Paper: Rb24-2-5302; 2012/3/23

Mouth Movement Recognition Using Template Matching and its Implementation in an Intelligent Room

Kiyoshi Takita*, Takeshi Nagayasu*, Hidetsugu Asano, Kenji Terabayashi*, and Kazunori Umeda***

*Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
E-mail: takita@sensor.mech.chuo-u.ac.jp
**Pioneer Corporation
1-1 Shin-ogura, Saiwai-ku, Kawasaki-shi, Kanagawa 212-0031, Japan
[Received 00/00/00; accepted 00/00/00]

This paper proposes a method of recognizing movements of the mouth from images and implements the method in an intelligent room. The proposed method uses template matching and recognizes mouth movements for the purpose of indicating a target object in an intelligent room. First, the operator's face is detected. Then, the mouth region is extracted from the facial region using the result of template matching with a template image of the lips. Dynamic Programming (DP) matching is applied to a similarity measure that is obtained by template matching. The effectiveness of the proposed method is evaluated through experiments to recognize several names of common home appliances and operations.

Keywords: intelligent room, mouth movement recognition, template matching, image processing

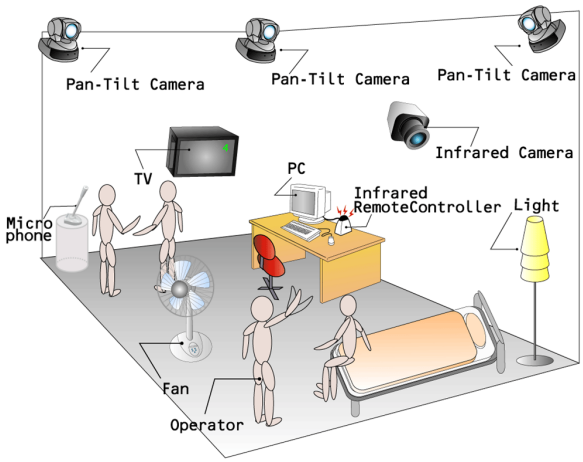


Fig. 1. Conceptual figure of an intelligent room.

1. Introduction

Home appliances have become essential to our everyday lives. However, their multiple functions and advanced functions often cause problems of operational complexity. There have been many studies on using human gestures to perform intuitive operations of home appliances or other products familiar to people [1–4]. As shown in **Fig. 1**, we set up cameras in the four corners of a room and built an intelligent room in which the gestures of the operator are recognized to operate home appliances [5, 6]. This system recognizes the direction in which the operator is pointing to select a home appliance to be operated. However, direction determination is difficult, and its stability depends on the camera angle. In addition, in situations in which both hands are full, such as in situations in the kitchen, the operator cannot gesture with his/her hands. We have therefore built a highly operable system that can meet the operator's requirements even if his/her body motion is limited. Such methods include voice recognition. However, since voice recognition is susceptible to noise, it is difficult to obtain stable recognition results in situations in which there is no microphone mounted. In terms of cost as well, a system that operates only with cameras is more desirable than one that needs cameras and microphones to determine the operator in the room. We have built an intelligent room [5, 6] equipped with pan-tilt-zoom cameras which capture the operator's face with sufficient size in a field of view of the cameras. A prior study [7] focuses on movements of the lips and used mouth movement recognition. The operator voices a function that he/she desires and, in so doing, selects an operation target and operates a home appliance. Some mouth movement recognition methods have already been proposed [7–17]. These can be categorized into the model-based methods [8–14] and the image-based methods [7, 14–17]. The model-based methods use geometric shapes to detect feature points, such as the corners and the contour of the lips, and thus obtain the area and width of the lips or the area of the open mouth region as features. This method has less data volume and is robust against environmental changes, but the difficulty of creating a model is a problem. On the other hand, as the image-based method uses an image of the area around the mouth, it does not require the building of a complicated model, and it is thereby easy to obtain data.

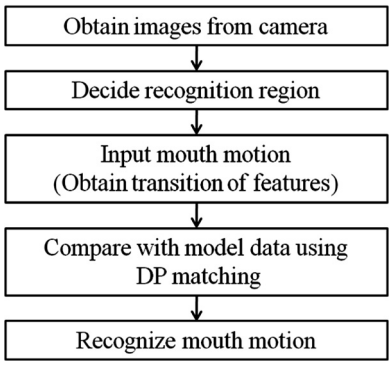


Fig. 2. Flow of mouth movement recognition.

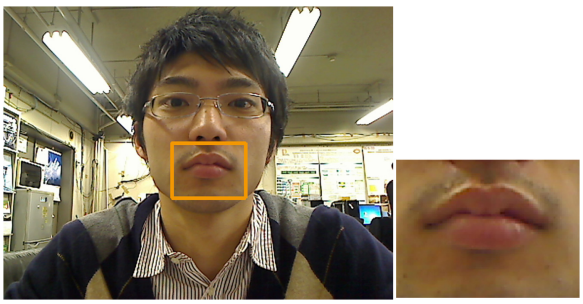
However, since this method is based on intensity information from images, it is susceptible to the position, size, and shade of the mouth in the image. In an effort to handle the problems inherent in the image-based method, Nakanishi et al. determine the recognition range according to the size of the operator's face and the centroid position of the region containing the open mouth in the image. This reduces variation in the size of the mouth. In addition, images of the recognition range are reduced in resolution so as to respond to the mouth position displacement that occurs during phonation [7]. In practice, for mouth movement recognition implemented in the intelligent room, we use the following features: a mouth image with reduced resolution, the shape of the lips, and the shape of the open mouth. Recognition using these features can achieve a high recognition rate as long as the images are captured by a fixed camera. However, if multiple network cameras are used, as in an intelligent room, images need to be compressed to reduce data volume. Compressed images have low image quality, so these features make it impossible to stably obtain the data. In addition, this type of recognition is also susceptible to changes in lighting due to changes in the position of the operator and in the lighting environment peculiar to rooms, such as changes caused by the source of illumination being on or off [14].

This research proposes mouth movement recognition using template matching as a method of obtaining the features even from compressed images. In addition, the usefulness of the proposed method is discussed through evaluation experiments. The proposed method is also implemented in an intelligent room, so its usefulness in a real environment is discussed.

2. Flow of Mouth Movement Recognition Processing

The flow of the proposed mouth movement recognition method is presented in **Fig. 2**. First, a range of recognition is determined. After the range is determined, the operator starts moving his/her mouth. These movements are recognized by performing DP matching, matching time-series data of the features with pre-registered model data.

(a) Distance from a camera: 1 m



(b) Distance from a camera: 0.3 m

Fig. 3. Region for recognition of mouth movements.

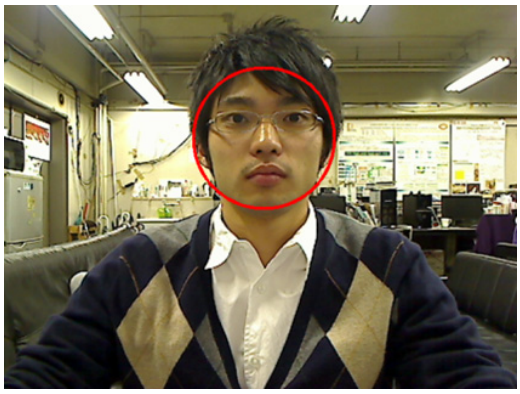


Fig. 4. Detection of facial region.

3. Recognition Range Determination Method

To recognize the mouth movements, face detection and lip detection are conducted on the obtained images in sequence to determine the recognition range. At that time, as shown in **Fig. 3**, the recognition range is determined so that the proportion of the lips in the recognition range is constant, even if the distance between the operator and the camera differs.

3.1. Face Detection

The OpenCV face detection module [a] is applied to the loaded image to determine the position and size of the face. We use the Viola and Jones' method [18]. This module encircles an area that is supposed to be a face, as seen in **Fig. 4**. The center of the circle is the center of the face, and the diameter of the circle is the size of the face.

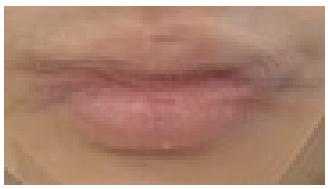


Fig. 5. Template image of lip.

Fig. 6. Result of lip detection.

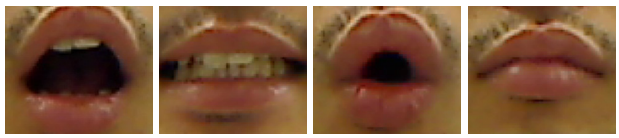
3.2. Lip Detection

The position of the mouth is determined from the detected center and size of the face. In addition, the area around the determined position is extracted as the mouth region. The mouth position is obtained empirically. The lips are extracted by performing template matching, matching the template of a lip image with the extracted region. The lip template is created by synthesizing multiple images of lips (**Fig. 5**), thereby allowing detection for any human subject. **Fig. 6** presents an example of lip detection. Multiple templates of different sizes are prepared, and the width and height of the template with the highest similarity are designated as the width and the height of the lips. This allows the proportion of the lips in the recognition range to be constant even if the distance from the camera changes. The similarity is obtained by normalized correlation of the following expression.

$$\bar{I} = \frac{1}{MN}\sum_{j=0}^{N-1}\sum_{i=0}^{M-1} I(i,j), \quad \bar{T} = \frac{1}{MN}\sum_{j=0}^{N-1}\sum_{i=0}^{M-1} T(i,j)$$

$$R_{ZNCC} = \frac{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}((I(i,j)-\bar{I})(T(i,j)-\bar{T}))}{\sqrt{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1}(I(i,j)-\bar{I})^2 \times \sum_{j=0}^{N-1}\sum_{i=0}^{M-1}(T(i,j)-\bar{T})^2}} . \quad (1)$$

Let $M \times N$ be the size of the template, $T(i,j)$ be the pixel value of the template at the position (i,j) and $I(i,j)$ be the pixel value of the target image that are overlapped with the template. The recognition range is determined from the center, width, and height of the detected lip template, and the image of the recognition range is resized to a specific size. The position and the size of the recognition range are fixed during the recognition of mouth movements.



(a) "a" (b) "i" (c) "u" (d) "m"

Fig. 7. Template images for recognition.

4. Mouth Movement Recognition Method

After the system determines the recognition range, the operator starts moving his/her mouth, time-series data of the features are compared with the pre-registered model data using DP (Dynamic Programming) matching [19, 20], and the entered word is recognized.

4.1. Recognition of the Beginning and End of Mouth Movements

For the system to automatically recognize the timing of the beginning and the end of mouth movements, after the recognition range is determined, the differences between the consecutive frames of the image that are converted to low-resolution are calculated. If the difference is equal to or greater than a threshold value, the system recognizes mouth movements as having started. After the mouth movements have started, if the difference is equal to or less than the threshold value, the system recognizes them as having ended.

4.2. Acquisition of Features

As features for recognizing mouth movements, we use the similarities that can be obtained by template matching. As a template, a total of four images with different mouth shapes are prepared. These are, more specifically, three images of the phonation of "a," "i," and "u," which are three of the five vowels used in the Japanese language, and one image of the state of "m," the closing of the mouth. The remaining two vowels, "e" and "o," are not used in this research because the mouth when these vowels are phonated is very similar to that when "i" and "u" are produced, respectively. Examples of individual images are presented in **Fig. 7**. The template is made smaller than that of the recognition range to remove the effect of mouth position displacement in the recognition range (**Fig. 8**). Each of the four images is used as a template in each frame, and template matching is performed once for each template. Four similarities obtained as a result of the template matching are obtained as features. The similarity is calculated using Eq. (1). As an example, the transition of similarity measured when the word "light" is said is shown in **Fig. 9**.

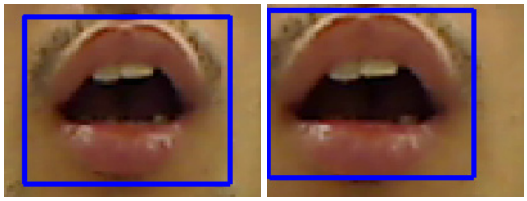


Fig. 8. Cancellation of position shift of mouth.

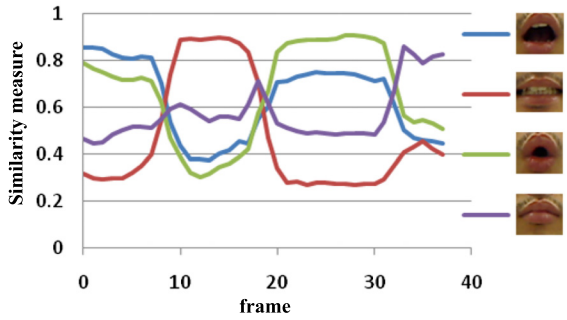


Fig. 9. Transition of similarity measure when "light" is said.

4.3. Comparison with Model Data

Comparison with model data is performed using DP matching [19, 20]. DP matching is a method that takes the expansion and contraction of a pattern into consideration, even if the number of input data and the number of model data are different due to a difference in phonation speed. The number of frames of model data and input data are denoted by M and N, respectively. The result of the calculation of the template matching of model data at the m^{th} frame with the t^{th} template of the four templates is denoted by $MTemp[m][t]$, and the result of the calculation of template matching of input data of the n^{th} frame with the t^{th} template of the four templates is denoted by $Temp[n][t]$. The distance $TPD[m][n][t]$ ($m=1,\ldots,M$, $n=1,\ldots,N$, $t=1,\ldots,4$) is then obtained by the following expression.

$$TPD[m][n][t] = \frac{|MTemp[m][t] - Temp[n][t]|}{\sqrt{MTemp[m][t])^2 + (Temp[n][t])^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The distance $TTD[m][n][t]$ between the data pair of the $(1,1)^{\text{th}}$ frame and the data pair of the $(m,n)^{\text{th}}$ frame in the initial state is obtained by the following expression.

$$\begin{aligned} TTD[m][n][t] = \min\{&TTD[m-1][n-1][t] \\ &+2TPD[m][n][t], \\ &TTD[m][n-1][t]+TPD[m][n][t], \\ &TTD[m-1][n][t]+TPD[m][n][t]\} \end{aligned} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Using Eq. (3), the distance $TTD[M][N][t]$ between the data pair of the $(1,1)^{\text{th}}$ frame in the initial state and the data pair of the $(M,N)^{\text{th}}$ frame in the terminal state is obtained. The distance obtained is normalized, and the distance $TValue[t]$ between the model data and the input data is obtained using the following expression.

$$TValue[t] = \frac{TTD[M][N][t]}{M+N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

The distance $TValue[t]$ for each feature obtained by Eq. (4) is added together using the following expression so that the distance $AllTValue$ of the total features is obtained.

$$AllTValue = \sqrt{TValue[1]^2+TValue[2]^2+TValue[3]^2+TValue[4]^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

The above processing is carried out for all the registered model data, and the model data with the minimum distance between the model data and the input data are used as a recognition result.

Fig. 10. Input image using PTZ camera.

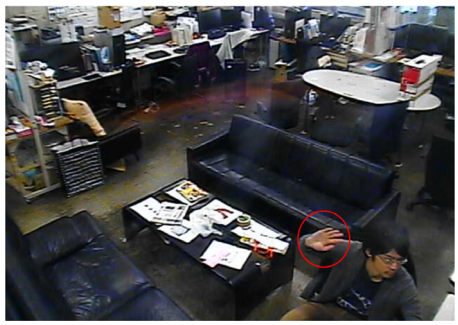


Fig. 11. Detection of waving hand.

5. Implementation in Intelligent Room

The mouth movement recognition system is implemented in the intelligent room [6] that we have built. In the intelligent room, the operator is determined by detecting a waving hand. Pan, tilt, and zoom are performed on the detected hand waving position, and the operator's hand is captured in the field of view of the camera at a size sufficiently large for the hand gesture to be recognized. The way in which this method is implemented is as follows. First, the operator is determined by a waving hand being detected. **Fig. 10** shows an image obtained from the camera, and **Fig. 11** shows detection of the waving hand. Next, pan, tilt, and zoom are preformed, directed at the

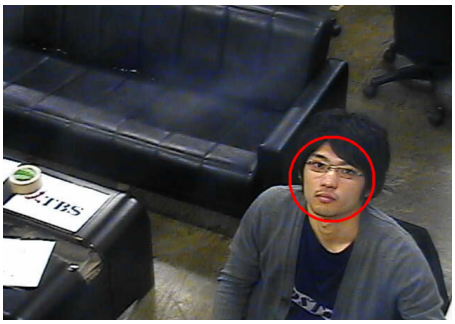


Fig. 12. Face detection after first zoom.

Fig. 13. Face detection after second zoom.

detected position of the waving hand. After the first zoom is finished, face detection is performed using the method described in Section 3.1 (**Fig. 12**). After that, the camera zooms in on the center of the operator's face a second time so that an image of sufficient size may be captured in the field of view of the camera. Finally, the recognition range is determined using the method described above, and the mouth motion is recognized. **Fig. 13** shows face detection after the camera zooms in on the face.

6. Experiments

To verify the usefulness of the mouth movement recognition system we have built, we conducted experiments. Section 6.1 discusses the usefulness of the system when fixed cameras are used, and Section 6.2 discusses the usefulness of the system when implemented in the intelligent room.

6.1. Experiments Using Fixed Cameras

6.1.1. Experiment System

Our experiment system is composed of a Webcamera C905m (Logicool, 640×480 pixels), a PC (Core i7 CPU 930 2.80 GHz, DDR3 6.00 GB), and image processing software (Intel OpenCV).

Parameters used for recognition range determination are set as follows. The face radius obtained by face detection is denoted by r, a range $(-r/2, r/2)$ in the lateral direction from the center of the face and $(0, r)$ in the longitudinal direction from the center of the face is extracted, and lip detection is performed on the extracted range. The

color	Resolution	Template Image
RGB	150 × 130	
	15 × 13	
Gray	150 × 130	
	15 × 13	

Fig. 14. Template image.

Table 1. Phonation time [ms].

	Video	TV	Light	Up	Down
Average time	488	606	550	533	377

template used for the lip detection is created by synthesizing seven images of lips. We create four templates with different sizes. We empirically set the sizes of the four templates: 80×44, 100×55, 115×63, and 130×77 pixels. The width and the height that are obtained by lip detection are denoted by w and h, respectively, and a range $(-3/4w,\ 3/4w)$ in the lateral direction from the center of the lips and $(-h/2-5,\ 9/8w-h/2-5)$ in the longitudinal direction from the center of the lips is extracted and designated as a recognition range. The image of the recognition range is resized to 200×150 pixels, and the size of the template used to obtain the features is set to 150×130 pixels, which is smaller than the recognition range. Mouth movement is considered started when 15 or more pixels have the pixel value difference between frames of 12 or more in an image with resolution reduced to 12×9 pixels. The sum of absolute values of each pixel value difference must also be 600 or more. Mouth movement is considered ended when there are 20 consecutive frames in which the sum of absolute values of each pixel value difference is 300 or less.

6.1.2. Recognition Experiments

We set up the camera in front of the subject at almost the same height as the subject's face, 0.4 m, and we conducted experiments with one subject under fluorescent light. Words to be recognized were "video," "TV," "light," "up," and "down." As model data, we used time-series data of the features when the subject spoke each word once. We used four types of templates (**Fig. 14**) that were prepared by combining two types of images, i.e., an RGB image and its gray-scale image of the recognition range, with two other types of images, i.e., a template with its resolution reduced to 15×13 pixels and a template without reduced resolution. After the subject performed each mouth movement 50 times, we examined the recognition rate. **Table 1** presents the average phonation time of each of the five words, and **Table 2** presents the processing time of template matching per frame and the processing time of DP matching.

First, we conducted an experiment using the templates and model data of the subject him/herself. The results are presented in **Table 3**. In the following tables, "color"

Table 2. Processing time of template matching and DP matching [ms].

matching [ms]

Method	Template matching				DP matching
Color	RGB	RGB	Gray	Gray	
Resolution	150×130	15×13	150×130	15×13	
Average time	18.92	1.12	8.74	0.96	3.11

Table 3. Results of subject using his/her own template and model data [%].

Feature / Word		Video	TV	Light	Up	Down	Average
Color	Resolution						
RGB	150×130	98	98	98	96	100	98.0
RGB	15×13	98	100	98	100	100	99.2
Gray	150×130	98	100	90	90	100	94.4
Gray	15×13	100	100	94	100	100	98.8

Table 4. Results of subject using another's template and model data [%].

Feature / Word		Video	TV	Light	Up	Down	Average
Color	Resolution						
RGB	150×130	100	68	60	94	18	68.0
RGB	15×13	82	76	78	32	86	70.8
Gray	150×130	58	98	78	90	76	80.0
Gray	15×13	98	98	96	38	80	82.0

represents color information where an RGB image is represented by "RGB," a gray-scale image is represented by "gray," and "resolution" represents the size of an image.

In this case, we obtained the highest average recognition rate, 99.2%; even the lowest was 94.4%. These are high recognition results regardless of the type of template. This is because the features were obtained stably. When using templates and model data of the subject him/herself, the proposed mouth movement recognition system is useful.

Next, we conducted an experiment using templates and model data of a person other than the subject in order to examine the recognition rate when the subject and the person who produced the model data were different. The results are presented in **Table 4**.

In this case, the recognition rate was lower than that when templates and model data of the subject him/herself were used. This is because the use of another person's templates may produce higher similarity of a template when a different sound is produced, even if the operator and the template assume the same shape when producing the same sound. Thus, it was not possible to obtain stable features. Among them, the highest recognition result was 82.0%. This was the average recognition rate, obtained when a template with gray-scale recognition range and low resolution was used. This is because the gray-scale and low resolution reduced the difference in the shape of the mouth shape between individuals. Considering the recognition results by word at this time, when saying "video," "TV," "light," and "down," relatively high recognition results were obtained. On the other hand, when "up" was said, a low recognition result was obtained. Results of other templates indicate that there are words with high recognition rates and others with low recognition rates. This suggests that there are great differences in recognition rates between phonated words.

6.1.3. Recognition Experiment with Increased Number of Categories

We conducted an experiment with ten words for recognition, "video," "TV," "light," "up," "down," "channel," "volume," "telephone," "music," and "OK." We used model data and templates from the subject him/herself. The "RGB, 15×13" template type, which exhibited the highest recognition rate in the experiment in Section 6.1.2, was used. After the subject performed each mouth movement 50 times, we examined the recognition rate. The results are presented in **Table 5**. The average phonation time and DP matching processing time of the five newly added words are presented in **Table 6**.

In this experiment, the average recognition rate was 94.4%. Compared to when five words were targeted, the recognition rate was lower but still high. The recognition rate was lower than that in the experiment with five words because an increase in the number of words resulted in an increase in words with similar time-series data of the features. This is best understood by the result of the input of the word "up." "Up" has the lowest recognition rate and was misidentified as "channel" in the most erroneous recognition. This is because both "up" and "channel" have a similar progression of mouth shapes when they are spoken; thus there is less difference in time-series data of the features. The shape progression starts from a state of "a" with the mouth wide open, then a state of "m" with the mouth once closed, and lastly a state of "u" with the mouth slightly open. As this shows, if there are words with a similar progression of shapes of the mouth, the recognition rate is lowered.

The above results indicate that the proposed method can be used practically as long as the number of words to be recognized is about ten.

6.2. Experiments Using Cameras in the Intelligent Room

6.2.1. Experimental System

We used the camera that was set up in the intelligent room, an AXIS 233D pan-tilt-zoom camera (640×480 pixels). Other details were the same as those outlined in Section 6.1.1.

6.2.2. Recognition Experiments

We conducted the experiments under fluorescent light. The height of the subject's face was -1.4 m, front 2 m, 3 m, and 4 m with respect to the pan-tilt-zoom camera.

Table 5. Results of subject using his/her own template and model data [%].

Output / Input	Video	TV	Light	Up	Down	Channel	Volume	Telephone	Music	OK
Video	100	0	0	0	0	0	0	0	0	0
TV	0	100	0	0	0	0	0	0	0	0
Light	0	0	88	0	0	6	0	2	4	0
Up	0	0	0	72	0	26	2	0	0	0
Down	0	0	0	0	100	0	0	0	0	0
Channel	0	0	0	0	0	98	0	0	2	0
Volume	0	0	0	0	0	0	98	0	0	2
Telephone	0	0	0	0	0	0	0	94	0	0
Music	0	0	0	0	0	4	0	0	96	0
OK	0	0	0	0	0	0	0	2	0	98

Table 6. Phonation time and processing time of DP matching [ms].

	Channel	Volume	Telephone	Music	OK	DP matching
Average time	854	750	756	729	725	6.40

(a) 2 m (b) 3 m

(c) 4 m

Fig. 15. Images of intelligent room camera.

Table 7. Results of subject using his/her own template and model data with the PTZ camera in the intelligent room.

(a) Distance: 2m [%]

Feature: Color	Feature: Resolution	Word: Video	TV	Light	Up	Down	Average
RGB	150×130	98	100	92	84	96	93.2
RGB	15×13	90	98	96	94	94	94.4
Gray	150×130	82	100	96	82	96	91.2
Gray	15×13	90	100	80	98	82	90.0

(b) Distance: 3m [%]

Feature: Color	Feature: Resolution	Word: Video	TV	Light	Up	Down	Average
RGB	150×130	86	94	70	94	56	80.0
RGB	15×13	88	98	86	24	90	77.2
Gray	150×130	92	88	100	40	96	83.2
Gray	15×13	100	100	78	2	94	74.8

(c) Distance: 4m [%]

Feature: Color	Feature: Resolution	Word: Video	TV	Light	Up	Down	Average
RGB	150×130	92	62	42	96	16	61.4
RGB	15×13	82	74	6	0	90	50.4
Gray	150×130	52	98	14	96	4	52.8
Gray	15×13	84	100	38	66	84	74.4

At all the positions, the subject was facing the camera. **Fig. 15** shows images captured by the camera at each position. As model data, we used time-series data of the features when the subject spoke each word once. After one subject produced each mouth movement 50 times, we examined the recognition rate. Words to be recognized were "video," "TV," "light," "up," and "down." We used model data and templates of the subject him/herself at a position where the camera was 2 m in front of the subject. As outlined in Section 6.1.2, we used four types of templates that were prepared by combining two types of images, i.e., an RGB image and its gray-scale image of the recognition range, with two types of images, i.e., a template with its resolution reduced to 15×13 pixels and a template with no reduction in resolution. The results are presented in **Table 7**.

At the 2 m position, high recognition results were obtained regardless of the templates. This is because the distance at which the templates and the model data were obtained and the distance at which the recognition experiment was conducted were the same, and thus time-series data of the features similar to the model data were obtained.

At the 3 m position, higher recognition results were obtained using templates without reduced resolution. This is because, although the position at which the experiment was conducted and the position at which the template and the model data were obtained were different, there was less difference from the template image, and thus time-series data of the features similar to the model data were obtained without reduced resolution.

At the 4 m position, the highest recognition result was obtained when using the gray-scale, low resolution template. This is because the gray-scale, low resolution tem-

plate best handled the difference in lighting and the angle of the mouth, that are due to differences in the distance from which the template was obtained. This indicates that the proposed method will also work in the real environment of an intelligent room. However, the type of template that will attain the highest recognition rate depends on the position of the operator. Therefore, by changing the type of template in accordance with the distance, a system with a high recognition rate can be built regardless of operator position. In addition, while this experiment assumed a state in which the operator faced the camera, there may be times when the operator has no choice but to be at an angle to the camera in the room. If this happens, the apparent difference from the template becomes great. However, the results of this experiment indicate that the use of low-resolution template images can compensate for the apparent differences to some extent.

7. Conclusions

This study has proposed mouth movement recognition using template matching as a method that allows for stable features to be obtained. In addition, a mouth movement recognition system using the proposed method has been implemented in an intelligent room. With the proposed method, the operator's face is detected, then lip detection is performed using template matching from an extracted mouth region, and after that a recognition range is determined. Four prepared images of different mouth shapes are used as templates to obtain four similarities by means of template matching. The prepared templates of the mouth shapes are smaller than the recognition range so that mouth position displacement is handled successfully. Time-series data of similarities are used as input data and phonated words are recognized by means of DP matching.

The usefulness of the proposed method has been verified through experiments using a fixed camera and cameras in an intelligent room. The experiments using the fixed camera indicate that a high recognition rate can be obtained using model data and templates of the subject him/herself. In addition, a high recognition rate can be obtained even if the number of words to be recognized is increased to ten. Operations of home appliances can be hierarchized so that the number of necessary choices is reduced to ten or less. Therefore, this method can be satisfactory for the operation of home appliances in everyday life. The experiments using the cameras in the intelligent room indicate that the type of template that gives a high recognition results depends on the distance between the operator and the camera.

In the future, we intend to build a system that achieves a stable, high recognition rate by combining face recognitions to determine individuals, using model data and templates of the operator him/herself, and selecting templates in accordance with the distance between the operator and camera.

Acknowledgements

This work was supported by KAKENHI (19100004).

References:

[1] T. Mori and T. Sato, "Robotic Room: Its Concept and Realization, Robotics and Autonomous Systems," Robotics and Autonomous Systems, Vol.28, No.2, pp. 141-144, 1999.

[2] J. H. Lee and H. Hashimoto, "Intelligent Space – Concept and Contents –," Advanced Robotics, Vol.16, No.4, pp. 265-280, 2002.

[3] T. Mori, H. Noguchi, and T. Sato, "Sensing room – Room-type behavior measurement and accumulation environment –," J. of the Robotics Society of Japan, Vol.23, No.6, pp. 665-669, 2005.

[4] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer, "EasyLiving: Technologies for Intelligent Environments," Proc. Int. Symp. on Handheld and Ubiquitous Computing, pp. 12-27, 2000.

[5] K. Irie, N. Wakamura, and K. Umeda, "Construction of an intelligent room based on gesture recognition – operation of electric appliances with hand gestures," Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 193-198, 2004.

[6] K. Irie, N. Wakamura, and K. Umeda, "Construction of an Intelligent Room Based on Gesture Recognition," Trans. of the Japan Society of Mechanical Engineers, C, Vol.73, No.725, pp. 258-265, 2007 (in Japanese).

[7] T. Nakanishi, K. Terabayashi, and K. Umeda, "Mouth Motion Recognition for Intelligent Room Using DP Matching," IEEJ Trans. EIS Vol.129, No.5, pp. 940-946, 2009 (in Japanese).

[8] T. Saitoh and R. Konishi, "Lip Reading Based on Trajectory Feature," Trans. IEICE Vol.J90-D, No.4, pp. 1105-1114, 2007 (in Japanese).

[9] T. Wark and S. Sridharan, "A syntactic approach to automatic lip feature extraction for speaker identification," Proc. IEEE ICASSP, Vol.6, pp. 3693-3696, 1998.

[10] R. W. Frischholz and U. Dieckmann, "Bioid: A Multimodal Biometric Identification System," IEEE Computer, Vol.33, No.2, pp. 64-68, 2000.

[11] L. G. ves da Silveira, J. Facon, and D. L. Borges, "Visual speech recognition: A solution from feature extraction to words classification," Proc. XVI Brazilian Symposium on Computer Graphics and Image Processing, pp. 399-405, 2003.

[12] M. J. Lyons, C.-H. Chan, and N. Tetsutani, "Mouthtype: text entry by hand and mouth," Proc. Conf. on Human Factors in Computing Systems, pp. 1383-1386, 2004.

[13] Y. Ogoshi, H. Ide, C. Araki, and H. Kimura, "Active Lip Contour Using Hue Characteristics Energy Model for A Lip Reading System," Trans. IEEJ, Vol.128, No.5, pp. 811-812, 2008.

[14] K. Takita, T. Nagayasu, H. Asano, K. Terabayashi, and K. Umeda, "An Investigation into Feature for Mouth Motion Recognition Using DP matching," Dynamic Image processing for real Application 2011, pp. 302-307, 2011 (in Japanese).

[15] C. Bregler and Y. Konig "'Eigenlips' for robust speech recognition," Proc. Int. Conf. Acoust. Speech Signal Process (ICASSP), pp. 669-672, 1994.

[16] O. Vanegas, K. Tokuda, and T. Kitamura, "Lip location normalized training for visual speech recognition," IEICE Trans. Inf. & Syst., Vol.E83-D, No.11, pp. 1969-1977, Nov. 2000.

[17] J. Kim, J. Lee, and K. Shirai, "An efficient lip-reading method robust to illumination variation," IEICE Trans. Fundamentals, Vol.E85-A, No.9, pp. 2164-2168, Sept. 2002.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Vol.1, pp. 511-518, 2001.

[19] T. Nishimura, T. Mukai, S. Nozaki, and R. Oka, "Spotting Recognition of Gestures Performed by People form a Single Time-Varying Image Using Low-Resolution Features," Trans. IEICE, Vol.J80-D-II, No.6, pp. 1563-1570, 1997.

[20] S. Uchida and H. Sakoe, "Analytical DP Matching," Trans. IEICE, Vol.J90-D, No.8, pp. 2137-2146, 2007.

Supporting Online Materials:

[a] http://opencv.jp/opencv-1.1.0/document/index.html

Name:
Kiyoshi Takita

Affiliation:
Course of Precision Engineering, Graduate School of Science and Engineering, Chuo University

Address:
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Brief Biographical History:
2011 Received B.Eng. in Precision Mechanics from Chuo University

Membership in Academic Societies:
• The Japan Society of Mechanical Engineers (JSME)

Name:
Takeshi Nagayasu

Affiliation:
Course of Precision Engineering, Graduate School of Science and Engineering, Chuo University

Address:
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Brief Biographical History:
2010 Received B.Eng. in Precision Mechanics from Chuo University

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)

Name:
Hidetsugu Asano

Affiliation:
Pioneer Corporation

Address:
1-1 Shin-ogura, Saiwai-ku, Kawasaki-shi, Kanagawa 212-0031, Japan

Brief Biographical History:
2002 Received M.Eng. degree in Electrical and Computer Engineering from Yokohama National University
2002- Researcher, Pioneer Corporation

Name:
Kenji Terabayashi

Affiliation:
Assistant Professor, Department of Precision Mechanics, Chuo University

Address:
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Brief Biographical History:
2004 M.Eng. in Systems and Information Eng. from Hokkaido Univ.
2008 Ph.D. in Precision Eng. from The Univ. of Tokyo
2008- Assistant Professor, Chuo University

Main Works:
• K. Terabayashi, N. Miyata, and J. Ota, "Grasp Strategy when Experiencing Hands of Various Sizes," eMinds: Int. J. on Human-Computer Interaction, Vol.1, No.4, pp. 55-74, 2008.
• K. Terabayashi, H. Mitsumoto, T. Morita, Y. Aragaki, N. Shimomura, and K. Umeda, "Measurement of Three Dimensional Environment with a Fish-eye Camera Based on Structure From Motion – Error Analysis," J. Robotics and Mechatronics, Vol.21, No.6, pp. 680-688, 2009.

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)
• The Japan Society for Precision Engineering (JSPE)
• The Japan Society of Mechanical Engineers (JSME)
• The Virtual Reality Society of Japan (VRSJ)
• The Institute of Image Electronics Engineers of Japan (IIEEJ)
• The Institute of Electrical and Electronics Engineers (IEEE)

Name:
Kazunori Umeda

Affiliation:
Professor, Department of Precision Mechanics, Chuo University

Address:
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Brief Biographical History:
1994 Received Ph.D. in Precision Machinery Engineering from The University of Tokyo
1994- Lecturer, Chuo University
2003-2004 Visiting Worker, National Research Council of Canada
2006- Professor, Chuo University

Main Works:
• M. Shinozaki, M. Kusanagi, K. Umeda, G. Godin, and M. Rioux, "Correction of color information of a 3D model using a range intensity image," Computer Vision and Image Understanding, Vol.113, No.11, pp. 1170-1179, Nov. 2009.
• K. Terabayashi, H. Mitsumoto, T. Morita, Y. Aragaki, N. Shimomura, and K. Umeda, "Measurement of Three Dimensional Environment with a Fish-eye Camera Based on Structure From Motion – Error Analysis," J. Robotics and Mechatronics, Vol.21, No.6, pp. 680-688, Dec. 2009.
• T. Kuroki, K. Terabayashi, and K. Umeda, "Construction of a Compact Range Image Sensor Using Multi-Slit Laser Projector and Obstacle Detection of a Humanoid with the Sensor," 2010 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2010), pp. 5972-5977, Oct. 2010.

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)
• The Japan Society for Precision Engineering (JSPE)
• The Japan Society of Mechanical Engineers (JSME)
• The Horological Institute of Japan (HIJ)
• The Institute of Electronics, Information and Communication Engineers
• Information Processing Society of Japan (IPSJ)
• The Institute of Electrical and Electronics Engineers (IEEE)