https://doi.org/10.2352/ISSN.2470-1173.2018.09.IRIACV-237

An Image Processing Based Method for Chewing Detection Using Variable-intensity Template

Atsuto FUJIMOTO, Takaaki OHKAWAUCHI ,Junji YAMATO ‡and Jun OHYA

Department of Modern Mechanical Engineering, Waseda University, Tokyo, Japan

Abstract

To contribute to the actualization of the care worker assistance robot, this paper proposes a method for detecting whether the care receiver is chewing from the video sequence acquired by the camera that observes that receiver. The proposed method detects the receiver's face and areas for both cheeks and chin. After applying some normalization to the areas, chewing detection that uses a variable-intensity template is performed, where the template consists of shape models, interest points and intensity distribution model. A likelihood based on the variable-intensity template is computed so that the receiver is judged whether the receiver is chewing. Experiments using seven subjects are conducted. As a result, the accuracy of chewing detection by the proposed method is 83%, which is quite promising.

Introduction

The recent statics reported by the World Health Organization [1] says that the disabled population occupies about 15% of the world population and is increasing due to population aging and increase in chronic health condition. Actually, much more than this 15% people are suffering from serious functional disabilities [2], which could cause difficulties in their daily lives.

The above-mentioned situation is considered to get more serious from now on. This means that disabled people who need to be supported by care workers increase rapidly, which results in serious shortage of care workers. To solve this problem, the nursing robot, which can support disabled people on behalf of human care workers, is desired to be actualized.

Among care workers' many assistances, the meal assistance is one of the most important assistances. Care workers carefully observe whether the disabled persons (care receivers) eat meals safely and comfortably, because care receivers, in particular, aged people, could fail to swallow the food appropriately at high probabilities, which could cause the death in the worst case. For the safety and comfortableness for the care receiver, care workers should give the next food after the receiver's mouth gets empty: i.e. care workers should give the food to the care receiver using a spoon, see if the receiver finishes chewing the food, and then give the next food at appropriate timings to the receiver. Among the nursing robot's technologies needed for the meal assistance, this paper focuses on a method that can judge whether the receiver finishes chewing the food.

Conventional methods of chewing detection are categorized into two types: contact and non-contact types. The contact type lets the receiver wear devices equipped with a pressure sensor [3] or a sound sensor [4][5], but wearing these devices is burdensome for elderly people. The non-contact type analyses the video sequence acquired by the camera that observes the receiver. For example, Synno et al. [6] developed a chewing detection method that calculates changes in the distance between the upper lip and the lower jaw by utilizing image processing technologies. However, since the above-mentioned distance change tends to be smaller than the receiver's head and/or body movements, the accuracy for detecting the change is not high. In addition, in order to apply noise filtering of time series data to the obtained output result of mastication detection, it is necessary to know the chewing cycle in advance, which is a strong constraint in general.

This paper proposes a non-contact type chewing detection method that can solve the problems present in the conventional method [6]. In this paper, we apply variable-intensity template[7][8] modeling changes in luminance of the left and right cheeks and chin areas during chewing so that we realize a chewing detection method which is robust against human pose fluctuation and not dependent on individual's chewing cycle.

In the remainder of this paper, we first describe the outline of the proposed method in Section 3. Then, we introduce preprocessing of the proposed method in Section 4. Furthermore, we introduce the variable-intensity template, which is the essence of the proposed method in Section 5. Finally, we evaluate our work and conclude the paper in Section 6 and 7 respectively.

Algorithm

Figure 1 illustrates the proposed method. The proposed method consists of two parts: pre-processing and chewing detection. In the pre-processing part, first, we conduct face detection in each frame of the video sequence acquired by the camera that observes the care receiver's face and extract the areas of the left and right cheeks as well as chin regions. Second, in order to ensure robustness against changes in lighting conditions, we normalize the brightness of those three areas using the average and variance of the brightness of the whole face image. Furthermore, luminance histogram smoothing is performed in order to enhance luminance changes in the above-mentioned three areas during chewing.

In the chewing detection part, before the chewing detection, a variable-intensity template is created by the training phase. Here, the template consists of shape models of the areas for the left and right cheeks as well as chin regions when the receiver does not chew, interest points extracted from these areas, and luminance distribution models when not chewing. During the time the receiver is chewing, the chewing detection part detects the chewing by matching each frame of the input video sequence with the trained model.

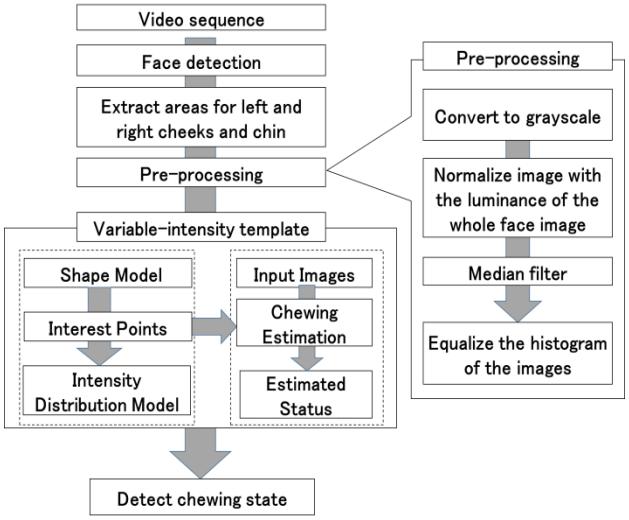


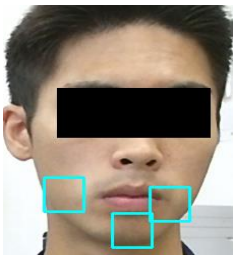
Fig.1 Overview of the proposed method

Pre-processing

First, the care receiver's face is detected from each frame of the input video sequence by using face detection library prepared by Kinect V2 , which is realized by learning a large amount of 3d skeleton model .Then we extract areas of the left and right cheeks as well as chin, by getting the coordinates of the left and right cheeks and chin by using the same face detection library and cutting 40×40 pixels images centered on the coordinates as shown in Fig. 2(a). Second, we convert these three areas to gray-scale images. Fig. 2 (b) shows the gray-scale image of the left cheek area. Third, to make the chewing detection process (detailed below) robust against changes in the camera position and lighting conditions, we normalize the luminance (gray-level) of the left and right cheeks as well as chin so that the luminance in the three areas match the average and standard deviation of the luminance of the whole face image. The normalized luminance I '(m, n) at pixel, whose horizontal and vertical coordinates are m and n, respectively, is computed by the following equation. Here, m and n respectively indicate the x and y coordinates of the image. In this process, we calculate independently for each of the three areas.

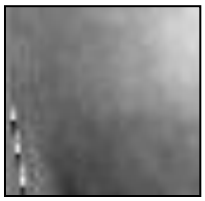
$$\mathbf{I'(m,n)} = \boldsymbol{\mu_0} + \frac{\boldsymbol{I(m,n)-\mu}}{\boldsymbol{\sigma}} * \boldsymbol{\sigma_0} \quad (1)$$

where I(m,n) is the luminance before normalization, $\boldsymbol{\mu}$ is the mean value of the three areas before normalization , $\boldsymbol{\sigma}$ is the standard deviation of the three areas before normalization, $\boldsymbol{\mu_0}$ is the mean value of whole face image, $\boldsymbol{\sigma_0}$ is the standard deviation of the whole face image. Fig.2 (c) shows the normalized right cheek area. Furthermore, a median filter is applied to remove the granular noise as shown in Fig.2(d). Finally, the contrast is enhanced by flattening the histogram of the area so that the change in the luminance during chewing is clarified as shown in Fig. 2 (e).

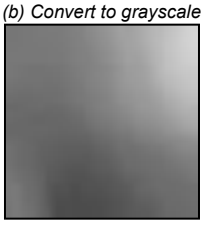


(a)Extract areas for left and right cheeks and chin

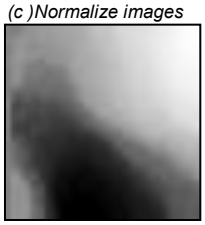
(b) Convert to grayscale



(c)Normalize images



(d) Median filter



(e) Equalize the histogram

Fig.2 Pre-processing

Variable-intensity template

the variable-intensity template M is composed of the following three elements,

$$\boldsymbol{M} = \{\boldsymbol{S}, \boldsymbol{P}, \boldsymbol{L}\} \quad (2)$$

where S is the shape models of the areas of the left and right cheeks as well as chin when not chewing, P is the set of the coordinates of points of interest, L is the luminance distribution model when not chewing. We detail these three elements in the following.

Shape model S

As described earlier, this paper builds the shape model S for the areas of the left and right cheeks as well as chin from a face that is not chewing and could change its orientation. We get the shape model S from a video sequence containing the "not chewing" face (for the training) by the procedure shown in Figure 3. First, a differential image of the frame (for the area) at time *t* and the frame at *t-1* is acquired for the areas of the left and right cheeks as well as chin, where the time for the first frame of the video sequence is 0. Then, we binarize the differential images using the threshold T, where 0 and 1 correspond to smaller and larger than T, respectively. Notice that the number of 1-value pixels of the binarized image during this chewing is significantly different from that in other conditions. If the number of 1-value pixels is smaller than the threshold value T1, that area is judged to be "not chewing"; otherwise "chewing". If the area at *t* is judged as "non-chewing", the area in an image format is included in the shape model. We hold model S in image format. To be robust against

changes in the orientation of the "non-chewing" face, the above-mentioned process is repeated till N data is included in the shape model S.

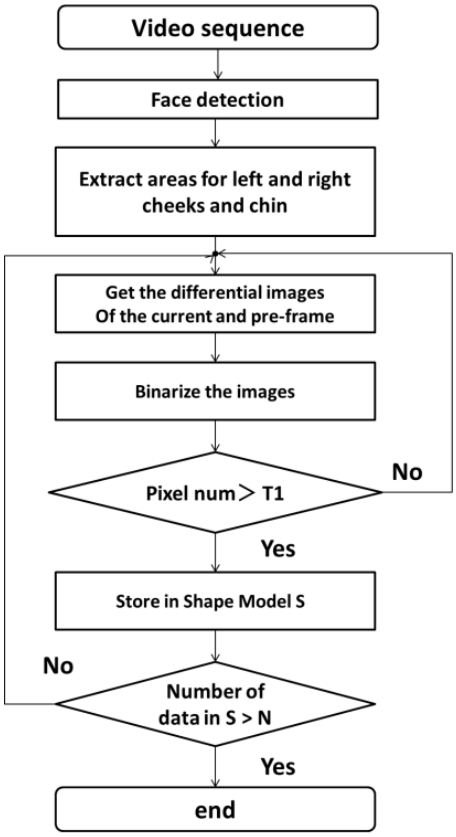


Fig.3 Flow chart of the Shape Model

Interest points P

Interest points are detected in the differential image of the frame at time t and t-1 for the right and left cheeks as well as chin. In case of "not chewing", almost all of the pixel values of the differential images are 0, while in case of "chewing", the pixel values of the differential images change intensively. First, in the differential image we detect candidate points whose pixel value is larger than the threshold value T2. Then, the initial points are obtained by selecting the candidate pixels with the first to k-th largest pixel values, where if a candidate pixel is not R pixels apart from one or more selected interest points, that candidate pixel is not selected as an interest point. The horizontal and vertical image coordinates of the k interest points are stored in P. Figure 4 shows the image of Interest Points.

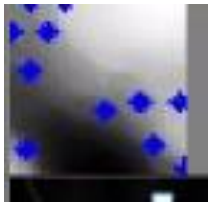


Fig.4 Interest Points

Intensity distribution model L

The chewing luminance distribution model is a model of how the luminance of each interest point changes during chewing. In this paper, assuming the independence of the luminance of each interest point, each luminance is represented by a normal distribution, and the luminance of each interest point during chewing changes according to that interest point's distribution. It is based on the idea that brightness is represented by a normal distribution because the brightness spreads due to multiple factors such as shape model error and imaging system noise. Owing to this chewing luminance distribution, this paper's chewing detection can be considered to be robustness against changes in the posture of the human body during chewing. Specifically, the chewing luminance distribution model is expressed as follows.

$$\boldsymbol{L} = \{\boldsymbol{P_1} \dots \boldsymbol{P_n}\}, \boldsymbol{P_k} = \boldsymbol{N}\left(\boldsymbol{\mu(e)}, \boldsymbol{\sigma^2(e)}\right), \boldsymbol{\sigma} = \boldsymbol{c} * \boldsymbol{\mu} \qquad (3)$$

where, $\boldsymbol{P_k}(\mathbf{k} = \mathbf{1} \dots \mathbf{n})$ are interest points described before. $\boldsymbol{N}\left(\boldsymbol{\mu(e)}, \boldsymbol{\sigma^2(e)}\right)$ is the normal distribution with mean μ and standard deviation σ . σ is defined as a constant multiple of μ , and e = {0},0 represents no chewing. Figure 5 shows the image of Intensity distribution Model.

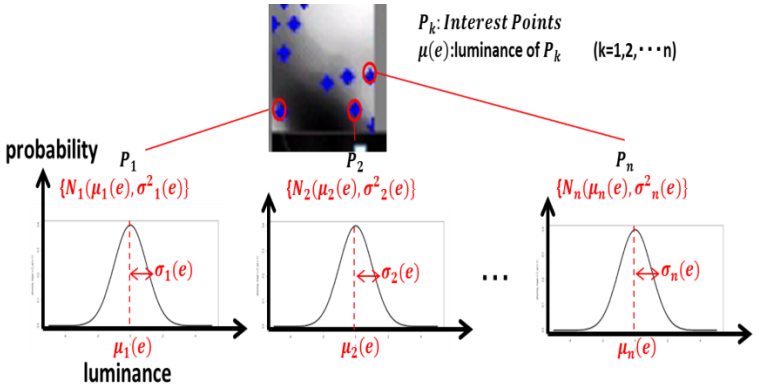


Fig.5 Intensity distribution model

Chewing detection using the likelihood of chewing

Figure 6 shows the chewing detection using the likelihood of chewing. The likelihood P (z | e) of the areas z for the left and right cheek as well as chin in a state e is represented assuming the independence of each attention point,

$$\mathbf{P(z|e)} = \prod_{i=1\dots k} \boldsymbol{P(z_i|e)} \qquad (4)$$

where, the likelihood $P(z_i|e)$ of the luminance which is a normal distribution is defined as follows.

$$\boldsymbol{P(z_i|e)} = \frac{1}{\sqrt{2\pi}\sigma_i(e)} \boldsymbol{exp}\left[-\frac{(z_i-\mu_i(e))^2}{2\sigma_i{}^2(e)}\right] \qquad (5)$$

In Eq. (5), since the shape model corresponds to an image in a stationary state without chewing, high likelihood means a state of no chewing and low likelihood means a state of chewing. Therefore, a threshold value T3 is set for the likelihood P (z | e), so that chewing is detected if P (z | e) is smaller than T3. As a matter

of fact, within the areas of the left and right cheeks as well as chin, we evaluate the chewing detection results of the two regions according to the orientation of the face against the camera. If the face is oriented to the right / left to the camera, we evaluate the chewing detection result of the areas for the left /right cheek and chin. Here, the chewing detection process causes a slight time delay. Therefore, we determine the output result as chewing if chewing is detected in the current frame or one frame before or one frame after in the two regions. Also, we do not detect anything for 0.5 seconds after the previous detection to prevent multiple detections during one chewing.

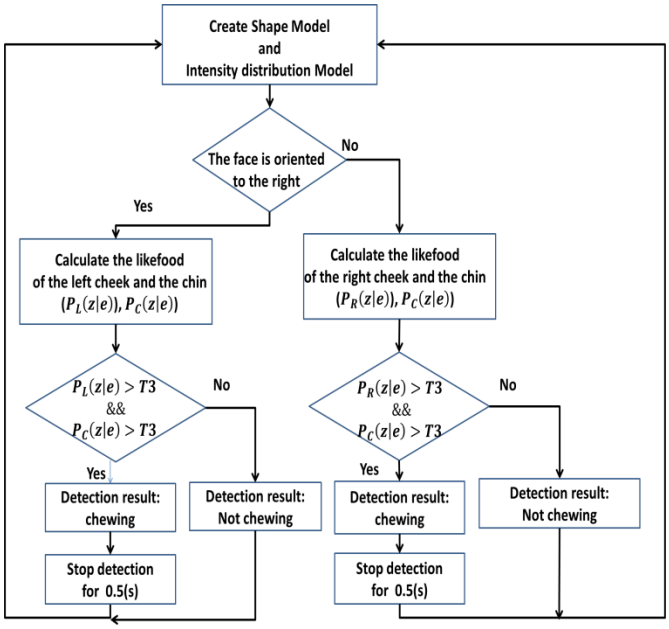


Fig.6 chewing detection

Evaluation

In order to confirm the effectiveness of the proposed method, we captured video sequences at 15 fps of seven subjects, six men and one woman in their 20s, who are eating snacks. Figure 7 shows the experiment environment.

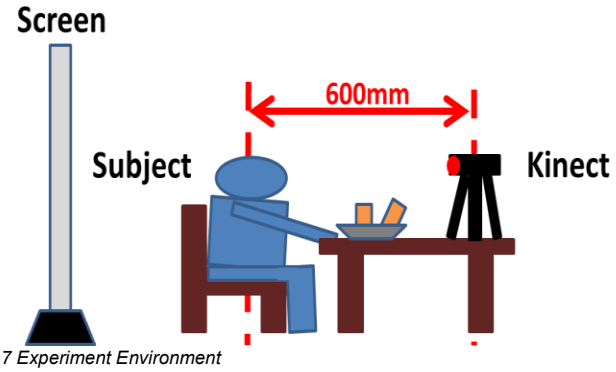


Fig.7 Experiment Environment

The subject sits at 600 mm away from the Kinect. Three snacks are placed in front of subjects and they eat at their own time. We placed a screen behind the subject to reduce background noise. Tbale.1 shows Machine Spec we used in the experiment.

Table.1 Machine Spec

PC	Dell XPS 8700
OS	Windows 8.1
CPU	Intel® Core^{TM} i7-4790 @3.60GHz
Memory	16.0 G
Programming Language	C++

In the experiments we used 40×40[pixel] images for the areas for the left and right cheeks as well as chin, respectively. In order to evaluate whether the chewing detection is performed correctly at the time when the subjects is chewing, we asked them to press the button when they chewed. In Fig. 8, the red line shows the result of chewing detection, where if the likelihood $P(z_i|e)$ is smaller than T3, the red line is 1; otherwise, 0. Here this line shows the evaluation of the chewing detection results of the two regions. In Fig. 8, the blue line shows the reaction (1: push the button; 0.2: no push) of the subjects. In addition, in order to investigate whether the proposed method detects chewing when the subjects do not chew, we asked them to stay calm (do not chew) for three seconds after finishing chewing.

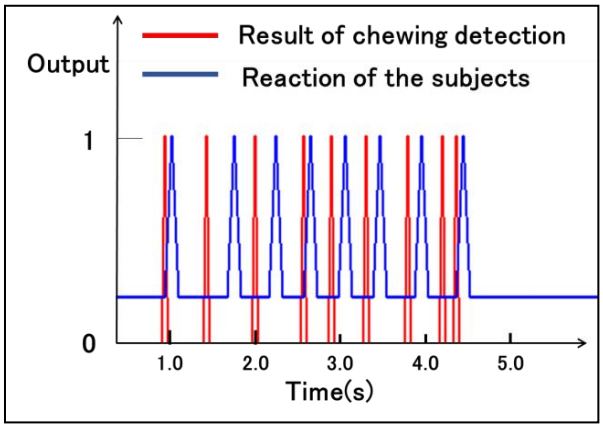


Fig.8 The result of chewing detecdtion and the reaction of the subjects

Evaluation is based on the calculation result of the recognition rate and the false detection rate. The recognition rate evaluates the number of times chewing was detected at the timing when the subjects pressed the chewing button. We defined the number of times the subject pressed the button as TN, and the number of times chewing was detected as GN. Therefore, if the recognition rate is Accuracy, the following expression is obtained.

$$Accuracy = {}^{GN}/_{TN} \quad (6)$$

Also, we defined the number of times chewing was detected without the subjects chewing as FN. Then, if the false detection rate is False, the following expression is obtained.

$$False = {}^{FN}/_{TN + FN} \quad (7)$$

Taking into consideration the mismatch of the response time when subjects push the button according with chewing, we define the output result as a correct detection in the case that chewing is detected within 60ms before or after subjects push the button.

Table.2 the experiment result

subgect	TN	GN	FN	Accuracy(%)	False(%)
A	27	26	3	96.2	10.0
B	78	70	4	89.7	4.88
C	51	24	15	47.1	22.7
D	102	73	15	71.6	12.8
E	47	45	5	95.7	9.62
F	102	100	11	98.0	9.73
G	73	60	10	82.2	12.0

Table.2 shows the result of detecting chewing. The sum of the seven subject's chewing frequency (times) was 480, and the sum of the times of the chewing detected by the proposed method was 398. The accuracy of chewing detection is about 83%, which is quite good accuracy. Also, the average of False is 11.7%.

Conclusion

This paper has proposed a method for detecting whether the care receiver is chewing, from the video sequence of that receiver. The proposed method detects the receiver's face and areas for both cheeks and chin. After applying some normalization to the areas, chewing detection that uses a variable-intensity template is performed, where the template consists of shape models, interest points and intensity distribution model. A likelihood based on the variable-intensity template is computed so that the receiver is judged whether the receiver is chewing. Experiments using seven subjects are conducted. As a result, the accuracy of chewing detection by the proposed method is 83%, which is quite promising. The result of chewing detection of Subject C was not good. This is because the change in the luminance of the left and right cheeks and the chin area was small because the test subject's chewing was weak. In the future, it will be necessary to construct a system that can evaluate the strength of chewing and devise such as changing the threshold of the likelihood in the variable intensity template according to the strength of chewing. In addition, in this experiment, hard food was used in the experiment as much as possible for easier detection of chewing, but the strength of chewing depends on food. Therefore, in order to make it possible to detect chewing for all foods, it will be necessary to add the information of the hardness of food to the system.

Acknowledgement

I would like to thank Prof. Itao for useful discussion. I am grateful to Mr. Yano and Mr. Kanda for assistance with the experiment.

References

[1] World Health Media Centre, "World Health Organization," December,2014.[Online]Available:

http://www.who.int/mediacentre/factsheets/fs352/en/

[2] S.Hartley, V. Ilagan, R. Madden, A Officer, A. Posarac, K Seelman, T. Shakespare, S. Sipos, and Swanson, "WORLD REPORT ON DISABILITY" World Health Organaization, Malta ,2011

[3] Syuji UNO, Jun KAWAI, Sigeo KANEDA,"Proposal of swallowing disorder inspection method using detection of chewing / swallowing coordinated movement",Information Processing Society of Japan,2012

[4] Syuji UNO, Ryo Arizumi, Sigeo KANEDA, Hirohide Haga, "Advising the Number of Mastication by Using Bone-Conduction Microphone", Annual Conference of the Japanese Society for Artificial Inteligence,2010

[5] Hidekazu TANAKA, Naoya KOIZUMI, Yuji UEMA, Kouta MINAMIZAWA, Masahiko INAMI, "Augmened food texture rendering system using chewing detection device", Proceedings of the Japan Virtual Reality Society Conference Proceedings , 2011

[6] Tuyosi SYINNO, Hirotoshi AMEMIYA, Hirohide HAGA, Sigeo KANEDA "Proposal of a chewing frequency guidance system using moving image processing", Information Processing Society of Japan, 2010

[7] Shiro KUMANO, Kazuhiro OTSUKA, Junji YAMATO, Eisaku MAEDA, and Yoichi SATO, "Robust Facial Expression Recognition Method Using Variable-Intensity Template for Head Pose Variation",MIRU, 2007

[8] Yasuharu MATUBARA and Takeshi SHAKUNAGA,"Sparse Template Matching andIts Application to Real-time Object Tracking",CVIM 11,2005

Author Biography

Atsuto FUJIMOTO is now a current Master Degree in the department of MME (Modern Mechanical Engineering) in Waseda University. He got the Bachelor Degree in Waseda University (2016). He is currently working on the visual functions of Meal assistance robot.

Dr. Jun OHYA is a professor at the Department of Modern Mechanical Engineering, Waseda University, Japan. He earned his B.S., M.S., and Ph.D. degrees in Precision Machinery Engineering from the University of Tokyo in 1977, 1979, and 1988, respectively. Dr. OHYA is a member of IEEE, IEICE, the Information Processing Society of Japan, etc. His research fields include image processing, computer vision, virtual reality, multimedia, pattern recognition.

Junji Yamato(M'90-SM'05)received B.E. and M.E. degrees from the University of Tokyo, Japan, in 1988 and1990, respectively. He received an M.S. degree from the MIT in 1998, and a Ph.D. degree from the Universityof Tokyo in 2001. He is currently the Executive Manager of Media Information Laboratory, NTT Communication Science Laboratories. His research interests include computer vision, gesture recognition and human-robot interaction.

Takaaki Ohkawauchi earned his Ph.D. degrees in Global Information and Telecommunication Studies from the University of Waseda in 2013. He was a program coordinator of Rikkyo University Social Information Education Research Center. He is now a lecturer at Teikyo University General Education Center