A Robust Gesture Recognition Using Depth Data

Hironori Takimoto, Jaemin Lee, and Akihiro Kanagawa

***Abstract*—In this paper, we propose a novel gesture recognition system using depth data captured by Kinect sensor. Conventionally, the features which have been used for hand gesture recognition are divided into two parts, hand shape features and arm movement features. However, these traditional features are not robust for environmental changing such as individual differences in body size, camera position and so on. In this paper, we propose a novel hand gesture recognition system using depth data, which is robust for environmental changing. Our approach involves an extraction of hand shape features based on gradient value instead of conventional 2D shape features, and arm movement features based on angles between each joints. In order to show the effectiveness of the proposed method, a performance is evaluated comparing with the conventional method by using Japanese sign language.**

***Index Terms*—Image processing, hand gesture recognition, depth sensor, HMM.**

I. Introduction

To achieve natural human-computer interaction (HCI), a human hand can be considered as an input device. In particular, hand gesture is a powerful and natural human-to-human communication modality involving two major factors: hand shape and arm movement. Building natural interaction between human and computer required accurate hand gesture recognition system as an interface, where the recognition gestures can employ for controlling a robot or conveying meaningful information [1,2].

However, the human hand is a complex articulated object consisting of many connected parts and joints. From the viewpoint of engineering systems, hand shape with fingers intricately vary because of the multiple joint structure, and estimation of the hand shape is very difficult because masked or occluded regions are produced by the palm of the hand and each finger [3]. Therefore, many researches have tried with different instruments and equipment to measure hand movements like gloves, cameras, 3D scanners and depth sensors [3]-[9]. Conventional methods are categorized into two major approaches: 3D model-based and 2D appearance-based approaches. In typical 3D model-based methods, data glove-based devices such as the CyberGlove have been used to capture human hand motion and shape [4]. However, the gloves and their attached wires are still cumbersome and awkward for users to wear.

Manuscript received March 2, 2013; revised April 25, 2013. This research was supported by Telecommunications Advancement Foundation and JSPS, Grant-in-Aid for Scientific Research (Grant-in-Aid for Scientific Research (C)), Japan.

Hironori Takimoto and Akihiro Kanagawa are with Okayama Prefectural University, 111, Kuboki, Soja, Okayama, 719-1197, Japan (e-mail: takimoto@c.oka-pu.ac.jp; kanagawa@c.oka-pu.ac.jp).

Lee Jaemin is with SuperSoftware Co., Ltd., 1-7-13, Ebisu, Shibuya-ku, Tokyo, 150-0013, Japan (e-mail: zaemin2@gmail.com).

In recent years, gesture recognition methods using depth sensor such as Kinect sensor and TOF sensor have been proposed. Conventionally, features which have been used for hand gesture recognition are divided into two parts, hand shape and arm movement [2]. Sato *et al.* proposed Japanese sign language recognition system using TOF sensor [5]. In this method, only the ratio and contour based on 2D feature are used as hand shape feature for gesture recognition. However, it is difficult to classify a hand shape by only using 2D hand features because the human hand is a complex articulated object consisting of many connected parts and joints. Furthermore, conventional arm movement features are not robust for environmental changing such as individual differences in body size, distance between sensor and human's hand, because only coordinates of centroid of both hand region are used as arm movement feature.

In this paper, we propose a novel hand gesture recognition system using depth data which is robust for environmental changing. Our approach involves an extraction of hand shape features based on gradient value instead of conventional 2D shape features, and arm movement features based on angles and angular velocity between each joints. In order to show the effectiveness of the proposed system, the performance for the proposed method is evaluated by comparing with the conventional method. In addition, we perform an evaluation experiment by using Japanese sign language.

II. Related Works

Many hand gesture recognition methods have been widely studied. Generally, conventional methods are categorized into two major approaches: 3D model-based and 2D appearance-based approaches. In 3D model-based method, a glove which mechanical or optical sensors were attached is employed to measure a finger flection [6]. However, the gloves and their attached wires are still cumbersome and awkward for users to wear. In current state-of-the-art vision-based hand tracking and gesture recognition, the research is more focused on tracking the bare hand and recognizing hand gestures without the help of any markers or gloves.

A 3D scanner can be used to extract features from hand shape [7]. However, 3D scanners are often too expensive for regular users, and moreover, the measurement of hand shape using a 3D scanner requires a lot of time. Takimoto et al. proposed hand posture recognition system using multiple camera system [8]. But, it is difficult to detect hand regions from images captured under complex background. Although Ueda *et al.* [9] prepared a voxel model for determining hand-finger posture using multiple cameras, it is difficult to estimate hand posture with a complex background.

In recent years, gesture recognition methods using depth sensor such as Kinect sensor and TOF sensor have been

DOI: 10.7763/IJMLC.2013.V3.312

proposed. Furukawa *et al.* proposed hand and fingertip tracking using depth data obtained from Kinect sensor [10]. Also, Sato et al. proposed Japanese sign language recognition system using 2D appearance feature and TOF sensor. In this method, only the ratio and contour based on 2D feature are used for hand shape feature. However, it is difficult to classify a hand shape by only using 2D hand features because the human hand is a complex articulated object consisting of many connected parts and joints. Considering the global hand pose and each finger joint, human hand motion has roughly 27 degrees of freedom (DOF) [11]. Furthermore, conventional arm movement features are less robust for environmental changing such as individual differences in body size, distance between sensor and hand, because coordinates of centroid of both hand are used as arm movement feature.

III. Proposed Gesture Recognition Method

Fig. 1 shows the procedure of the proposed gesture recognition system, which consists of four main phases. We obtain depth data and 3D coordinates of six joints from sensor. By using 3D coordinate of hand joints, a hand region is cropped from depth data. Next, we extract hand shape features and arm movement features. Finally, hand gestures are recognized by using HMM. In this section, each process of the proposed method is described.

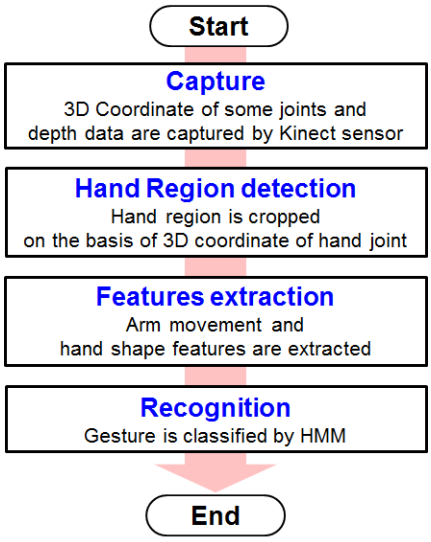
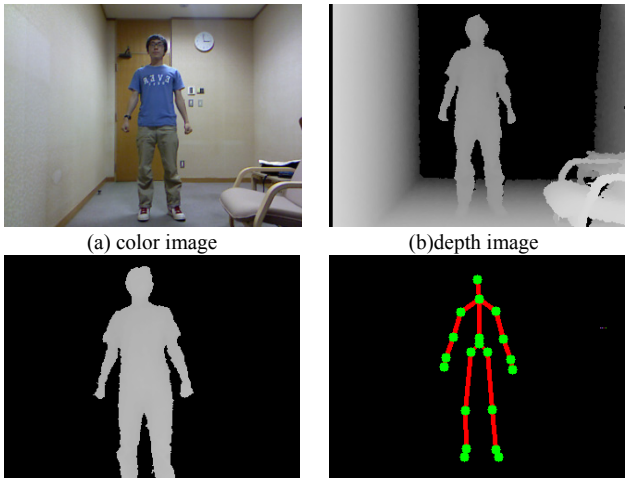


Fig. 1. Procedure of gesture recognition

A. Input Device

In this paper, we use Kinect sensor as input depth sensor. The Kinect sensor is comprised of IR projector, depth image CMOS for depth and color image CMOS for RGB data. Sensors compute the depth of the space while the RGB camera is used to capture the images. The depth data and RGB image of the object could be obtained at the same time. The specification of Kinect sensor is shown in Table I. The depth output of this sensor is of 11 bit with 2048 levels of sensitivity.

We use Kinect SDK which is a software development kit released from Microsoft. Fig. 2 shows RGB image, visualized depth data, user image and skeleton model obtained from SDK. The skeleton model consists of 3D-coordinates of 20 joints shown in Table II. In [12], the accuracy of joints estimation in skeleton model is 73%. In our study, we use six joints which are Hand Left, Hand Right, Wrist Left, Wrist Right, Shoulder Left and Shoulder Right obtained from skeleton model. In addition, the user image without background pixels is used for gesture recognition.



(a) color image (b)depth image

(c) user image (d) skeleton model

Fig. 2. The obtained data from Kinect sensor

TABLE I: Specification of Depth Sensor

Resolution of color image	640×480 pix.
Resolution of depth image	320×240 pix.
Resolution (z axis)	10mm @ 2000mm
Frame rate	30 fps
Horizontal angle	43 degree
Vertical angle	57 degree

TABLE II: Joints of Skeleton Model

Head	Shoulder Center	Spine
Hip Right	Knee Right	Ankle Right
Hip Left	Knee Left	Ankle Left
Elbow Right	Shoulder Right	Wrist Right
Elbow Left	Shoulder Left	Wrist Left
Hip Center	Foot Right	Foot Left
Hand Right	Hand Left	

B. Hand Detection

In the proposed system, both hands regions are cropped on the basis of centroid coordinates of both hands obtained from skeleton model. Example of proposed hand cropping is shown in Fig. 3(a) and (b). However, when subject's hand exists on far from the sensor, the captured hand region is small in the captured image. Therefore, we dynamically determine a window size of hand cropping according to the distance between subject's hand and a sensor. The field of view of this sensor depends on the horizontal and vertical angles which are shown in Table I. Here, we measure the distance between the sensor and hand because space resolution depends on the distance between the sensor and the hand. Then, the range that each pixel expresses is calculated by:

$$x_{world} = 0.00354 \times z_{world} \quad (1)$$

where z_{world} [mm] is depth value from the sensor and x_{world} [mm] is length that each pixel expresses in real world. We define a hand size in real-world as 250×250 [mm]. The window size of hand region is defined by

$$k_{size} = \frac{250}{x_{world}} \tag{2}$$

The result of hand extraction is shown in Fig. 3(b). However, un-hand elements which are body and head are included in cropped hand region. In order to remove un-hand pixels, we create a histogram which is depth distribution in cropped image Fig. 3(b). The created histogram is shown in Fig 4. In this histogram, we assume that the block which are near to the sensor to be a hand. Therefore, we classify into two classes by using Otsu's threshold [14]. After that, un-hand groups except the nearest group are removed from Fig. 3(b). As a result, we obtain the optimum hand region which is shown in Fig. 3(c).

(a) user image acquired from SDK (b) before (c) after

Fig. 3. Example of hand detection by Otsu threshold

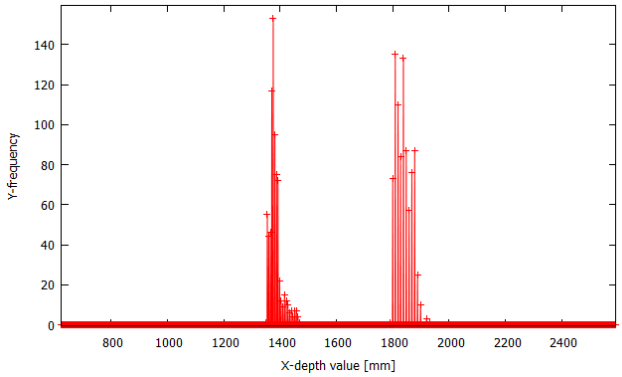


Fig. 4. Frequency of depth pixel in extracted hand region (Fig. 3(b))

C. Feature Extraction

The features for gestures recognition are divided into two parts: hand shape and arm movement. Conventionally, the position and movement velocity of both hands have been used as arm movement features. However, these features are less robust for environmental changing such as individual differences in body size, camera position, and so on. On the other hand, the ratio or contour of hand has been used as hand shape features. However, it is not sufficient to correctly represent the hand shape for hand posture recognition.

In this paper, we propose a novel and robust arm movement and hand shape features based on depth data.

1) Arm movement features

Angles between selected six joints (hand, elbow and shoulder) are defined as arm movement features. Fig. 5 shows example of two angles between elbow and hand. In this case, elbow is defined as origin of feature space. Here, arm movement features $\theta_1^{elbow-hand}$ and $\theta_2^{elbow-hand}$ are defined by:

$$\theta_1^{elbow-hand} = \tan^{-1}\left(\frac{y_3 - y_2}{x_3 - x_2}\right) \tag{3}$$

$$\theta_2^{elbow-hand} = \tan^{-1}\left(\frac{z_3 - z_2}{x_3 - x_2}\right) \tag{4}$$

In the same way, angles between elbow and shoulder are defined by:

$$\theta_3^{shoulder-elbow} = \tan^{-1}\left(\frac{y_2 - y_1}{x_2 - x_1}\right) \tag{5}$$

$$\theta_4^{shoulder-elbow} = \tan^{-1}\left(\frac{z_2 - z_1}{x_2 - x_1}\right) \tag{6}$$

It is thought conventionally that the movement velocity of hand is important to recognize hand gestures. Thus, angular velocities are employed for arm movement features.

$$\theta_i^{velocity} = \theta_i^t - \theta_i^{t-1} \qquad (i = 1,...,4) \tag{7}$$

where t is frame number. Finally, we obtain 16 dimensional feature vectors in both arms.

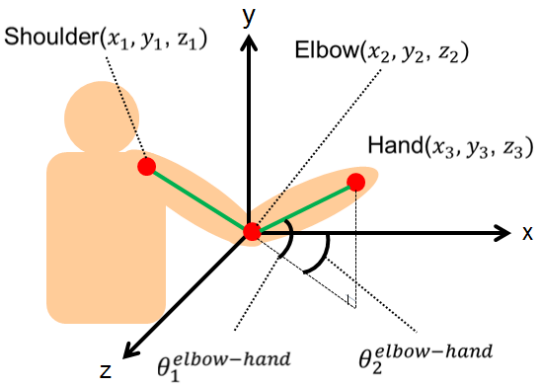


Fig. 5. The detail of proposed arm angle features

2) Hand shape features

Generally, a distance between sensor and hand causes the scale transform of captured hand shape. In addition, a movement of wrist joint causes the rotation transform of captured hand shape. Therefore, a scale and rotation invariant feature is required for robust hand gesture recognition.

In this study, we propose a novel feature extraction method by utilizing a concept of Speed Up Robust Feature (SURF) [15] which is known as a robust local feature descriptor in various environment. The process of SURF is categorized into two major groups: keypoint detection and feature description. Keypoint is defined as maxima and minima of the result of determinant of Hessian blob detector applied in scale-space. Our keypoint detection is not required for point of scale-space extrema in image because the cropped image is not large. We use the coordinates of hand joint as a keypoint, without keypoint detection.

On the other hand, in SRUF algorithm, the calculation of orientation is preformed in order to obtain invariant feature of rotation in feature description. However, we use an angle between hand and elbow as orientation, without conventional orientation estimation. The detail of proposed orientation is

shown in Fig. 6. After that, the cropped image is rotated on the basis of estimated orientation, which is shown in Fig. 7. Therefore, we use only an idea of feature description in SURF.

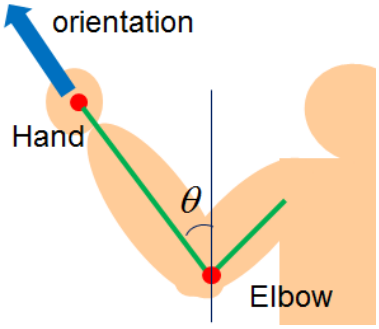
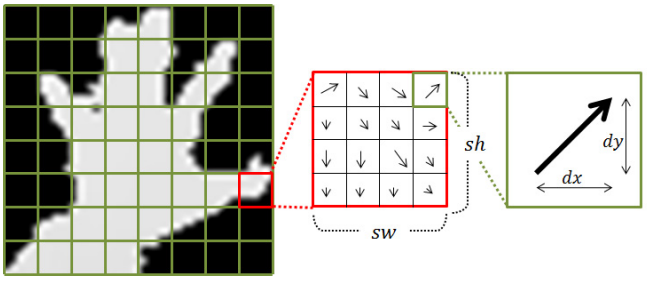


Fig. 6. The detail of proposed orientation estimation



(a) Cropped hand region (b) Segmented block (c) Pixel

Fig. 7. Description of proposed hand shape feature

The hand region is normalized by rotating and fitting hand pixels, which is shown in Fig 7. And then, the normalized hand region is divided into 8×8 blocks. The proposed gradient feature description is defined by:

$$\frac{1}{sh \times sw} \sum_{(x,y)} dx(x,y) \tag{8}$$

$$\frac{1}{sh \times sw} \sum_{(x,y)} dy(x,y) \tag{9}$$

$$\frac{1}{sh \times sw} \sum_{(x,y)} |dx(x,y)| \tag{10}$$

$$\frac{1}{sh \times sw} \sum_{(x,y)} |dy(x,y)| \tag{11}$$

where *sh* and *sw* are height and width of block, respectively. In order to obtain scale-invariant feature, all features are normalized by block size. Therefore, 256 (8 blocks × 8blocks × 4features) dimensional feature vectors in each hand are obtained. Then, these hand shape features are compressed to 8 dimensional vectors by principal component analysis. As a result of feature extraction, we obtain 16 dimensional feature vectors in both hands for hand shape features.

D. Classification

We use HMM with mixture of gaussians outputs as a classifier in order to recognize hand gesture, which is a statistical model assumed to be a Markov process with unobserved (hidden) states. Fig. 8 shows a left-right Markov model which consists of s_i states, a_i transition probabilities and b_i output probabilities. To classify a sequence into one of k classes, we train up parameter of k HMMs. A sequence of input gestures is classified by computing output probabilities in each state.

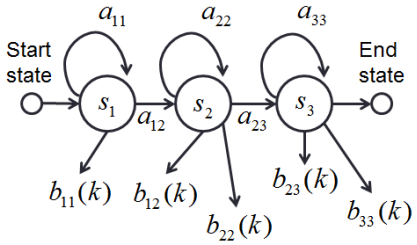


Fig. 8. Hidden markov model

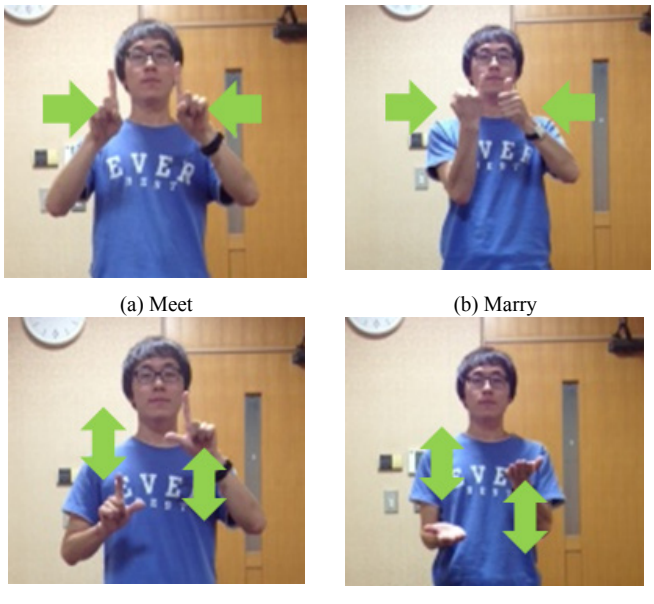
IV. Experiment and Discussion

In order to show the effectiveness of the proposed method, a performance for proposed hand gesture recognition method is evaluated by comparing with the conventional method using depth sensor [5].

In this experiment, we use 14 Japanese sign languages which are shown in Table III, as test data set. Example of this data set is shown in Fig. 9. To evaluate effectiveness of proposed hand shape features, sign languages which involve similar arm movement are employed. Thus, there is a possibility that Fig. 9(a) "meet" is recognized as Fig. 9(b) "marry" when the proposed system recognizes sign language "meet" using only arm movement features.

TABLE III: Sign Language for Evaluational Experiment

meet – marry	study – cheer up
size – which one - compare	
tanabata – shougatu	heavy - west
cute / come	place



(a) Meet (b) Marry

(c) Size (d) Compare

Fig. 9. Example of similar sign language

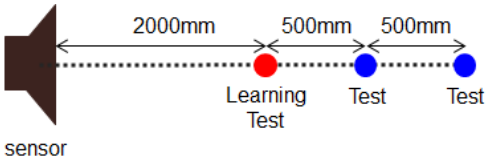


Fig. 10. Detail of capturing learning and test data set

Test data set is captured from four subjects. As a restriction, a distance between sensor and subject's hand is set to about 2000mm, 2500mm and 3000mm. For each subject, 3 times per each sign language is captured. The cross-validation method is used for this evaluational experiment. In the case 1, as a training and test dataset, 2000mm gesture data captured at red point is used. In the case 2, 2000mm gesture data captured at red point is used for training data. However, different distance data (2500mm and 3000mm) is used for test data, in order to evaluate robustness for environmental changing. The parameter of mixtures is set to 15 gaussians, and states of HMM is set to 13, experimentally. The number of dimensions in feature vector of conventional method is 58, and the number of dimensions in feature vector of proposed method is 36.

As a result, we show the recognition accuracies of proposed method and conventional method in Table IV. For recognition, the conventional method uses 2D hand features such as ratio and ellipse approximation of extracted hand shape. In addition, only 2D coordinates of both hands are used for arm movement feature. From table IV, although recognition accuracy of the conventional method in case 2 decreased from case 1, the accuracy of proposed of method is same in both cases. Therefore, we confirmed that effectiveness and robustness of the proposed method.

TABLE IV: EXAMPLE OF MISRECOGNITION

(%)	Case 1	Case 2
Proposed method	81.8	81.9
Conventional method [10]	77.6	68.4

V. CONCLUSIONS

In this paper, we proposed the novel hand gesture recognition system using depth data which is robust for environmental changing. Our approach involves an extraction of hand shape features based on gradient value instead of conventional 2D shape features, and arm movement features based on angles between each joints. In order to show the effectiveness of the proposed method, a performance for proposed hand gesture recognition system was evaluated by using Japanese sign language. As a result, we confirmed the effectiveness of the proposed method by comparing with conventional method.

REFERENCES

[1] G. R. S. Murthy and R. S. Jadon, "A Review of Vision Based Hand Gestures Recognition," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 405-410, 2009

[2] M. Turk, "Gesture Recognition in Handbook of Virtual Environment Technology," *Lawrence Erlbaum Associates*, 2001.

[3] K. Hoshino and T. Tanimoto, "Realtime hand posture estimation with self-organization map for stable robot control," *IEICE Trans. Inf. and Syst.*, vol. E89-D, no. 6, pp. 1813-1819, 2006.

[4] L. Dipietro, A. M. Sabatini, and P. Dario, "Survey of Glove-Based Systems and their applications," *IEEE Trans. on systems, Man and Cybernetics*, vol. 38, no. 4, pp. 461-482, 2008.

[5] A. Sato, K. Shinoda and S. Furui, "Sign Language Recognition Using Time-of-Flight Camera (in japanese)," in *Proc. of Meeting on Image Recognition and Understanding*, IS3-44, pp. 1861-1868, 2010.

[6] L. Dipietro, A. M. Sabatini, and P. Dario, "A Survey of Glove-Based Systems and Their Applications," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 38, no. 4, pp. 461-482, 2008.

[7] K. Higashiyama, S. Ono, Y. Wang, and S. Nakayama, "Finger character recognition using 3-dimensional template matching (in Japanese)," *IEEJ Trans. EIS*, vol. 125, no. 9, pp. 1444-1454, 2005.

[8] H. Takimoto, S. Yoshimori, Y. Mitsukura, and M. Fukumi, "Hand Posture Recognition Robust for Posture Changing in Complex Background," *Journal of Signal Processing*, vol. 14, no. 6, pp. 483-490, 2010.

[9] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara, "Hand pose estimation using multi-viewpoint silhouette images," in *Proc. 2001 IEEE/RSJ Int. Conf. Intelligent Robots and Syst.*, pp. 1989-1996, 2001.

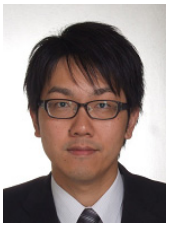
[10] H. Takimoto, T. Furukawa, M. Kishihara, and K. Okubo, "Robust Fingertip Tracking for Constructing an Intelligent Room", *Proc. of 21th IEEE International Symposium in RO-MAN2012*, pp. 759-763, 2012.

[11] Y. Wu and T. S. Huang, "Hand modeling analysis and recognition for vision-based human computer interaction," *IEEE Signal Processing Magazine, Special Issue on Immersive Interactive Technology*, vol. 18, no. 3, pp. 51-60, 2001.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images," in *Proc. of 24th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1297-1304, 2011.

[13] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: SpeedUp Robust Features," *Computer vision and Image Understanding(CVIU)*, vol. 110, no. 3, pp. 346-359, 2008.



Hironori Takimoto received his B.E., M.E., and Ph. D. degrees from the University of Tokushima in 2002, 2004, and 2007, respectively. He joined the Sasebo National College of Technology in 2005. Since 2009, he has been an Assistant Professor at Okayama Prefectural University. His research interests include image processing, and human sensing. He is member of IEEJ, IEICE, SICE, and IEEE.

Lee Jaemin received his B.E. and M.E. degrees from Okayama Prefectural University in 2011 and 2013. Since April 2013, he has joined SuperSoftware Co., Ltd. His research interests include image processing, and human sensing.

Akihiro Kanagawa received his B.E. degree in systems engineering from Kobe University in 1983, and Ph. D. degree in industrial engineering from the University of Osaka Prefecture in 1991. From 1993 to 2007, he was an Associate Professor at Okayama Prefectural University. Currently, he is a Professor in the Faculty of Computer Science & Systems Engineering, Okayama Prefectural University, Japan.