Appearance-Based Gaze Estimation for Digital Signage Considering Head Pose

Hiroki Yoshimura, Maiya Hori, Tadaaki Shimizu, and Yoshio Iwai

Abstract—**Digital signage often uses a large display with the camera placed on the outer frame and set with a wide angle to capture the face of the viewer. In some cases the viewer's iris is obstructed due to the gaze position of the viewer. In this study, we present an appearance-based gaze estimation method that can be used for digital signage even when the iris is not fully visible. The proposed approach uses the angle of the head and the image of the eye area as features for a neural network machine learning algorithm. Our subject experiments confirm that we achieve accurate focus-point gaze estimation.**

Index Terms—**Gaze estimation, digital signage, head pose, neural network.**

I. Introduction

Posters attached to physical structures (e.g., walls) have thus far been the main medium for indoor and outdoor advertising and information displays. These paper posters are costly to update and are not conducive to presenting information tailored to the location or time. The electronic billboard, referred to as digital signage (Fig. 1), is emerging to solve these problems. Digital signage is capable of real-time operation by using communication networks. Content changes are easily accomplished, and these updates can be delivered at any time. As a result, digital signage has been spreading rapidly.

Fig. 1. An example of digital signage.

Our objective is to provide a system that activates ad distribution in digital signage. Our design requirements include: tailoring content to the interests of the viewers, determining the optimal signage positioning, and using the viewer's gaze for reactive avatars and effects. These features lead to increased consumer attention and are thus highly beneficial for advertisers. In order to provide such a system, information about the gaze direction, head position, and behavior of the audience is vital. Our work focuses on gaze estimation, which is the fundamental element of these systems.

Existing gaze-estimation methods can be grouped into three categories. The first type is an infrared irradiation method [1]-[3]. The gaze direction is estimated by detecting the position of the pupil from the light reflected by the corneal surface in the eye (Purkinje image).This method boasts highly accurate results to approximately one degree. However, the placement of the infrared camera and light source is problematic since the eye must be illuminated by the infrared light. Moreover, this method requires high-resolution images of the eye region to detect the Purkinje image. Finally, device configuration is complicated as it entails multiple components interacting. The second gaze-estimation method type is the model-based approach [4]-[7]. This category can be further subdivided into two types. One approach estimates the gaze direction from the ellipse fitting parameters for the observed elliptical iris area. The other estimates gaze direction as a three-dimensional vector connecting the center of the iris and the eyeball center. of these approaches perform gaze estimation using a three-dimensional model of the eyeball. They require very accurate measurements taken from high-resolution images of the eye area. Therefore, the viewer placement is constrained to the vicinity of the camera. Furthermore, the computation of the eyeball rotation angle in the head entails an increased processing time. The third main gaze estimation type is the appearance-based approach. These gaze direction estimation techniques use machine learning algorithms such as nearest neighbors and neural networks to perform pattern recognition on images of the eye area [8]-[11]. With appearance-based methods, it is possible to estimate the gaze direction with a sufficient degree of accuracy even when using a low-resolution image since a three-dimensional model of the eyeball is not required. The viewer is not as constrained with respect to position in relation camera. However, appearance-based methods are affected by changes in the relative position of the face and the attitude of the head.

As described, existing methods have significant drawbacks for practical applications. Infrared irradiation methods are difficult to adapt f or digital signage gaze estimation since an infrared emitter is required. Both the model-based approach and infrared irradiation approach suffer from a reduction in estimation accuracy if the iris or reflected infrared light are not fully visible in the image. Since digital signage typically uses a large display, the camera must be installed outside the

Manuscript received October 7, 2014; revised August 2, 2015.

The authors are with the Department of Information and Electronics, Graduate school of Engineering, Tottori University, 4-101 Koyama-minami, Tottori, Japan (e-mail: yosimura@ike.tottori-u.ac.jp, tadaaki@ike.tottori-u.ac.jp, hori@ike.tottori-u.ac.jp, iwai@ike.tottori-u.ac.jp).

display area. As a result of the often acute angle between the camera and the viewer, the iris can be occluded according to the gaze direction.

In this paper, we propose a gaze estimation method for digital signage that overcomes many of the problems with existing systems. Since it is appearance-based, our system does not require any infrared equipment and is less perturbed by the relative positions of the viewer and camera. We estimate the head pose to address the effect of head-attitude changes on existing appearance-based methods. A feature of the proposed method consists in using the angle information of the face. We use a neural network machine learning algorithm with the image of the eye region and the angle of the face as input features. The angle of the face is obtained from the head pose estimation. We thus achieve a robust gaze estimation for digital signage.

II. Proposed Method

A. Overview of Proposed Method

The main objective of our work (see Fig. 2 for an overview) is to determine which of 12 subdivided areas of the digital signage the viewer is watching (see Fig. 1 for an example of digital signage, see Fig. 3 for the area divisions used). In this study, the granularity of the divided areas was considered sufficient for digital signage applications. We determine head pose, as well as the feature points and angle of the face from the camera image. The area of interest (the eye area) is cropped based on the feature points corresponding to the eyes. The angle of the face and the pixel intensities of the eye area are the input features to the neural networks.

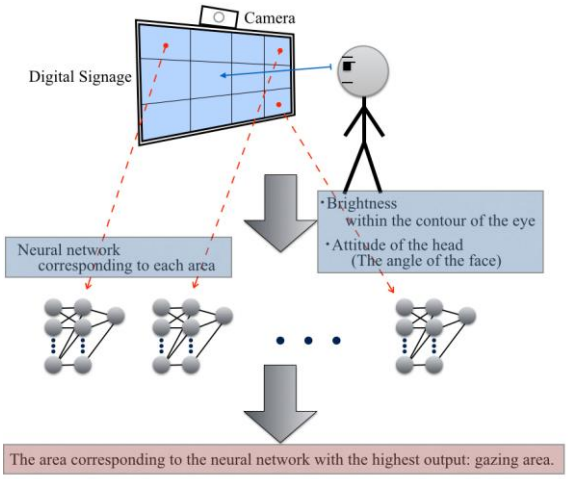


Fig. 2. Overview of proposed method.

B. Head Pose Estimation Using Faceapi

In this work, we use Seeing Machines' (an Australian company) face API face tracking tool for head pose estimation. The face API performs face tracking, feature point selection, and face angle detection. The face angle is obtained from the head pose estimation.

The face angle is given by pitch, yaw, and roll as shown in Fig. 4(a). Pitch is the angle of the x-axis, yaw is the angle of the y-axis, and roll is the angle of the z-axis (see Fig. 4(b)).The angle ranges from face API are−30 to 60 degrees for pitch, −90 to 90 degrees for yaw, and−90 to90 degrees for roll.

C. Extraction of the Eye Area

We determine the area of interest from the eye feature points obtained from face API. Fig. 5(a) shows the face feature points which are be obtained from face API. The eight feature points of the eye contour are then extracted (Fig. 5(b)).

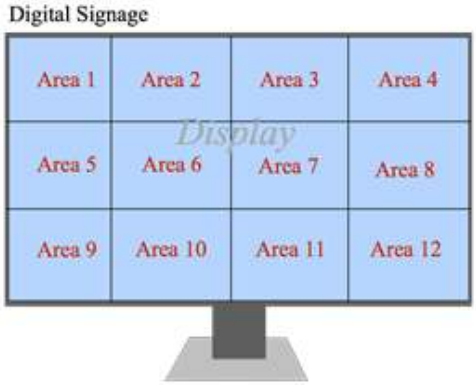
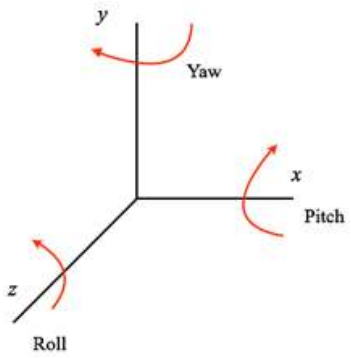
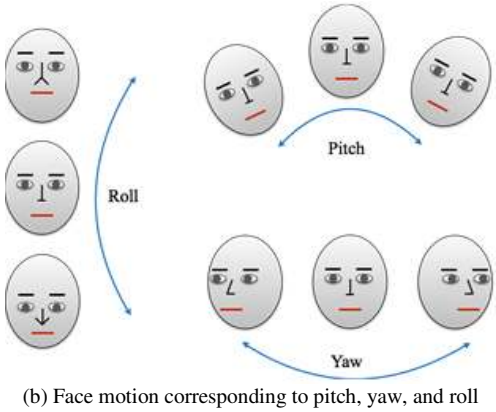


Fig. 3. Digital signage area divisions used.

Fig. 6 shows the normalization process to prepare the input for the neural networks. Straight lines connect the eye feature points of each eye, forming the approximation of the eye contour. Next, we determine the pixel intensities within the resulting rhombus-shaped contour of the eye. We then surround each eye by a70 × 22 pixel rectangle (including the feature points labeled in green in Fig. 6). These dimensions were determined based on the maximum face image width and height. The lower edge of the rectangle passing through the feature point on the lower eyelid is parallel to the line connecting the points on the outer and inner corners of each eye. The area outside of the eye within the rectangles (colored blue in Fig. 6) is assigned a zero intensity value.

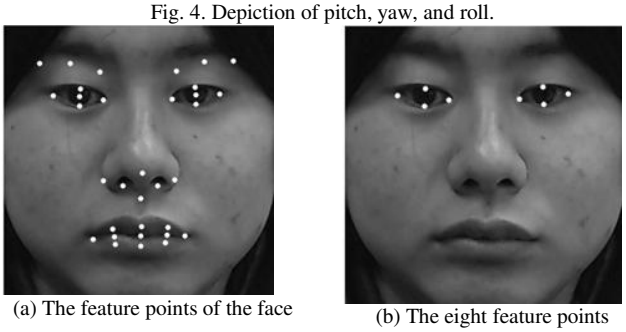


(a) Pitch, yaw, and roll axes



(b) Face motion corresponding to pitch, yaw, and roll

Fig. 4. Depiction of pitch, yaw, and roll.



(a) The feature points of the face

(b) The eight feature points corresponding to the eyes

Fig. 5. Feature points obtained from face API.

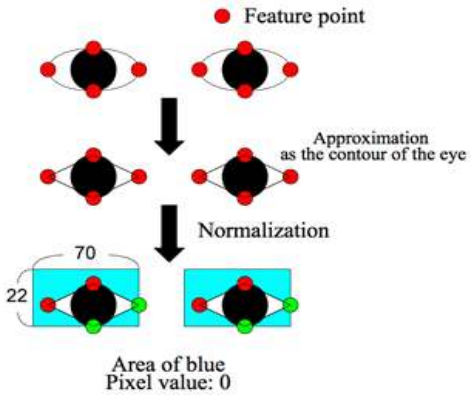


Fig. 6. Eye-area normalization steps.

D. Proportional Conversion to Prepare Inputs for Neural Networks

The input features to the neural networks are the contour of the eye (Fig. 6) and the face angle (pitch, yaw, and roll). The angle range differs between pitch (60 to −30 degrees) and yaw and roll (−90 to 90 degrees). Furthermore, the pixel intensity values in the eye contour range between 0and 255. The difference in ranges affects the influence of each feature on the neural network. Therefore, we perform a proportional conversion to equalize the influence between features. We set the maximum value to 0.9 and the minimum value to 0.1 for pitch, yaw, and roll in the training data. The maximum and minimum values are not 1 and 0 respectively since there is a possibility that the values of the minimum and maximum in the training data are not equivalent to those in the test data. In this case, the test-set face angle could exceed 1 or be below 0.Proportional conversion of pixel intensities within the eye contour is performed separately for each eye in each frame. The minimum pixel intensity is set to 0 and the maximum is 1.The values of the pixels outside of the eye contour are always 0, and these pixels are disregarded for proportional conversion.

III. Experiments

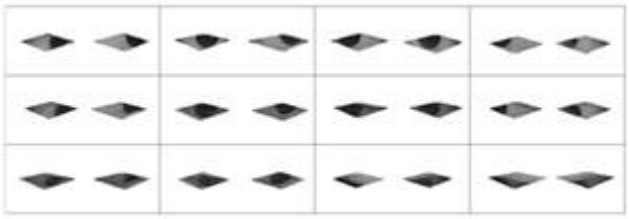
A. Experimental Method

This section describes how we obtain the face images we use throughout our experiments. We prepare digital signage images divided into 12 areas (three vertical ×four horizontal, see Fig. 3) with a red dot in the center of one area. The subject focuses on the red dot for a set period of time and the system captures an image of his/her face. We collect two categories of subject data:

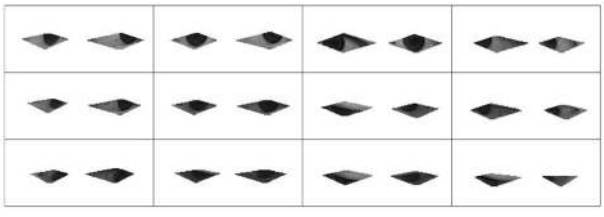
1) The subject looks at the red dot without moving the face (Fig. 7 (a)).
2) The subject looks at the red dot with natural unrestricted face movement (Fig. 7 (b)).

(a) No face movement

(b) Unrestricted natural face movement



(c) Eyes regions for (a)



(d) Eyes regions for (b)

Fig. 7. Examples of the face images.

B. Experimental Setup

We use the following equipment for our experiments:

- Digital signage:

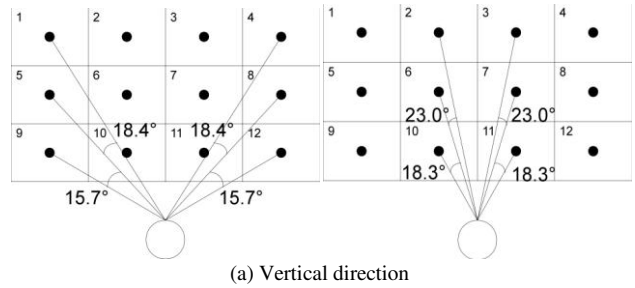
We use a 60-inch horizontally oriented information display from Sharp Corporation (PN-A601). The dimensions of the display are 133.4 × 75.3cm. The digital signage is mounted such that the lower edge is 65cm above the floor.

- RGB Camera:

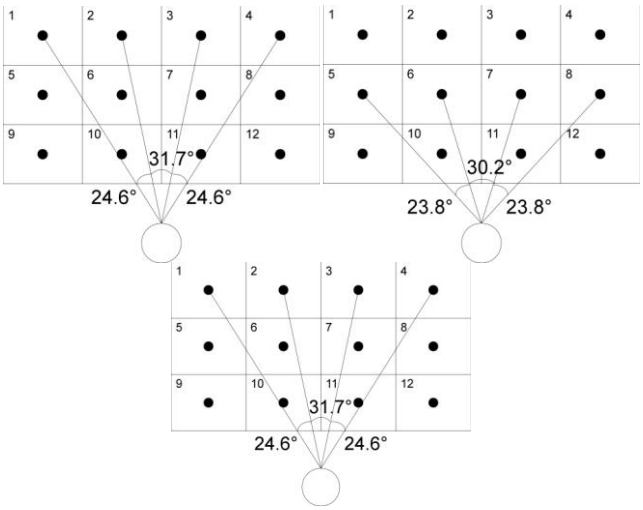
We use a Logicool HD Pro Webcam C920 from Logitech International S.A. with are solution of 1920 × 1080 pixels. This camera is placed at the center of the top of the digital signage with a 25 degree angle looking downward.

Fig. 8. Experimental setup.

The viewer is 60 cm from the digital signage. This is a reasonable generalization for the typical distance in real-world scenarios. The height of the eye is 20cm below the top of the digital signage. The subject is not physically restrained but is instructed to consciously limit face movement as much possible for the constrained experiment (Fig. 7(a)). The viewing angle in the vertical direction between the area of the display and the subject is shown in Fig. 9(a), the average is 18.8°. The viewing angle in the horizontal direction between the area of the display and the subject is shown in Fig. 9(b), the average is 25.1°.



(a) Vertical direction



(b) Horizontal direction

Fig. 9. The viewing angles for the display areas.

C. Gaze Direction Estimation Method

We use 12 neural networks, one corresponding to each area of the digital signage. Each frame of the data is one input to the neural network. The area corresponding to the neural network with the highest output is chosen as the gaze focus area. We use 10 cross-validations to evaluate our results. In total, our data set consists of 1440 frames (120 frames for each of the 12 areas).We use1296 frames for training (108 frames for each area) and 144 frames for testing (12 frames for each area).

D. Estimation Results

We use neural networks with three layers. The input layer consists of 3083 units (the number of pixels in the eye area, 22 × 70 pixels for each eye: 3080; pitch, yaw, and roll: 3; total: 3083).The output layer consists of one unit that is used to determine whether or not this neural network corresponds to the gazing area. The hidden layer consists of five units. A squared error of $E \leq 0.001$ is used to signal learning completion. We use a learning coefficient of $\eta = 0.1$. The learning coefficient η and the number of units in the hidden layer are determined empirically.

Table I(A) shows the success rate for the scenarios where the viewer moves only the eyes and constrains the rest of the face and head. The average accuracy rate for this case is 93.8%. Table I(B) shows the success rate for the scenarios where the viewer is allowed natural face and head movement in addition to the eye movements. The average accuracy rate for this case is 96.8%.We obtain a 3% higher accuracy rate for the unconstrained case than the constrained experiment.

TABLE I: EXPERIMENTAL ACCURACY RATES

(A) CONSTRAINED SCENARIO: EYE MOVEMENT ONLY

Correct \ Estimation	1	2	3	4	5	6	7	8	9	10	11	12
1	95.8	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	3.3	0.0	0.0
2	0.0	97.5	0.0	0.0	0.0	1.6	0.0	0.0	0.8	0.0	0.0	0.0
3	0.0	0.0	98.3	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.8
4	0.0	0.0	0.0	97.5	0.0	0.0	0.0	0.0	0.8	1.6	0.0	0.0
5	0.0	0.0	0.0	0.0	92.5	0.0	0.0	0.0	0.0	5.0	0.8	1.6
6	0.0	0.0	0.0	0.0	0.0	90.8	0.0	0.0	1.6	6.6	0.0	0.8
7	0.0	0.0	0.0	0.0	0.0	0.0	65.0	24.1	0.0	1.6	5.0	4.1
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.1	0.0	0.8	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	89.1	5.8	0.8	4.1
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

(B) UNCONSTRAINED SCENARIO: EYE AND NATURAL FACE MOVEMENT

Correct \ Estimation	1	2	3	4	5	6	7	8	9	10	11	12
1	91.6	3.3	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.8	0.8	0.0
2	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0
3	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	96.6	0.0	0.0	0.0	1.6	0.0	0.0	1.6	0.0
5	0.0	0.0	0.0	0.0	97.5	0.0	0.0	0.0	1.6	0.8	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	90.0	0.0	0.0	0.0	8.3	1.6	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	88.3	1.6	0.0	0.0	10.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.1	0.8	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

IV. DISCUSSION

Overall, our experimental results show a high success rate for gaze direction estimation. To understand where the

failures occurred, we analyzed the results frame by frame. The main reason for failure is attributed to blinking of the eyes (blinking caused 49.2% of the failures in the constrained case and 93.3% of the failures in the unconstrained case).When the gazing area was determined incorrectly, it was often the area below the correct one that was chosen (see Table I(A) and Table I(B)). This is because the contour of the eye shrinks during a blink and the image of the eye areas becomes extremely similar for two vertically adjacent areas.

We verified the effect of blinking by conducting an experiment after thinning the frames where the blinks occur and we saw increased estimation accuracy. However, in real-world applications, viewers are expected to blink naturally. Still, even with the presence of blinking, our proposed method provides sufficiently accurate results.

In failures not attributed to blinking, the neighboring areas were often chosen instead of the correct ones. This occurred more frequently in the constrained case. In the unconstrained case, it appears that the additional input of the face angle aided in gaze estimation. This is promising since viewers will inevitably use natural face movement in practical applications.

V. Conclusion

We present a system which successfully performs gaze estimation for digital signage divided into 12 areas. We obtain a high accuracy rate in both constrained and unconstrained cases and most of the failures were attributed to the viewer blinking. In future work, we aim to improve the accuracy rate by using a neural network to detect blinking. In this paper, we did not consider camera calibration and we plan to apply the proposed method without doing so to simplify real-world set-up.

References

[1] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *Proc. The 2006 Symposium on Eye Tracking Research & Application*s, 2006, pp. 87-94.

[2] T. Nagamatsu, J. Kamahara, and N. Tanaka, "Calibration-free gaze tracking using a binocular 3D eye model," in *Proc. CHI EA '09 CHI '09 Extended Abstracts on Human Factors in Computing Systems*, 2009, pp. 3613-3618.

[3] A. Villanueva and R. Cabeza, "A novel gaze estimation system with one calibration point," *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 38, no. 4, pp. 1123-1138, 2008.

[4] J. Wanga, E. Sungb, and R. Venkateswarlu, "Estimating the eye gaze from one eye," *Computer Vision and Image Understanding*, vol. 98, issue 1, pp. 83-103, 2005.

[5] Y. Kitagawa, H. Wu, T. Wada, and T. Kato, "On eye-model personalization for automatic visual line estimation," in *Proc. PRMU2007*, 2007, vol. 106, no. 469, pp. 55-60.

[6] Z. Wen, T. N. Zhang, and S. J. Chang, "Eye gaze estimation from the elliptical features of one iris," *Optical Engineering*, vol. 50, no. 4, 2011.

[7] K. Arai and R. Mardiyanto, "Camera mouse and keyboard for handicap person with trouble shooting capability, recovery and complete mouse events," *International Journal of Human Computer Interaction*, vol. 1, no. 3, pp. 46-56, 2010.

[8] C. Morimoto, A. Amir, and M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," in *Proc. Int', l Conf. Pattern Recognition*, 2002.

[9] Y. Ono, T. Okabe, and Y. Sato, "Gaze estimation from low resolution images," *Advances in Image and Video Technology Lecture Notes in Computer Science*, vol. 4319, pp. 178-188, 2006.

[10] J. Orozco, O. Rudovic, J. Gonzalez, and M. Pantic, "Hierarchical on-line appearance-based tracking for 3d head pose, eyebrows, lips, eyelid and irises," *Image and Vision Computing*, vol. 31, pp. 322-340, 2013.

[11] Y. Sugano, Y. Matsushita, and Y. Sato, "Unconstrained gaze estimation with learning from mouse operations," in *Proc. MIRU2009*, 2009, pp. 266-273.

Hiroki Yoshimura graduated from Tottori University in 1993 and completed the M.S. and doctoral programs in 1995 and 1998, respectively, also at Tottori University. He is currently an assistant professor at the Graduate School of Engineering at Tottori University. His main research interest is in machine learning.

Maiya Hori received his B.E. degree in systems science from Osaka University in 2005. He received his M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology in 2007 and 2011, respectively. He is currently a research fellow at Tottori University.

Tadaaki Shimizu received a Ph.D. degree in engineering from Osaka University in 2002. In 1987, he joined the Department of social systems engineering, Tottori University. He is currently an associate professor at the Graduate School of Engineering at Tottori University.

Iwai Yoshio graduated from Osaka University in 1992 and completed M.S. and doctoral programs in 1994 and 1997, respectively. He was then appointed as a research associate at the university, and he is subsequently becoming an associate professor. From 2004 to 2005, he was a visiting researcher at Cambridge University. He is currently a professor at the Graduate School of Engineering, Tottori University. He is engaged in studies relating to computer vision. He is a member of IEEE, the Information Processing Society, and the Japanese Robotics Society. He holds a D.Eng. degree.