fig. 8 Sensor locations of the $\mathcal{N}$ dataset

erated by processing all the registered point clouds with a voxel size of 0.05[m] and applying normals estimation with a radius of 0.075[m], then we activated the EDL shader from the qEDL plugin in Cloud Compare.

Figure 7 also shows a close up of a point cloud of $\mathcal{N}$ (G_2) and its location inside the rendering. Three sphere targets that were physically placed near the sensor are shown alongside its ground truth spheres shown in green. These are part of the ground truth spheres set. Additionally, sensor artifacts produced by moving objects in the scene are shown.

The ground truth spheres were generated by using the scanner's software in a user-guided process. The ground truth data spheres of the real dataset are highly precise since it was used to register the point clouds by manually labeling the spheres in the FARO Scene® point cloud processing software.



(a) Nearest sphere: 6.15[m], 4,958 points (b) Farthest sphere: 31[m], 169 points

Fig. 9 Example of sensed spheres and their respective ground truth 3D models of the R_4 (LIDAR) point cloud

Table 2 Datasets information

Dataset	Points[#]	BBDD[m]	CR[cm]
$\mathcal{M}_L$	217,088	8.77	1.00
$\mathcal{M}_M$	217,088	8.88	1.24
$\mathcal{M}_H$	217,088	8.99	1.47
$\mathcal{N}_G$	25,543,762.73 ±3,625,816.05	299.70±60.27	0.44±0.07
$\mathcal{N}_R$	18,261,275.62 ±2,615,343.03	344.15±18.81	0.30±0.04

The location of each categorized scan can be seen in **Fig. 8**. 26 of the point clouds were taken from ground level, numbered from G_1 to G_{26}, and marked with red boxes in Fig. 8. Eight point clouds were taken from the roof of the building, numbered from R_1 to R_8, and marked with blue boxes in Fig. 8.

The ground truth spheres of $\mathcal{N}$ have fixed positions, but depending on the distance from the sensor, the density of their points and their noise can vary enormously. **Figure 9** shows the effect of a sphere sensed from 6.15[m] and 31[m], there is a big difference in the density of their points when they are near or far from the sensor. Therefore, a challenging feature of these long-range point clouds, is their variations in density, which makes it difficult to validate hypothetical spheres by only counting their inliers.

In **Table 2**, we summarize statistical information of both datasets. The synthetic dataset consists of 10 point clouds that resemble the ones gathered from low-cost sensors. The table shows the data from the synthetic clouds with low ($\mathcal{M}_L$), mid ($\mathcal{M}_M$), and high ($\mathcal{M}_H$) levels of Gaussian noise when $\mu = 0$ and σ become [0.004, 0.02, 0.04], respectively. Also, we divided the real point clouds $\mathcal{N}$ dataset into two parts, ground ($\mathcal{N}_G$) and roof ($\mathcal{N}_R$), which separates the point clouds that were scanned at

the ground level and over the building. CR stands for *Cloud Resolution*, which is a measure of the average density of the point cloud and is defined as the mean distance between each point and its nearest neighbor.

4.2 Experimental setup

To evaluate the accuracy, we performed matching of the ground truth spheres S_{gt} and the detected spheres S_{det} coefficients. A match occurs if a sphere in S_{det} is found in S_{gt} by thresholding its Euclidean distance in $\mathbb{R}^4$,

$$\left|S_{det_i} - S_{gt_j}\right| < 0.1[\text{m}]. \tag{13}$$

This threshold provides an average maximum error tolerance of 0.025[m] for the four parameters of a sphere model. To decide this threshold, we ran EFRANSAC on the noisiest point cloud of $\mathcal{M}$ and adjusted it such that a visible bias coming from noise, and affecting the position and radius of spheres do not alter the true positives count TP when most of its inliers are close to the sphere (shown later in visual assessment).

All the detected spheres that satisfy Eq. (13) are TP, then we can define the following metrics from information retrieval[30] precision γ, recall ζ, and F_1 score η,

$$\gamma = \frac{TP}{\#S_{det}} \tag{14}$$

$$\zeta = \frac{TP}{\#S_{gt}} \tag{15}$$

$$\eta = \frac{2TP}{\#S_{det} + \#S_{gt}} \tag{16}$$

where # denotes set cardinality. Precision will max at 1 when we detected only correct spheres, while recall will max at 1 when all the ground truth spheres were detected. η is the harmonic mean between γ and ζ.

4.3 Experiments results and discussion

Both algorithms implementations are in C++, compiled with gcc 7.5 and O3 optimizations. We used the EFRANSAC implementation from the CGAL library 4.11 [25], while the proposed method was implemented mostly using routines from the PCL Library 1.9.1[24]. Since EFRANSAC is nondeterministic, we executed the experiments 50 times and measured both mean and standard deviation.

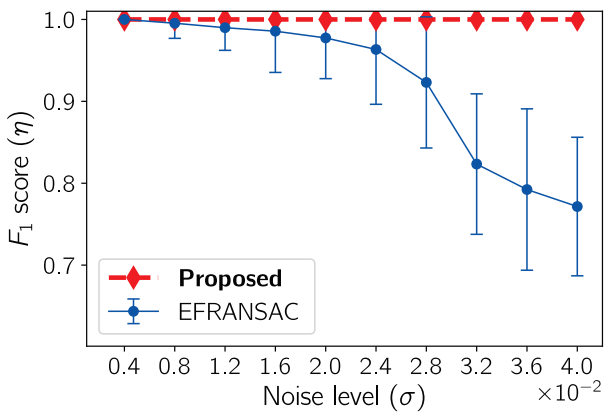
We adjusted the parameters of the evaluated methods based on preliminary experiments to get the most accurate results, following to adjustments in improving processing time without deteriorating their accuracy. **Table 3** and **Table 4** show EFRANSAC and the proposed method parameters, respectively. For each dataset, we provide results of both visual and numerical assessment. For the synthetic datasets, we calculated η for each level of noise when using EFRANSAC and the proposed method. **Figure 10** (a) shows the accuracy of the

Table 3 EFRANSAC parameters

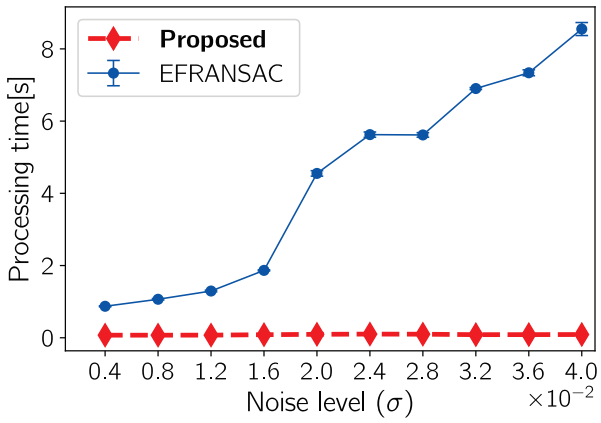
Name	ϵ	C_ϵ	N_{th}	Min. support	z	Normals radius
$\mathcal{M}_L$	0.016	0.80	0.95	700	0.00010	0.04
$\mathcal{M}_M$	0.020	0.80	0.95	1100	0.00005	0.14
$\mathcal{M}_H$	0.040	0.80	0.95	1100	0.00005	0.20
$\mathcal{N}_G$	0.020	0.02	0.10	200	0.00005	0.05
$\mathcal{N}_R$	0.020	0.02	0.10	200	0.00005	0.05

Table 4 Proposed method parameters

Name	V_s	V_{min}	Acc_{res}	ϵ	$Th_{\delta\mu}$	K	B_{bins}
$\mathcal{M}_L$	0.16	20	0.05	0.020	0.4	0.1	50
$\mathcal{M}_M$	0.16	20	0.05	0.020	0.4	0.1	50
$\mathcal{M}_H$	0.18	30	0.05	0.020	0.4	0.1	50
$\mathcal{N}_G$	0.05	50	0.01	0.002	0.3	0.1	25
$\mathcal{N}_R$	0.05	50	0.01	0.002	0.3	0.1	25



(a) η experiment results



(b) Processing time[s] experiment results

Fig. 10 η and processing time experiment results of the $\mathcal{M}$ dataset

Table 5 Detailed results per ground truth sphere of experiments with synthetic data of $\mathcal{M}$.

Sphere	Mean error in $\mathbb{R}^4$[m] (EFRANSAC)	Mean error in $\mathbb{R}^4$[m] (proposed)	Detection rate[%] (EFRANSAC)	Detection rate[%] (proposed)	Note
1	0.1105	**0.0321**	89.8	**100.0**	Small radius (0.2[m])
2	**0.0196**	0.0233	**100.0**	**100.0**	Nearest
3	**0.0155**	0.0173	**100.0**	**100.0**	Occluded
4	0.0219	**0.0138**	99.8	**100.0**	100% Visibility
5	0.0389	**0.0172**	98.0	**100.0**	50% Visibility
6	0.0886	**0.0292**	73.2	**100.0**	25% Visibility

evaluated methods. A value of one means the proposed method detected all the ground truth spheres without false positives, η goes down when either false positives or false negatives drop. The EFRANSAC accuracy drops as the noise worsen, even when we increased its iterations and the radius of normal vectors. On the other hand, the proposed method detected successfully all the spheres with high accuracy. Figure 10 (b) is showing that the proposed method is drastically more efficient than EFRANSAC, which is requiring more processing power as noise increases due to the necessity of expanding the support radius of normal vectors estimation and its iterations.

Table 5 shows the results of per-ground truth sphere in the dataset $\mathcal{M}$. The index of the sphere is shown in the left-most column. The two following columns show their mean error with respect to their closest match of detected spheres in $\mathbb{R}^4$ by matching the ground truth spheres with the one that minimizes the left term of Eq. (13). Also, we highlighted in bold the ones that showed the best results.

EFRANSAC has a slightly better detection error than the proposed method only when its detection rate is competitive. Hence, in this step, EFRANSAC is filtering points that do not agree with the normal vector of the hypothetical model. However, the heavy computational costs associated with computing point-wise normal vectors (see Fig. 10(b)) are not worth a few millimeters of better precision.

On the other hand, in the detection rate of Table 5, we can notice that EFRANSAC fails to detect the sphere with the least visibility. This happens because the normal vectors near the border of the plane wall and the sphere are imprecise. Hence, further reducing the available surface to detect the sphere number six.

Figure 11 shows the η results of the evaluated methods for each of the point cloud groups of the $\mathcal{N}$ dataset. The y-axis of the plot is divided into two because of the poor results of EFRANSAC since it had drastically lower

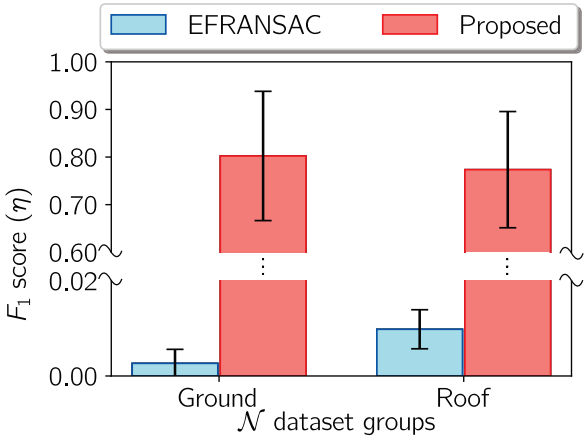


Fig. 11 η results using the $\mathcal{N}$ dataset, higher is better

precision and negatively affected its η. EFRANSAC validates a model by thresholding its inliers count without considering their geometry and completeness. Therefore, the inliers ratio ϕ_r of those point clouds is quite low, and near the machine epsilon, it is obvious that a method that relies only on random sampling like EFRANSAC is failing very badly in the results.

Moreover, a sphere of the dataset usually contains a few hundred of points while the whole point cloud accounts for dozens of millions. This situation forces EFRANSAC to set a lower value of inliers count that supports a sphere candidate, thus detecting numerous false positives and diminishing its precision. The standard approach to address this issue is to increase RANSAC iterations by modifying its probability parameter described in Eq. (2). However, this value becomes dangerously near the machine epsilon for single float precision and thus numerically unstable. Furthermore, its performance decreases in several orders of magnitude.

On the other hand, the proposed method opts for a smarter strategy by robustly measuring the likelihood of regions to be spherical, converging hypothetical spheres from highly spherical regions into an accumulator and filtering by their completeness. It demonstrated superior accuracy even in massive point clouds with very low ϕ_r.

We roughly estimated that this ratio for the $\mathcal{N}$ dataset is less than 0.00005.

Figure 12 illustrates the processing time of the evaluation methods for each point cloud of the $\mathcal{N}$ dataset. The y-axis is divided into two sections with different scales to visualize the processing time results. For better visualization, and unlike the experiments with synthetic data, we are not including the normals vectors estimation time in these results. If we included them, it would be extremely difficult to visualize and compare the high efficiency of the proposed method, since it is astronomically better than EFRANSAC.

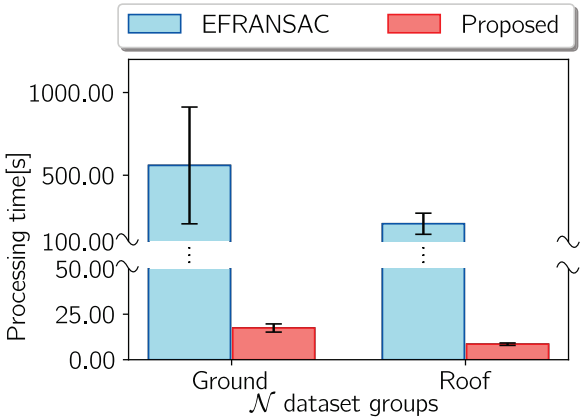


Fig. 12 Processing time[s] evaluation results using the $\mathcal{N}$ dataset, lower is better

Figure 13, **Fig. 14**, and **Fig. 15** show the detailed results of EFRANSAC and the proposed method against the $\mathcal{N}$ dataset. Figure 13 shows the proposed method has superior accuracy due to the low precision of EFRANSAC shown in Fig. 14 because EFRANSAC detected numerous false positives in the $\mathcal{N}$ dataset. Figure 15 is showing how many of the ground truth spheres were detected without considering false positives. The proposed method showed a superior recall in all scenarios. Noticeably, at the mark Ⓐ in Fig. 15, EFRANSAC recall was extremely low, and the proposed method recall was remarkably high. EFRANSAC failed because the ground truth spheres of the point cloud G_{11} are relatively far from the sensor and has a lower density, all the spherical points account for 0.0024% of the points, and a ground truth sphere was partially occluded.

At the mark Ⓑ in Fig. 15, EFRANSAC recall was very close to the proposed method. **Figure 16**(a) shows the R_3 point cloud from $\mathcal{N}$. The green triangle represents the sensor location, and the blue circles the ground truth spheres. One of the ground truth spheres is extremely far from the sensor, its distance is represented by the red dashed line and is about 38.33[m]. Both methods failed

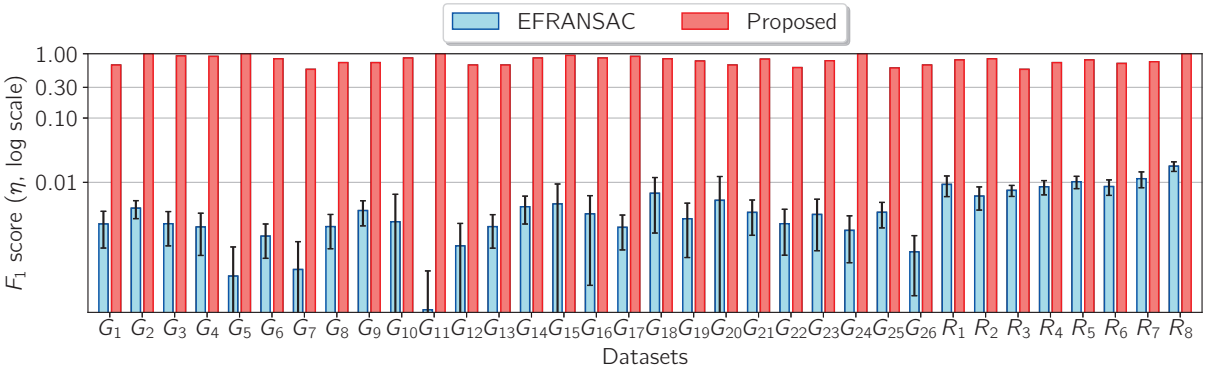


Fig. 13 F_1 score (η) results using the $\mathcal{N}$ dataset, higher is better

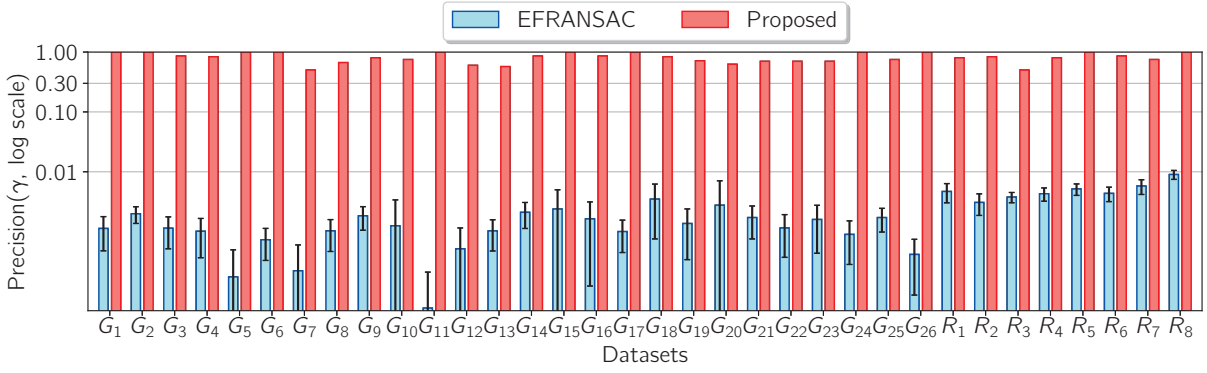


Fig. 14 Precision (γ) results using the $\mathcal{N}$ dataset, higher is better

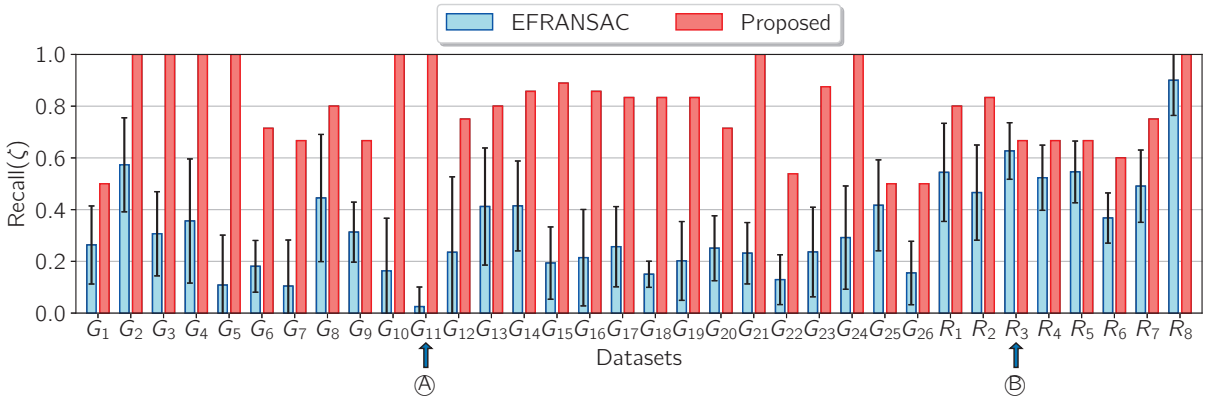


Fig. 15 Recall (ζ) results using the $\mathcal{N}$ dataset, higher is better

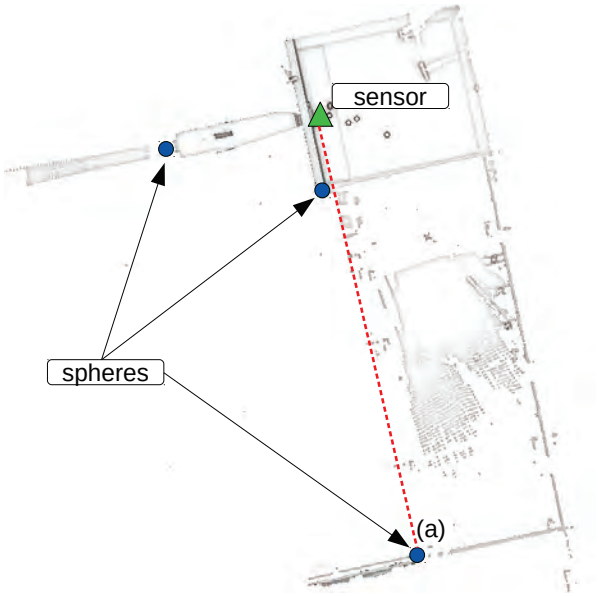


Fig. 16 R_3 point cloud from the $\mathcal{N}$ dataset

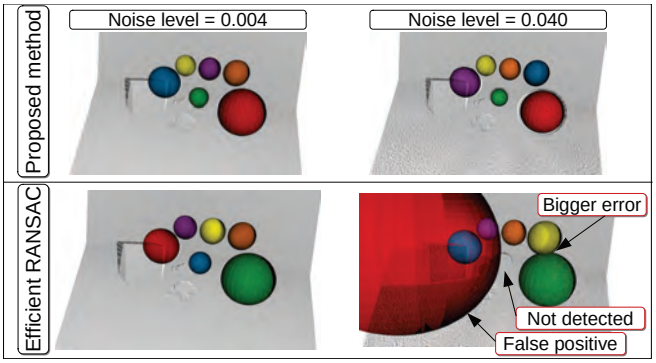


Fig. 17 Graphical comparison: $\mathcal{M}$ dataset

to detect this sphere since it is barely recognizable due to its extremely low density.

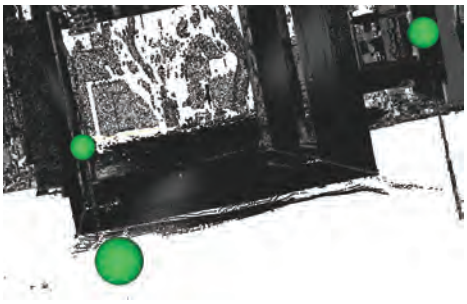
We also show a visual assessment of the experiment results. In **Fig. 17**, we can observe the results of executing the evaluated methods on the synthetic point cloud with the least ($\sigma = 0.004$) and most noise ($\sigma = 0.04$). We assigned a different color to each detected sphere in the order they are given by the evaluated methods. The proposed method detected the spheres flawlessly with high accuracy. EFRANSAC results are very good with low noise scenarios, but as we test for higher levels of noise the number of misdetections and false positives arise. Moreover, as we need to increase thresholds and the support radius for normals estimation, the smallest spheres tend to be undetected, and bigger errors in the spheres coefficients become visible.

Figure 18 shows the ground truth and results of spheres detection of the point cloud G_1. Three ground truth spheres are not visible in Fig. 18(a), which were placed in the central part of the building with a radius of 0.0381[m], neither the proposed method nor EFRANSAC were able to detect. Figure 18(a) shows the detected spheres of both the proposed method and EFRANSAC. Due to the numerous false positives and for better visualization, we omitted EFRANSAC spheres larger than 0.3[m]. Even though the extra filtering we can notice EFRANSAC failed to detect one target sphere and detected several false positives in the background.

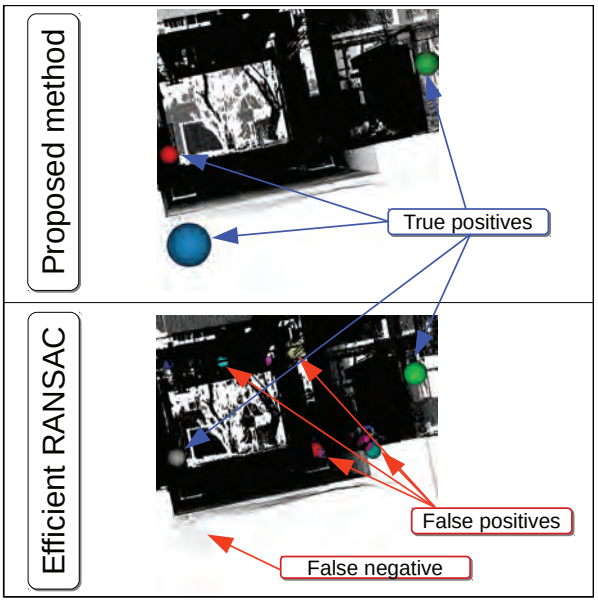
5. Conclusions and Future Works

Sphere detection is a core technique in point cloud processing with applications in computer vision, reconstruction, modeling, among others. However, existing algorithms work with various drawbacks such as fixed radius, low efficiency, and poor accuracy in noisy data with a numerous amount of outliers. To solve these problems, in this paper, we proposed a new sphere detection method based on sliding voxels and Hough voting.

Through experiments we found that the proposed method achieved 50 times faster processing time and is 1.08 times more accurate than EFRANSAC for the synthetic dataset $\mathcal{M}$. In the real 3D LIDAR dataset $\mathcal{N}$, the proposed method achieved 31 times faster processing time and is 183 times more accurate (F-score) than EFRANSAC without including normals estimation processing time. This is because the proposed method can



(a) Ground truth spheres render of G_1



(b) Detection results

Fig. 18 Graphical comparison: $\mathcal{N}$ dataset (G_1)

analyze all the points by employing sliding voxels even for big point clouds with numerous amount of points.

As our future works, we should investigate adaptive settings of the voxel size for various density levels in massive point clouds to further improve its accuracy. Also, we need more comprehensive experiments to analyze the relationship among processing time, total number of points and size of the spheres to give us better insight on the performance and scalability of the sphere detection methods. Furthermore, as promising extensions of this work for real-world applications, we are considering to detect nearly spherical shapes, such as human heads, bone junction, and so on.

References

1) R. Schnabel, R. Wahl, R. Klein: "Efficient RANSAC for Point-Cloud Shape Detection", Proc. of Computer Graphics Forum, Vol. 26, No. 2, pp. 214–226 (2007).
2) J. Chen, H. Lai, C. Lin: "Point Cloud Modeling Using Algebraic Template", International Journal of Innovative Computing, Information and Control, Vol. 7, No. 4, pp. 1521–1532 (2011).
3) C.F. Jekel, G. Venter, M.P.Venter: "Obtaining Non-Linear Orthotropic Material Models for Pvc-Coated Polyester via Inverse Bubble Inflation", Structural and Multidisciplinary Optimization, Vol. 54, No. 4, pp. 927–935 (2016).
4) B. Muralikrishnan, P. Rachakonda, V. Lee, M. Shilling, D. Sawyer, G. Cheok, L. Cournoyer: "Relative Range Error Evaluation of Terrestrial Laser Scanners Using a Plate, a Sphere, and a Novel Dual-Sphere-Plate Target", Measurement, Vol. 111, pp. 60–68 (2017).
5) M. van der Glas, F.M. Vos, C.P. Botha, A.M. Vossepoel: "Determination of Position and Radius of Ball Joints", Proc. of Medical Imaging: Image Processing, pp. 1571–1577 (2002).
6) M. Franaszek, G. S. Cheok, C. Witzgall: "Fast Automatic Registration of Range Images from 3D Imaging Systems Using Sphere Targets", Automation in Construction, Vol. 18, No. 3, pp. 265–274 (2009).
7) Y. Wang, H. Shi, Y. Zhang, D. Zhang: "Automatic Registration of Laser Point Cloud Using Precisely Located Sphere Targets", Journal of applied remote sensing, Vol. 8, No. 1, pp. 1–14 (2014).
8) D. Yun, S. Kim, H. Heo, K.H. Ko: "Automated Registration of Multi-View Point Clouds Using Sphere Targets", Advanced Engineering Informatics, Vol. 29, No. 4, pp. 930–939 (2015).
9) T. Ogundana, C.R. Coggrave, R. Burguete, J.M. Huntley: "Fast Hough Transform for Automated Detection of Spheres in Three-Dimensional Point Clouds", Optical Engineering, Vol. 46, No. 5,pp. 1–11 (2007).
10) A. Abuzaina, M.S. Nixon, J.N. Carter: "Sphere Detection in Kinect Point Clouds via the 3D Hough Transform", International Conference on Computer Analysis of Images and Patterns, pp. 290–297 (2013).
11) M. Camurri, R. Vezzani, R. Cucchiara: "3D Hough Transform for Sphere Recognition on Point Clouds", Machine Vision and Applications, Vol. 25, No. 7, pp. 1877–1891 (2014).
12) C. Sommer, Y. Sun, E. Bylow, D. Cremers: "PrimiTect: Fast Continuous Hough Voting for Primitive Detection", Proc. of the International Conference of Robotics and Automation (ICRA), arXiv preprint arXiv:2005.07457 (2020).
13) P.VC. Hough: "Method and Means for Recognizing Complex Patterns", US Patent 3,069,654 (1962).
14) D.H. Ballard: "Generalizing the Hough Transform to Detect Arbitrary Shapes", Readings in Computer Vision, pp. 714–725 (1987).
15) J. Illingworth, J. Kittler: "A Survey of the Hough Transform", Computer Vision, Graphics, and Image Processing, Vol. 44, No. 1, pp. 87–116 (1988).
16) A. Kaiser, J.A. Ybanez Zepeda, T. Boubekeur: "A Survey of Simple Geometric Primitives Detection Methods for Captured 3D Data", Computer Graphics Forum, Vol. 38, No. 1, pp. 167–196 (2019).
17) M. Kharbat, N. Aouf, A. Tsourdos, B. White: "Sphere Detection and Tracking for a Space Capturing Operation", Proc. of the IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 182–187 (2007).
18) N. Kiryati, Y. Eldar, A.M. Bruckstein: "A Probabilistic Hough Transform", Pattern Recognition, Vol. 24, No. 4, pp. 303–316 (1991).
19) B. Drost, M. Ulrich, N. Navab, S. Ilic: "Model Globally, Match Locally: Efficient and Robust 3D Object Recognition", Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 998–1005 (2010).
20) M. Fischler, R. Bolles: "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Communications of the

ACM, Vol. 24, No. 6, pp. 381–395 (1981).
21) L. Wang, C. Shen, F. Duan, K. Lu: "Energy-Based Automatic Recognition of Multiple Spheres in Three-Dimensional Point Cloud", Pattern Recognition Letters, Vol. 83, pp. 287–293 (2016).
22) T. Birdal, B. Busam, N. Navab, S. Ilic, P. Sturm: "A Minimalist Approach to Type-Agnostic Detection of Quadrics in Point Clouds", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3530–3540 (2018).
23) G. Roth, M.D. Levine: "Extracting Geometric Primitives" CVGIP: Image Understanding, Vol. 58, No. 1, pp. 1–22 (1993).
24) R. Rusu, S. Cousins: "3D is Here: Point Cloud Library (PCL)", Proc. of IEEE International Conference on Robotics and Automation (ICRA), pp. 1–4 (2011).
25) CGAL: Computational Geometry Algorithms Library, https://www.cgal.org (2019).
26) P.H. Torr, A. Zisserman: "MLESAC: A New Robust Estimator with Application to Estimating Image Geometry", Computer Vision and Image Understanding, Vol. 78, No. 1, pp. 138–156 (2000).
27) C.M. Shakarji: "Least-Squares Fitting Algorithms of the NIST Algorithm Testing System", Journal of Research of the National Institute of Standards and Technology, Vol. 103, No. 6, pp. 633–641 (1998).
28) D. Marquardt: "An Algorithm for Least-Squares Estimation of Nonlinear Parameters", Journal of the Society for Industrial and Applied Mathematics, Vol. 11, No. 2, pp. 431–441 (1963).
29) M. Gschwandtner, R. Kwitt, A. Uhl, W. Pree: "BlenSor: Blender Sensor Simulation Toolbox", Proc. of the International Symposium on Visual Computing, pp. 199–208 (2011).
30) C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Natural Language Engineering, Cambridge University Press (2010).
31) Distance Between 3D Lines and Segments: http://geomalgorithms.com/a07-_distance.html (2020).

Appendix A Sphere estimation from two oriented points

It starts by detecting the shortest segment between two lines defined by two oriented points P_1 and P_2 using Sunday's approach[31]. **Figure A. 1** illustrates how the center $C = \{C_x, C_y, C_z\}$ of the sphere can be calculated from

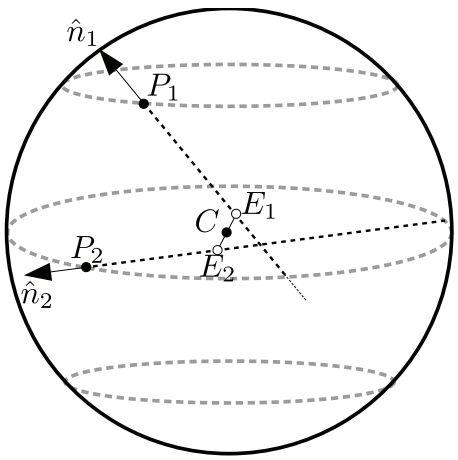


Figure A. 1 Sphere estimated with two points and their normal vectors

the line segments $L_1 = \overline{P_1\hat{n_1}}$ and $L_2 = \overline{P_2\hat{n_2}}$. C is the midpoint between the shortest line segment $L^* = \overline{E_1E_2}$ between L_1 and L_2. L^* is the shortest line segment if and only if it is orthogonal to L_1 and L_2 simultaneously. To estimate the endpoints E_1 and E_2, we need to parameterize L_1 and L_2 as follows,

$$\begin{aligned} E_1(s) &= P_1 + s\hat{n}_1, \text{ where } s \in \mathbb{R}, \\ E_2(t) &= P_2 + t\hat{n}_2, \text{ where } t \in \mathbb{R}. \end{aligned} \tag{A–1}$$

Then, we need to estimate s^* and t^* such that the resulting line segment

$$\begin{aligned} L^*(s^*, t^*) &= E_1(s^*) - E_2(t^*) \\ L^*(s^*, t^*) &= P_1 - P_2 + s^*\hat{n}_1 - t^*\hat{n}_2 \end{aligned} \tag{A–2}$$

is parallel to both L_1 and L_2.

The angle between L^* and $\{L_1, L_2\}$ can be derived from the equations

$$\begin{aligned} \hat{n}_1 \cdot L^*(s^*, t^*) &= \|\hat{n}_1\|_2 \, \|L^*(s^*, t^*)\|_2 \cos\theta \text{ and} \\ \hat{n}_2 \cdot L^*(s^*, t^*) &= \|\hat{n}_2\|_2 \, \|L^*(s^*, t^*)\|_2 \cos\phi. \end{aligned} \tag{A–3}$$

Since we need to fix $\theta = \phi = 90[deg]$, then substituting $\cos(90[deg]) = 0$ in Eq. (A–3) results in the following equations

$$\hat{n}_1 \cdot L^*(s^*, t^*) = 0, \tag{A–4}$$

$$\hat{n}_2 \cdot L^*(s^*, t^*) = 0. \tag{A–5}$$

In other words, given the parameterized points $E_1(s)$ and $E_2(t)$ from Eq. (A–1), we need to find s and t such that the conditions of Eq. (A–4) and Eq. (A–5) apply simultaneously and will name them s^* and t^*.

By substituting Eq. (A–1) in Eq. (A–4) and Eq. (A–5), we get

$$s(\hat{n}_1 \cdot \hat{n}_1) - t(\hat{n}_1 \cdot \hat{n}_2) = -\hat{n}_1(P_1 - P_2) \text{ and} \tag{A–6}$$

$$s(\hat{n}_1 \cdot \hat{n}_2) - t(\hat{n}_2 \cdot \hat{n}_2) = -\hat{n}_2(P_1 - P_2). \tag{A–7}$$

For the sake of simplicity we are going to apply the following substitutions to Eq. (A–6) and Eq. (A–7)

$$\begin{aligned} a &= \hat{n}_1 \cdot \hat{n}_1 \\ b &= \hat{n}_1 \cdot \hat{n}_2 \\ c &= \hat{n}_2 \cdot \hat{n}_2 \\ d &= \hat{n}_1(P_1 - P_2) \\ e &= \hat{n}_2(P_1 - P_2). \end{aligned} \tag{A–8}$$

Then Eq. (A–6) and Eq. (A–7) become

$$sa - tb = -d \tag{A–9}$$

$$sb - tc = -e. \tag{A–10}$$

To solve s and t we apply Cramer's rule and obtain

$$s = \frac{cd - be}{b^2 - ac} \tag{A–11}$$

$$t = \frac{db - ae}{b^2 - ac}. \tag{A–12}$$

Noticeably, if $\hat{n}_1$ and $\hat{n}_2$ are nearly parallel, the denominator $b^2 - ac$ tends to 0, and the solution is undefined. In this case, we ignore it if $b^2 - ac$ is close to the machine epsilon.

After we estimate appropriate values of s and t, the center of the sphere C is

$$C = \frac{1}{2}(E_1 + E_2), \tag{A–13}$$

and its radius

$$r = \frac{1}{2}(\|P_1 - C\|_2 + \|P_2 - C\|_2). \tag{A–14}$$

(Received May 8, 2020)
(Revised June 20, 2020)

Jaime SANDOVAL
(*Student Member*)
He received his B.E. degree in Computer Systems Engineering from the Universidad del Valle del Fuerte (Mexico) in 2009, and his M.E. degree in Electrical and Electronic Engineering from Shinshu University in 2017. Currently, he is a Ph.D. student in the Interdisciplinary Graduate School of Science and Technology of Shinshu University with a major in Systems Development Engineering. His research interests are 3D Point Clouds Processing, Computer Vision and Image Processing.

Kazuma UENISHI (*Member*)
He received his B.E. degree in Computer Science in 2008, and his M.E. degree in Mathematics and Computer Science in 2014 from the National Defense Academy of Japan. Currently, he is enrolled as a Ph.D. student in the Interdisciplinary Graduate School of Science and Technology of Shinshu University with a major in Systems Development Engineering. He is pursuing research in 3D Point Clouds Processing.

Munetoshi IWAKIRI (*Member*)
He received his B. E. degree in Computer Science in 1993, and received his M. E. degree in Mathematics and Computer Science from the National Defense Academy of Japan in 1998. In 1999, he joined the Department of Computer Science, National Defense Academy of Japan, as a Research Associate. In 2002, he received Dr. Eng. degree from Keio University, Tokyo, Japan. In 2005 he became Lecturer and in 2015 he became Associate Professor in the same institution. He is pursuing research related to Multimedia Processing and Information Security. He is a member of the Information Processing Society of Japan.

Kiyoshi TANAKA (*Fellow*)
He received his B.S and M.S. degrees in Electrical Engineering and Operations Research from the National Defense Academy, Yokosuka, Japan, in 1984 and 1989, respectively. In 1992, he received the Dr. Eng. degree from Keio University, Tokyo, Japan. In 1995, he joined the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano, Japan, and currently, he is a full professor in the academic assembly (Institute of Engineering) of Shinshu University. He is the Vice-President of Shinshu University as well as the director of the Center for Global Education and Collaboration of Shinshu University. His research interests include image and video processing, 3D point cloud processing, information hiding, human visual perception, evolutionary computation, multi-objective optimization, smart grid, and their applications. Currently, he is the president of IIEEJ, a fellow of IIEEJ, a member of IEEE, IEICE, IPSJ, JSEC, and so on.

Call for Papers

Special Issue on
Image-related Technology for Realizing Immersive Media

IIEEJ Editorial Committee

Research on immersive media, which is an advanced system with such as all-sky video (360-degree video), VR, AR, and MR, is being actively conducted. In addition, the development of a head-mounted display (HMD) and the study of specific user interfaces are underway to fully demonstrate the appeal and the benefit of immersive media. In ISO / IEC, Coded Representation of Immersive Media (MPEG-I) has been studied for realizing immersive media in MPEG team, and though its standardization is being completed, further technologies are expected to be standardized. ITU-T SG16 is also promoting the recommendation of ultra-high presence live experience (ILE: Immersive Live Experience), and the momentum for systematization is increasing.

In this special issue, we are looking for a wide range of papers on elemental technologies that realize immersive media and system development papers on research to realize immersive high-presence systems by applying these technologies.

1. Topics covered include but not limited to

   All-sky video (360-degree video), Free viewpoint Image, Point cloud, Light field, Holography, Head-mounted display (HMD), User interface, User experience, Usability, Interaction, Immersive content, VR, AR, MR, SR, xR, CG, Image processing, Image coding, Computer vision, Deep learning, MPEG-I, ILE

2. Treatment of papers

   Submission paper style format and double-blind peer review process are the same as an ordinary contributed paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as an ordinary contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:

   IEEJ Transactions on Image Electronics and Visual Computing Vo.9, No.2 (December 2021)

4. Submission Deadline

   Monday, May 31, 2021

5. Contact details for Inquires:

   IIEEJ Office E-mail: hensyu@iieej.org

6. Online Submission URL: http://www.editorialmanager.com/iieej/

Call for Papers

Special Issue on CG & Image Processing Technologies Supporting and Expanding Human Creativities

IIEEJ Editorial Committee

The AI technology is expected to become a key technology for solving social problems and SDGs (Sustainable Development Goals), such as declining birthrate, aging population, shortage of labor, depopulation area, and so on, to which mature society, especially Japan is facing. The application scope of the technology in the field of image processing is expanding beyond the image generation and object recognition to include areas related to creativity, such as attribute conditioned image generation, super-resolution, image colorization, and line art coloring.

On the other hand, deep learning techniques don't always have good nature at interpretability and explainability of results required for creativity processing. Also, there is still issues to be addressed in AI technology, such as the ineffectiveness to the human-computer interface field by a simple application. So, technologies to support and expand the human creativity are gathering much concern, and quite promising in wide variety of applications.

This special issue of the paper targets various image-related technologies that support and expand creativities, and calls for papers and system development papers that cover not only applications of deep learning but also other technologies.

1. Topics covered include but not limited to
   Image Processing, Image Recognition, Image Detection, Pattern Recognition,
   Computer Graphics, Visualization, Binocular Vision, 3D image processing
   Computer Vision, Big Data, Image Data Bases,
   Machine Learning, Deep Learning, Understandability, Explainability,
   Creativity, Usability, Interpretability, Human Interface and Interaction, User Experience, Ubiquitous,
   Other related fundamental / application / systemized technologies.

2. Treatment of papers
   Submission paper style format and double-blind peer review process are the same as an ordinary contributed paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as an ordinary contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:
   IEEJ Transactions on Image Electronics and Visual Computing Vo.10, No.1 (June 2022)

4. Submission Deadline:
   Tuesday, November 30, 2021

5. Contact details for Inquires:
   IIEEJ Office E-mail: hensyu@iieej.org

6. Online Submission URL: http://www.editorialmanager.com/iieej/

First Call for Papers

The 7th IIEEJ International Conference on Image Electronics and Visual Computing 2021 (IEVC2021)

Shiretoko (Shari), Hokkaido, Japan / Sept. 8-11, 2021

https://www.iieej.org/en/ievc2021/

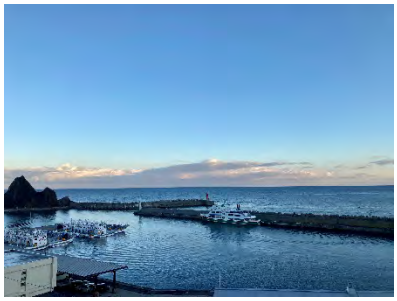
Purpose:

The International Conference on Image Electronics and Visual Computing 2021 (IEVC2021) will be held in Shiretoko (Shari), Hokkaido, Japan, on Sept. 8-11, 2021, as the 7th international academic event of the Institute of Image Electronics Engineers of Japan (IIEEJ) based on the great success of IEVC2007 held in Cairns, Australia, IEVC2010 held in Nice, France, IEVC2012 held in Kuching, Malaysia, IEVC2014 held in Koh Samui, Thailand, IEVC2017 in Danang, Vietnam, and IEVC2019 held in Bali, Indonesia. The conference aims to bring together researchers, engineers, developers, and students from various fields in both academia and industry for discussing the latest researches, standards, developments, implementations and application systems in all areas of image electronics and visual computing.

Topics:

The conference will cover a broad set of research topics including, but not limited to, the following:

- ✧ Image and video coding
- ✧ Image analysis and recognition
- ✧ Segmentation and classification
- ✧ Computer vision
- ✧ Image restoration
- ✧ 3D image processing
- ✧ Object detection
- ✧ Motion analysis
- ✧ Data hiding
- ✧ Bioinformatics and authentication
- ✧ Image and video retrieval
- ✧ Digital museum, digital archiving
- ✧ Content delivery network
- ✧ Image assessment
- ✧ Printing and display technologies
- ✧ Visual communication
- ✧ Mobile image communication
- ✧ Hardware and software implementation
- ✧ Modeling
- ✧ Rendering
- ✧ Visualization
- ✧ Animation
- ✧ Interaction
- ✧ Non-photorealistic rendering
- ✧ Content production
- ✧ Virtual reality, augmented reality, and mixed reality
- ✧ Artificial intelligence and deep learning
- ✧ Security and privacy
- ✧ Big data and Cloud computing
- ✧ International standards



Paper submission:

The official language is English, and authors should submit their papers as PDF through the online submission system, which will be available around Jan. 2021 at the following IEVC2021 official website:

https://www.iieej.org/en/ievc2021/

The paper submission guide and IEVC formats (TeX format / MS Word format) will be also provided at this site. The organizing committee particularly encourages graduate students to present their works in the special sessions that are now planned by the committee of the conference.

General Papers:

The general papers category is divided into two types: journal track and conference track.

✧ Journal Track:

Journal track aims to publish the papers on the journal in addition to the publishing in the conference, with a quick review process. This type of paper will appear in a special issue on "Journal Track Papers in IEVC2021" in the IIEEJ Transactions on Image Electronics and Visual Computing, Vol. 9, No. 2 (December 2021), if accepted through the journal review process. The authors have to prepare two types of papers different in the amount, the paper for the conference and the paper for the journal. The latter one is the extended version of the former one. **Note that the paper for the journal should follow the "guidance for paper submission" available from the website of IIEEJ, to be finally published in the IIEEJ Transactions.**

Important Dates

- Pre-Entry Submission (title, authors, 100 words abstract):	April 2, Friday, 2021
- Paper Submission (2-4 pages, for the conference):	April 16, Friday, 2021
- Paper Submission (6-8 pages, for the journal):	May 28, Friday, 2021
- Notification of Conference Acceptance:	June 18, Friday, 2021
- Camera-Ready Paper (2-4 pages, for the conference):	July 9, Friday, 2021

✧ Conference Track:

Conference track aims to present the papers about recent results and preliminary work at IEVC2021. The authors are required to submit a paper of which length is 2-4 pages. Accepted papers will be published both in the online proceedings of IEVC2021 (indexed by J-stage) and in the USB proceedings. Rejected papers in the conference track can be resubmitted as late breaking papers.

Important Dates

- Pre-Entry Submission (title, authors, 100 words abstract):	April 2, Friday, 2021
- Paper Submission (2-4 pages):	April 16, Friday, 2021
- Notification of Acceptance:	June 18, Friday, 2021
- Camera-Ready Paper (2-4 pages):	July 9, Friday, 2021

Late Breaking Papers:

All suitably submitted papers for this category will be accepted for the conference. The authors must submit an abstract of which length is 1-2 pages, and select one from the following two types: 1) Technical papers or 2) Art/Demo papers. All the registered papers as late breaking papers will be published only in the USB proceedings of IEVC2021.

Important Dates

- Pre-Entry Submission (title, authors):	June 23, Wednesday, 2021
- Abstract Submission (1-2 pages):	June 28, Monday, 2021
- Notification of Acceptance:	July 5, Monday, 2021
- Camera-Ready Paper (1-2 pages):	July 9, Friday, 2021

Further information:

After the conference, the Trans. on IEVC of IIEEJ is planning a forthcoming special issue on "Extended Papers Presented in IEVC2021", which will be published in June 2022. More detailed information will be notified on the IEVC2021 website and the Journal of IIEEJ.

Revised: January 6, 2017

Revised: July 6, 2018

Guidance for Paper Submission

1. Submission of Papers

(1) Preparation before submission

- The authors should download "Guidance for Paper Submission" and "Style Format" from the "Academic Journals", "English Journals" section of the Society website and prepare the paper for submission.
- Two versions of "Style Format" are available, TeX and MS Word. To reduce publishing costs and effort, use of TeX version is recommended.
- There are four categories of manuscripts as follows:
  - Ordinary paper: It should be a scholarly thesis on a unique study, development or investigation concerning image electronics engineering. This is an ordinary paper to propose new ideas and will be evaluated for novelty, utility, reliability and comprehensibility. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Short paper: It is not yet a completed full paper, but instead a quick report of the partial result obtained at the preliminary stage as well as the knowledge obtained from the said result. As a general rule, the authors are requested to summarize a paper within four pages.
  - System development paper: It is a paper that is a combination of existing technology or it has its own novelty in addition to the novelty and utility of an ordinary paper, and the development results are superior to conventional methods or can be applied to other systems and demonstrates new knowledge. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Data Paper: A summary of data obtained in the process of a survey, product development, test, application, and so on, which are the beneficial information for readers even though its novelty is not high. As a general rule, the authors are requested to summarize a paper within eight pages.
- To submit the manuscript for ordinsry paper, short paper, system development paper, or data paper, at least one of the authors must be a member or a student member of the society.
- We prohibit the duplicate submission of a paper. If a full paper, short paper, system development paper, or data paper with the same content has been published or submitted to other open publishing forums by the same author, or at least one of the co-authors, it shall not be accepted as a rule. Open publishing forum implies internal or external books, magazines, bulletins and newsletters from government offices, schools, company organizations, etc. This regulation does not apply to a preliminary draft to be used at an annual meeting, seminar, symposium, conference, and lecture meeting of our society or other societies (including overseas societies). A paper that was once approved as a short paper and being submitted again as the full paper after completion is not regarded as a duplicate submission.

(2) Submission stage of a paper

- Delete all author information at the time of submission. However, deletion of reference information is the author's discretion.
- At first, please register your name on the paper submission page of the following URL, and then log in again and fill in the necessary information. Use the "Style Format" to upload your manuscript. An applicant should use PDF format (converted from dvi of TeX or MS Word

format) for the manuscript. As a rule, charts (figures and tables) shall be inserted into the manuscript to use the "Style Format". (a different type of data file, such as audio and video, can be uploaded at the same time for reference.)

http://www.editorialmanager.com/iieej/

- If you have any questions regarding the submission, please consult the editor at our office.

Contact:
Person in charge of editing
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-Ku, Tokyo 116-0002, Japan
E-mail: hensyu@iieej.org
Tel: +81-3-5615-2893, Fax: +81-3-5615-2894

2. Review of Papers and Procedures

(1) Review of a paper

- A manuscript is reviewed by professional reviewers of the relevant field. The reviewer will deem the paper "acceptance", "conditionally acceptance" or "returned". The applicant is notified of the result of the review by E-mail.
- Evaluation method

  Ordinary papers are usually evaluated on the following criteria:
  - ✓ Novelty: The contents of the paper are novel.
  - ✓ Utility: The contents are useful for academic and industrial development.
  - ✓ Reliability: The contents are considered trustworthy by the reviewer.
  - ✓ Comprehensibility: The contents of the paper are clearly described and understood by the reviewer without misunderstanding.

  Apart from the novelty and utility of an ordinary paper, a short paper can be evaluated by having a quickness on the research content and evaluated to have new knowledge with results even if that is partial or for specific use.

  System development papers are evaluated based on the following criteria, apart from the novelty and utility of an ordinary paper.
  - ✓ Novelty of system development: Even when integrated with existing technologies, the novelty of the combination, novelty of the system, novelty of knowledge obtained from the developed system, etc. are recognized as the novelty of the system.
  - ✓ Utility of system development: It is comprehensively or partially superior compared to similar systems. Demonstrates a pioneering new application concept as a system. The combination has appropriate optimality for practical use. Demonstrates performance limitations and examples of performance of the system when put to practical use.

  Apart from the novelty and utility of an ordinary paper, a data paper is considered novel if new deliverables of test, application and manufacturing, the introduction of new technology and proposals in the worksite have any priority, even though they are not necessarily original. Also, if the new deliverables are superior compared to the existing technology and are useful for academic and industrial development, they should be evaluated.

(2) Procedure after a review

- In case of acceptance, the author prepares a final manuscript (as mentioned in 3.).
- In the case of acceptance with comments by the reviewer, the author may revise the paper in consideration of the reviewer's opinion and proceed to prepare the final manuscript (as

mentioned in 3.).

- In case of conditional acceptance, the author shall modify a paper based on the reviewer's requirements by a specified date (within 60 days), and submit the modified paper for approval. The corrected parts must be colored or underlined. A reply letter must be attached that carefully explains the corrections, assertions and future issues, etc., for all of the acceptance conditions.
- In case a paper is returned, the author cannot proceed to the next step. Please look at the reasons the reviewer lists for the return. We expect an applicant to try again after reviewing the content of the paper.

(3) Review request for a revised manuscript

- If you want to submit your paper after conditional acceptance, please submit the reply letter to the comments of the reviewers, and the revised manuscript with revision history to the submission site. Please note the designated date for submission. Revised manuscripts delayed more than the designated date be treated as new applications.
- In principle, a revised manuscript willl be reviewed by the same reviewer. It is judged either accceptance or returned.
- After the judgment, please follow the same procedure as (2).

3. Submission of final manuscript for publication

(1) Submission of a final manuscript

- An author, who has received the notice of "Acceptance", will receive an email regarding the creation of the final manuscript. The author shall prepare a complete set of the final manuscript (electronic data) following the instructions given and send it to the office by the designated date.
- The final manuscript shall contain a source file (TeX edition or MS Word version) and a PDF file, eps files for all drawings (including bmp, jpg, png), an eps file for author's photograph (eps or jpg file of more than 300 dpi with length and breadth ratio 3:2, upper part of the body) for authors' introduction. Please submit these in a compressed format, such as a zip file.
- In the final manuscript, write the name of the authors, name of an organizations, introduction of authors, and if necessary, an appreciation acknowledgment. (cancel macros in the Style file)
- An author whose paper is accepted shall pay a page charge before publishing. It is the author's decision to purchase offprints. (ref. page charge and offprint price information)

(2) Galley print proof

- The author is requested to check the galley (hard copy) a couple of weeks before the paper is published in the journal. Please check the galley by the designated date (within one week). After making any corrections, scan the data and prepare a PDF file, and send it to our office by email. At that time, fill in the Offprint Purchase Slip and Copyright Form and return the scanned data to our office in PDF file form.
- In principle, the copyrights of all articles published in our journal, including electronic form, belong to our society.
- You can download the Offprint Purchase Slip and the Copyright Form from the journal on our homepage. (ref. Attachment 2: Offprint Purchase Slip, Attachment 3: Copyright Form)

(3) Publication

- After final proofreading, a paper is published in the Academic journal or English transaction (both in electronic format) and will also be posted on our homepage.

IIEEJ Trans. on Image Electronics and Visual Computing

Vol. 8, No. 2 2020

Editor in Chief: Mei Kodama
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan

Print: ISSN 2188-1898
Online: ISSN 2188-1901
CD-ROM: ISSN 2188-191x