https://doi.org/10.2352/EI.2022.34.17.3DIA-235

Hand Authentication from RGB-D Video Based on Deep Neural Network

Ryogo Miyazaki, Department of Imaging Science, Chiba University, Chiba, Japan
Kazuya Sasaki, Magik Eye Ink, Tokyo, Japan
Norimichi Tsumura, Graduate School of Engineering, Chiba University, Chiba, Japan
Keita Hirai, Graduate School of Engineering, Chiba University, Chiba, Japan

Abstract

In recent years, behavioral biometrics authentication, which uses the habit of behavioral characteristics for personal authentication, has attracted attention as an authentication method with higher security since behavioral biometrics cannot mimic as fingerprint and face authentications. As the behavioral biometrics, many researches were performed on voiceprints. However, there are few authentication technologies that utilize the habits of hand and finger movements during hand gestures. Only either color images or depth images are used for hand gesture authentication in the conventional methods.

In the research, therefore, we propose to find individual habits from RGB-D images of finger movements and create a personal authentication system. 3D CNN, which is a deep learning-based network, is used to extract individual habits. An F-measure of 0.97 is achieved when rock-paper-scissors are used as the authentication operation. An F-measure of 0.97 is achieved when the disinfection operation is used. These results show the effectiveness of using RGB-D video for personal authentication.

1. Introduction

In recent years, biometric authentication using fingerprints, irises, faces, etc., has been used instead of passwords and PINs as a personal authentication method when logging in to a PC or unlocking a smartphone. Such personal authentication using the physical and behavioral characteristics of a person is called biometrics authentication. It is said to be more secure than a password because it is less likely to be lost or stolen. Biometrics authentication is beginning to be used not only for logging in to PC and smartphones but also for immigration control at airports and identity verification when paying with credit cards.

Biometrics authentication is classified into a static method using physical characteristics and a dynamic method using behavioral characteristics. Fingerprints [1], faces [2], irises [3], veins [4], etc., are used for authentication based on physical characteristics. Since these technologies use a part of one's body for authentication, they are convenient and have high authentication accuracy. However, once a feature is stolen, the physical quality cannot be changed, so there is a disadvantage that it cannot be used for authentication again. Also, fingerprints and faces may change over time, and even the person himself may be rejected. Furthermore, face recognition can pass the authentication even if it is another person by using the person's face photo and collecting and duplicating the fingerprint.

Authentication using behavioral characteristics can solve the disadvantages of authentication using the above physical characteristics because it is difficult to reproduce the person's behavioral habits even if the authentication operation is leaked. Therefore, it is attracting attention as a more secure authentication method. Gaits [5], voiceprints [6], signatures [7], etc., are used for authentication. Besides, the Japan Automatic Identification Systems Association [8] has found that the fingers' movements during hand gestures include individual characteristics. Gait authentication uses the whole body, so it is difficult to change the movement, but hand gestures are very convenient because it is easy to change the movement pattern. There are previous studies such as [9, 10] for authentication using hand gestures, but the number of examples is small compared to authentication based on other behavioral characteristics [11]. Only color images are used for hand gesture authentication in reference [9], and only depth images are used in the method in reference [10]. Therefore, it is considered possible to improve the authentication accuracy by performing authentication by combining a color image and a depth image that has three types of information, vertical, horizontal, and depth, and which easily produces motion information. In addition, the demand for non-contact technology is increasing due to the epidemic of COVID-19, and it is possible to perform authentication using hand gestures without contact.

Based on the above, the purpose of this paper is to obtain individual habits from RGB-D images of finger movements during hand gestures and perform personal authentication as a non-contact and highly-secure authentication method. As hand gestures used for authentication, (1) rock-paper-scissors movement and (2) movement of holding his hand over the disinfectant are examined, and the habit of finger movement is sought, and personal identification is performed. A 3D CNN called ECO [12] by Mohammadreza Zolfaghari et al. Is used to extract individual habits. In addition, we will verify the accuracy of personal authentication through experiments.

2. Related Works

2.1 Personal Authentication Using Hand Gestures

Here, we introduce related research on biometrics authentication using hand gestures. For authentication by hand gesture, a method using an accelerometer [13, 14] and a method using a color image or a depth image have been proposed. Here, the latter related method is introduced.

Azuma et al. [9] use color images of hand gestures and perform personal authentication using two features, one that focuses on the movement of the fingertips and the other that uses HLAC. As a result, they have achieved a true acceptance rate of 80% and a true rejection rate of 99.5%. Wu et al. [10] have developed two-stream-convolutional neural networks for personal authentication from optical flows that represent the depth images of a series of gestures and the movement of objects between frames as vectors. Depth images and optical flow are used for personal authentication. By inputting the depth image into the spatial convolutional network and the optical flow into the time-axis convolutional network and combining the outputs of the two networks, it is possible to identify the person even in the gesture image that is not in the training image.

2.2 ECO

Here, we explain the ECO used to obtain the features of finger movements. ECO is an abbreviation for Efficient Convolutional Network for Online Video Understanding and is a deep learning model for classifying behaviors in videos.

The ECO consists of the 2D Net module Inception-v2 [15] and the 3D Net module 3D ResNet [16]. The input is a 16-frame video with a length and width of 224 pixels, which is first processed by Inception-v2 to extract features. Next, the extracted 16 frames of features are combined and input to 3D ResNet. After that, the behavior class is classified in the fully connected layer from the output of 3D ResNet. However, in this paper, we do not identify the behavior class after extracting the features but use ECO to obtain the features from the moving images. In this paper, we do not perform training or fine-tuning with our own data but use the pre-trained model [17] published on the ECO developer's GitHub.

3. Data Measurement System and Preprocessing

3.1 Shooting Environment

Figure 1 shows the shooting environment for this experiment. The 3D sensor used for shooting is FX-1 [18], and the color camera is eMeet Nova. The 3D sensor and color camera were fixed on a desktop stand. The measurement data is shown in Table 1. The subjects were 9 males in their 20s, and the subjects performed hand gestures with their palms facing up, and their movements were photographed at 30 fps for each sensor. The number of shootings is 24 times per person for rock-paper-scissors movements and 30 times per person for disinfection movements. Figures 2 and 3 show examples of data taken with a 3D sensor and a color camera for rock-paper-scissors and disinfection operations.

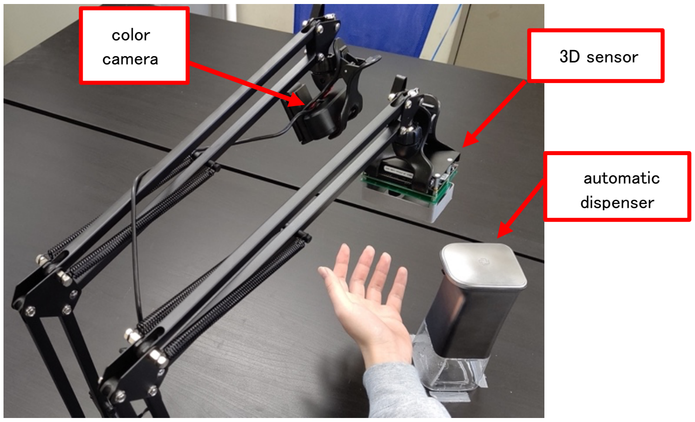


Figure 1. Shooting environment (left: color camera, right: 3D sensor)

Table 1 measurement data

authentication operation	subject	number of shots
Rock-paper-scissors	9 men in their 20s	24
disinfection	9 men in their 20s	30

3.2 Data Processing

First, both the shooting results of the color camera and the 3D sensor are trimmed for 128 frames from the detection of the hand to the end of the hand gesture. The shooting results with the 3D sensor are obtained as a point cloud, as shown in Fig. 2 and Fig. 4. Normalize this point cloud from -0.5 to 0.5. The shooting result of the color camera is converted into an image for each frame. Then resize the height and width to 224 pixels to fit the ECO input size.

4. Personal Authentication System Using 3D CNN

4.1 System Flowchart

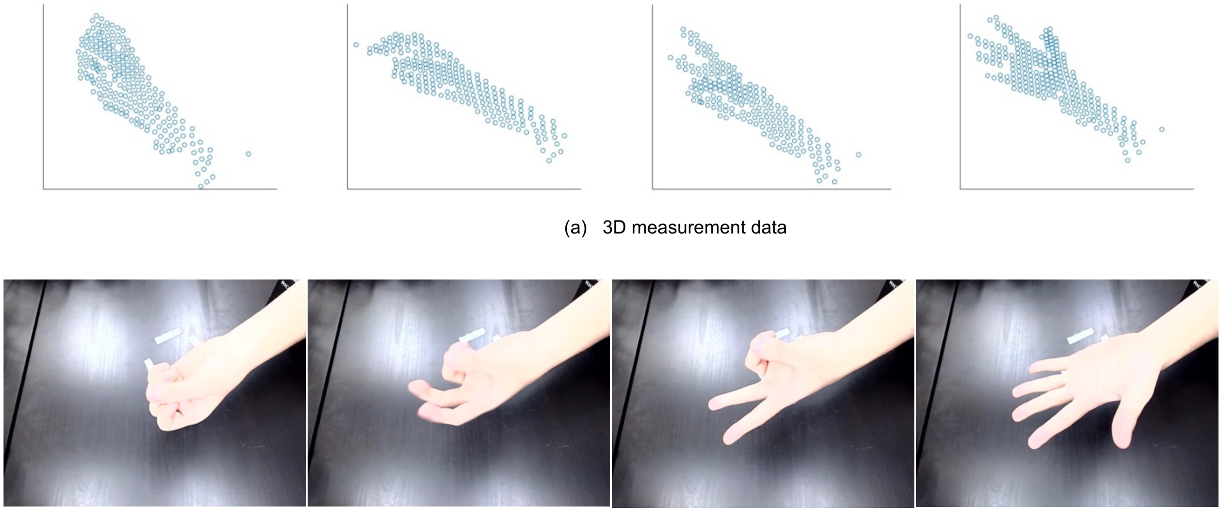
In this method, the feature amount of hand movement is obtained from the shooting result by 3D CNN. Figure 4 shows the flowchart of the proposed method. This method consists of two stages, learning, and authentication. First, during learning, the normalized 3D point cloud is converted into a depth image with a height and width of 224 pixels. The created depth image and color image are input to the 3D CNN for each frame, and the feature amount of the movement habit is calculated. In this paper, ECO [14] is used as a 3D CNN for finding features. A 512-dimensional feature is obtained as the output of ECO. Then, the features are input to the SVM, and a discriminative model for judging an individual from the habit of finger movement is created. At this time, the SVM inputs were verified using only depth images, color images only, and RGB-D images.

At the time of authentication, the normalized point cloud is changed to a depth image as in the case of learning. After that, it is input to the ECO for each frame, and the feature amount of the habit of hand gesture is calculated. After that, the feature amount is input to the SVM discrimination model created at the time of learning, and personal authentication is performed.

4.2 Network Details

ECO consists of a 2D network and a 3D network. The 2D network uses up to the inception-3c layer of the BN-Inception architecture [15]. This architecture has a pooling kernel with 2D filters and batch normalization. This architecture is said to be efficient. The input of the 2D network is a 16-frame image of size 224 x 224, and the output consists of 96 feature maps of size 28 x 28 with 16 frames.

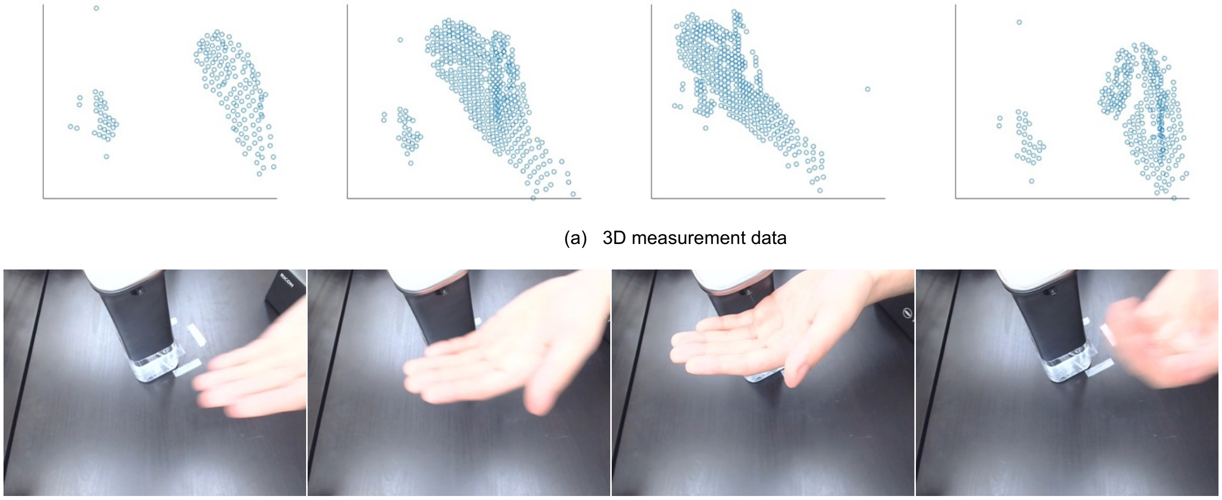
The 3D network uses 3D-ResNet18 [16]. 3D-ResNet18 is an efficient architecture used in many video classification tasks. The output is a feature size of 1x512.



(a) 3D measurement data

(b) RGB measurement data

Figure 2. Shooting data of rock-paper-scissors operation



(a) 3D measurement data

(b) RGB measurement data

Figure 3. Shooting data of disinfection operation

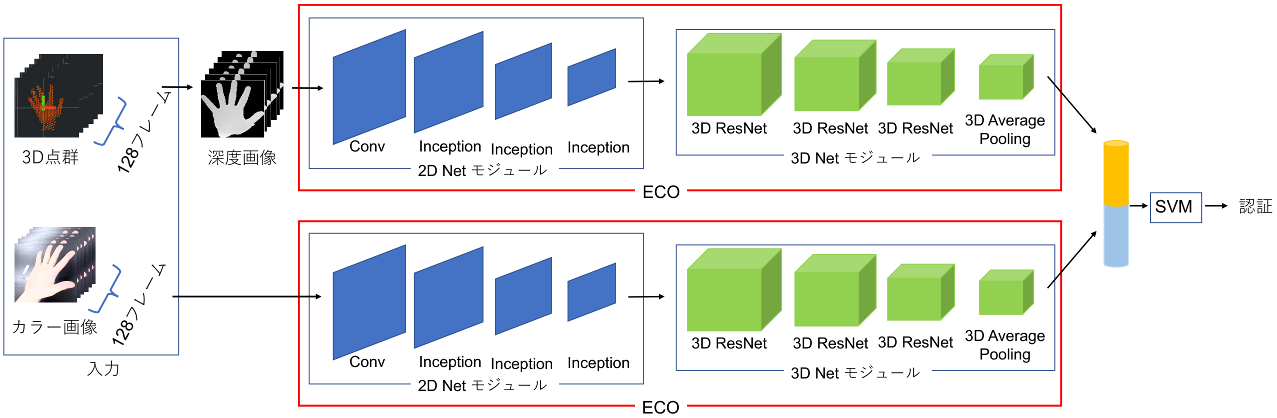


Figure 4. Flowchart of Proposal System

In this paper, we did not train ECO with our own data but used a pre-trained model. The pre-trained model is trained by the Momentum method [19] with an initial learning rate of 0.001, a batch size of 32, a weight decay of 0.0005, and a momentum of 0.9. The 2D net weights are initialized with the Kinetics pre-trained BN-Inception [16] provided in [20]. The weights of the 3D Net are initialized by the pre-trained model of the 3D ResNet18 provided in [21]. After that, it is trained with 10 epochs in the Kinetics dataset.

5. Experimental Results

Based on the measurement data shown in Table 1, the results of personal authentication of rock-paper-scissors operation and disinfection operation from depth images, color images, and RGB-D images by the leave-one-out method are shown in Table 2 and Table, respectively. Shown in 3.

The maximum accuracy of the precision rate is 0.98, and the maximum accuracy of the recall rate is 0.98. The accuracy of the previous study [9] using color images of hand gestures is that the acceptance rate of the person is 80%, and the rejection rate of others is 99.5%. In the accuracy evaluation index of this paper, the recall rate corresponds to the person acceptance rate, and the precision rate corresponds to the refusal rate of others. From this, the acceptance rate of the person in this method is 98%, and the rejection rate of others is 98%. The reproducibility exceeded the accuracy of the previous study, and the precision rate achieved the same accuracy as the previous study.

Looking at the average F-measure of personal authentication in rock-paper-scissors operation, it was 0.98 for depth images only, 0.94 for color images only, and 0.97 for RGB-D images. In the disinfection operation, it was 0.84 for the depth image only, 0.97 for the color image only, and 0.97 for the RGB-D image. It can be seen that the authentication with the depth image and the RGB-D image is more accurate than the authentication with the rock-paper-scissors operation only with the color image. It is considered that this is because the depth image can acquire the habit of hand and finger movement when the hand is changed from rock to scissors and scissors to paper in the 3D shape than in the color image. In addition, in Fig. 2 (a) and Fig. 3 (a), the shooting results with the color camera may not be able to perform stable image measurement due to the influence of lighting and external light, so the area and movement of the hand can be acquired appropriately. It may not be possible. Therefore, it is considered that the use of a 3D sensor can capture hand movements more stably in the shooting environment than with a color camera. On the other hand, in the case of disinfection operation, the F value of the color image and RGB-D image authentication is higher than that of the depth image alone. It is considered that this is because the disinfection movement only puts the hand in and out, and it is more difficult to distinguish the habit of finger movement such as bending and opening the finger than the rock-paper-scissors movement. However, the RGB-D image authentication is highly accurate in both the rock-paper-scissors operation and the disinfection operation. From the above, it is considered that stable authentication accuracy can be achieved by using RGB-D images for authentication, regardless of the shooting environment, even for hand gestures with few movement habits.

Table 2 Personal Authentication results of rock-paper-scissors operation

(a) depth video

Precision	Recall	F-measure	Accuracy
0.98	0.98	0.98	0.98

(b) RGB video

Precision	Recall	F-measure	Accuracy
0.95	0.94	0.94	0.94

(c) RGB-D video

Precision	Recall	F-measure	Accuracy
0.97	0.97	0.97	0.97

Table 3 Personal Authentication results of disinfection operation

(a) depth video

Precision	Recall	F-measure	Accuracy
0.84	0.84	0.84	0.84

(b) RGB video

Precision	Recall	F-measure	Accuracy
0.98	0.97	0.97	0.97

(c) RGB-D video

Precision	Recall	F-measure	Accuracy
0.97	0.97	0.97	0.97

6. Conclusion

In this paper, we used 3D CNN to find the movement habits of hand gestures for RGB-D image data and performed personal authentication. Authentication was performed using only depth images, only color images, and RGB-D images. As a result, it was confirmed that the depth image could be stably photographed with the hand movement regardless of the shooting environment, and the color image can be authenticated even with the movement that does not easily cause movement habits. Therefore, by using RGB-D images, stable authentication can be performed even for hand gestures that are not affected by the shooting environment and whose movement habits are unlikely to appear.

In this paper, 9 subjects were used for personal authentication. However, as an authentication system, since the number of subjects is small, it is necessary to confirm whether this method is effective even if the number of subjects is increased. In addition, although rock-paper-scissors and disinfection operations were used as personal authentication operations, it is necessary to consider whether other hand gestures also cause individual habits. It would be more convenient if authentication could be performed with shorter and simpler operations.

Furthermore, improving the authentication accuracy is a future issue. First, in order to improve accuracy, it is necessary to increase the shooting frame rate. In this paper, the authentication operation was shot at 30 fps, but by increasing the resolution on the time axis, it may be possible to shoot fine habits of hand and finger movements. Next, it is necessary to devise data preprocessing. In this paper, only data normalization is performed. Therefore, it is considered that the habit of hand gesture movement can be obtained by extracting only the hand region as preprocessing. Another example is to shoot hand gestures from two directions. In this paper, the image was taken from one direction, but by taking pictures from two directions, such as up and down, it is possible to reproduce the full 3D of the hand, and it is thought that it will be easier to find the habit of movement. In this paper, the deep learning used is a pre-trained model. Therefore, it is thought that the accuracy will be improved by learning the network with the data taken by oneself.

References

[1] W. Yang, S. Wang, J. Hu, G. Zheng, C. Valli, “Security and Accuracy of Fingerprint-Based Biometrics: A Review,” Symmetry, vol.11, no.2, pp.141, 2019.

[2] A. Rattani, R. Derakhshani, “A Survey Of mobile face biometrics,” Computers and Electrical Engineering, vol. 72, pp.39-52, 2018.

[3] K. W. Bowyer, K. Hollingsworth, P. J. Flynn, “Image understanding for iris biometrics: A survey,” Computer Vision and Image Understanding, vol. 110, no. 2, pp. 281-307, 2008.

[4] L. Wang, G. Leedham and D. Siu-Yeung Cho, ”Minutiae feature analysis for infrared hand vein pattern biometrics,” Pattern Recognit, vol. 41, no. 3, pp. 920-929, 2008.

[5] L. Wang, H. Ning, T. Tan and W. Hu, ”Fusion of static and dynamic body biometrics for gait recognition,” Proc. International Conference on Computer Vision, Nice, France, 2003.

[6] A. Boles and P. Rad, ”Voice biometrics: Deep learning-based voiceprint authentication system,” Proc. 12th Syst. Eng. Conf. (SoSE), 2017.

[7] M. A. Ferrer, M. Diaz-Cabrera and A. Morales, ”Static Signature Synthesis: A Neuromotor Inspired Approach for Biometrics,” Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 667-680, 2015.

[8] Japan Automatic Identification Systems Association, Yokuwakaru baiometorikusu no kiso, Ohmsya, 2005.

[9] A. Syota, ” Proposal and Experimental Evaluation of Gesture Authentication Using Motion Features,” Mie University, 2014

[10] J. Wu, P. Ishwar and J. Konrad, ”Two-stream CNNs for gesture based verification and identification: Learning user style,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 42-50, 2016.

[11] C. Liu, Y. Yang, X. Liu, L. Fang and W. Kang, “Dynamic-Hand-Gesture Authentication Dataset and Benchmark,” Proc. IEEE Transactions on Information Forensics and Security, vol. 16, 2021.

[12] M. Zolfaghari, K. Singh and T Brox, ”ECO: Efficient Convolutional Network for Online Video Understanding,” Proc. European Conference on Computer Vision (ECCV), 2018.

[13] X. Lu, Y. Fang, Q. Wu, J. Zhao, and W. Kang, ”A novel multiple distances based dynamic time warping method for online signature verification,” Proc. Chin. Conf. Biometric Recognit. Chan, Switzerland: Springer, pp. 645-652, 2018.

[14] D. Lu, D. Huang and A. Rai, ”FMHash: Deep hashing of In-Air-Handwriting for user identification,” Proc. ICC-IEEE Int. Conf. Commun. (ICC), pp. 1-7, 2019.

[15] S. Ioffe, C. Szegedy, ”Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” Proc. 32Nd International Conference on International Conference on Machine Learning, 2016.

[16] D. Tran, J. Ray, Z. Shou, S. Chang, M. Paluri, ”ConvNet Architecture Search for Spatiotemporal Feature Learning,” CoRR abs/1708.05038, 2017.

[17] ”GitHub - mzolfaghari/ECO-pytorch: PyTorch implementation for ”ECO: Efficient Convolutional Network for Online Video Understanding,” Proc. ECCV 2018”, https://github.com/mzolfaghari/ECO-pytorch, 18 January 2022.

[18] “Magik Eye Inc.," https://www.magik-eye.com, 18 January 2022.

[19] N. Qian, ”On the momentum term in gradient descent learning algorithms,” Proc. Neural networks : the official journal of the International Neural Network Society, 12(1), pp.145-151, 1999.

[20] L. Wang, Y. Xiong, Z. Wang, Y. Qian, D. Lin, X. Tang, L. Van Gool, ”Temporal Segment Networks: Towards Good Practices for Deep Action Recognition,” Proc. European Conference on Computer Vision (ECCV), 2016.

[21] L. Wang, W. Li, W. Li, L. Van Gool, ”Appearance-and-Relation Networks for Video Classification, ” CoRR abs/1711.09125.

Author Biography

Ryogo Miyazaki received a bachelor’s degree from Chiba University. Currently, He belongs to the Graduate School of Integrated Science and Engineering at Chiba University. The current research field is a material appearance and object texture reproduction using computers.

Kazuya Sasaki worked for a company as called Magik Eye Inc in Tokyo, Japan. He is a senior engineer in the fields of artificial intelligence with an expertise in computer vision, and robotics camera calibration.

Norimichi Tsumura is an associate professor in department of information and image science at Chiba University. He received a PhD from Osaka University and was also visiting associate professor at Dept. of electrical and computer Eng. in University of Rochester. His current research field is mainly color and appearance engineering, affective computing and deep learning.

Keita Hirai was an associate professor in department of information and image science at Chiba University. He was also a research fellow of Japan Society for the Promotion of Science from April 2009 to March 2010. It is so sad that he passed away suddenly in 1 May 2021. He was interested in the researches for visual information processing, color image processing, computer vision, and computer graphics.