Facial Expression Recognition of a Speaker Using Front-view Face Judgment, Vowel Judgment, and Thermal Image Processing

T. Fujimura[1], Y. Yoshitomi[2], T. Asada[2], and M. Tabuse[2]

[1]*Works Applications Co. Ltd., 1-12-32 Akasaka, Minato-ku, Tokyo 107-6019, Japan,*
[2]*Graduate School of Life and Environmental Sciences Kyoto Prefectural University,*
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan,
E-mail: yoshitomi@kpu.ac.jp

*Abstract***:** For facial expression recognition, we previously selected three images: (i) just before speaking, and speaking (ii) the first vowel and (iii) the last vowel in an utterance. In this study, as a pre-processing module, we added a judgment function to discriminate a front-view face for facial expression recognition. A frame of the front-view face in a dynamic image is selected by estimating the face direction. The judgment function measures four feature parameters using thermal image processing and selects the thermal images that have all the values of the feature parameters within limited ranges decided on the basis of training thermal images of front-view faces. As an initial investigation, we adopted the utterance of the Japanese name "Taro," which is semantically neutral. The mean judgment accuracy of the front-view face was 99.3% for six subjects who changed their face direction freely. Using the proposed method, the facial expressions of six subjects were discriminable with 87.7% accuracy when he or she exhibited one of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised." We expect the proposed method to be applicable for recognizing facial expressions in daily conversation.

Keywords: Facial expression recognition, Front-view face judgment, Speech recognition, Vowel judgment, Thermal image processing

I. INTRODUCTION

To better integrate robots into our society, a robot should be able to interact in a friendly manner with humans. The goal of our research is to develop a robot that can perceive human feelings and mental states.

The first stage is to develop a method for integrating the information of human expression. The basic information for integration is the visible ray (VR) image, thermal image, and voice. In this first stage, an automatic, real-time, interactive system is not necessary. It will be very difficult to equip a robot with a computer which can process the information inputted to it with the efficiency of the human brain. Therefore, we chose to enable the robot to use a type of information, such as thermal imaging, that the human brain cannot process. Thermal imaging is a good example because it is impossible for a human to perceive heat via the naked eye.

The second stage is to develop an automatic, real-time, interactive system that has the information integration of human expression as a processing characteristic. The third stage is to develop a robot that has a function developed from the synthesis of the first and second stages for use in our daily lives.

The present investigation investigates the first stage of development, in which a robot can visually detect human feelings or inner mental states. Although recognizing facial expressions has received considerable attention in the field of computer vision research, the mechanism of recognition still falls far short of human capability, especially from the viewpoint of robustness under widely varying lighting conditions. One of the reasons is that the nuances of shade, reflection, and local darkness influence the accuracy of facial expression recognition through the inevitable change of gray levels.

To avoid this problem and to develop a robust method for facial expression recognition applicable under widely varied lighting conditions, we did not use a VR image, as would be expected. Instead, we used an image produced by infrared rays (IR), which describe the thermal distribution of the face [1]–[9]. Although a human cannot detect IR, it is possible for a robot to process the information of the thermal images created by IR. Therefore, as a new mode of robot vision, thermal image processing is a practical method that is viable under natural conditions.

The timing of recognizing facial expressions is also important for a robot because the processing might be time-consuming. In a previous report, we adopted an utterance as the key to expressing human feelings or mental states because humans tend to say something when expressing a feeling [5]–[9].

Our reported method [5]–[9] is applicable only to a front-view face. However, in daily conversations, we often change our face direction, and therefore we need to select a frame of the front-view face in a dynamic image. In this paper, we propose a pre-processing module that has a judgment function to discriminate a front-view face. Using the module, a frame of the front-view face in a dynamic image is selected by estimating the face direction. Consequently, by using the proposed function for choosing the front-view face, facial expressions are discriminable even when a person changes his or her face direction freely.

II. IMAGE ACQUISITION

The principle behind thermal image generation is the Stefan–Boltzmann law, expressed as $W = \varepsilon\sigma T^4$, where ε is emissivity, σ is the Stefan–Boltzmann constant (=5.6705×10^{-12} W/cm^{2}K^4), and T is the temperature (K). For human skin, ε is estimated as 0.98 to 0.99 [10], [11]. In this study, the approximate value of 1 was used as ε for human skin because the value of ε for almost all substances is lower than that of human skin [10]. Consequently, the human face region is easily extracted from an image by using the value of 1 for ε when a range of skin temperatures is selected to produce a thermal image [1]–[9], [12]. Fig. 1 shows examples of face images of a male, obtained by VR and IR. We can obtain a thermal image of the face without light, even at night. In principle, the temperature measurement by IR does not depend on skin color [11], darkness, or lighting condition, and so the face region and its characteristics are easily extracted from a thermal image.

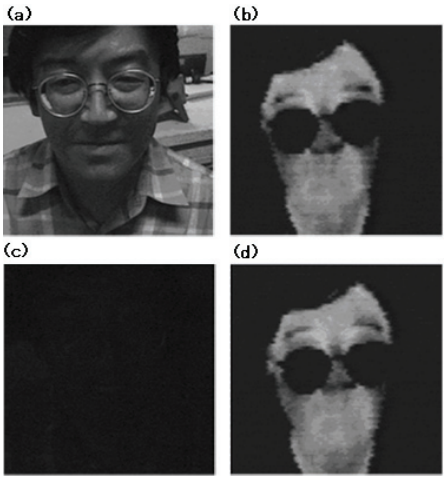

Fig. 1. Examples of a face image at night:
(a) VR with lighting, (b) IR with lighting, (c) VR without lighting, (d) IR without lighting [3]

III. PROPOSED METHOD

In this study, as a pre-processing module, we add a judgment function [12] of a front-view face to our reported method for facial expression recognition.

Fig. 2 illustrates the flow chart of our method. We have two modules in our system. The first is a module for speech recognition and dynamic image analysis, and the second is a module for learning and recognition. In the module for learning and recognition, we embed the module for front-view face judgment. The procedure including the proposed pre-processing module for front-view face judgment is explained in the following.

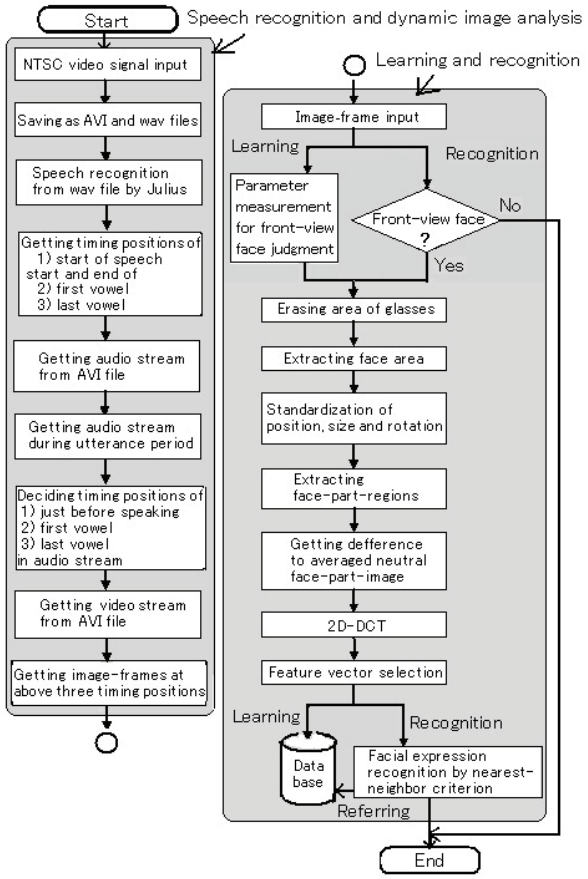

Fig. 2. Flow chart of our method

1. Front-view Face Judgment

We define the face rotation around the *X*, *Y*, and *Z* axes, as demonstrated in Fig. 3. The training data is calculated from the images of a front-view face. After normalizing the horizontal Feret's diameter of the segmented face image, the centerline of the face region in the limited region is drawn as the first standard line. Each pixel point in the centerline has the same number of pixels on both the left and the right sides of the point. Then, a straight line as the second standard line is drawn

from the upper to the lower edgepoints of the centerline. The two standard lines are used to estimate the deviation from the front-view face. To evaluate the face direction, we use four feature parameters (*nod*, *up*, *rotate*, *lean*), explained below.

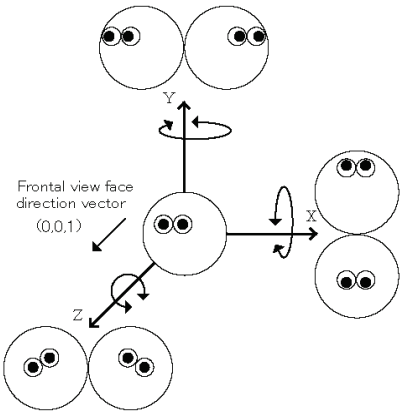

Fig. 3. The three-dimensional coordinate system

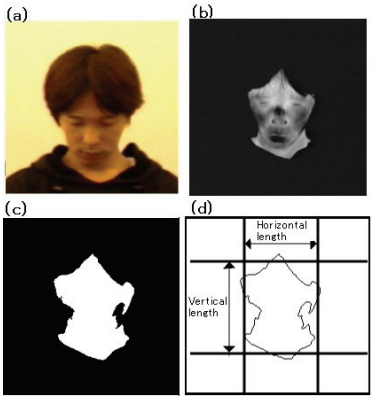

Fig. 4. Downward rotation; (a) visible image, (b) thermal image, (c) binary image, (d) schematic diagram for the parameter *nod*

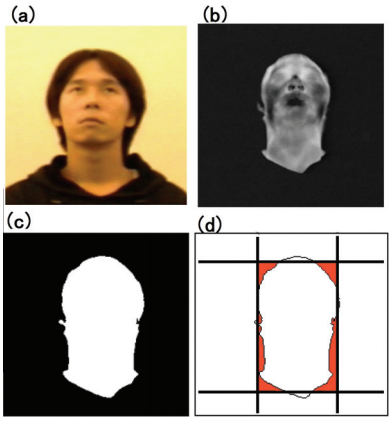

Fig. 5. Upward rotation; (a) visible image, (b) thermal image, (c) binary image, (d) schematic diagram for the parameter *up*

A. Parameter nod

The feature parameter *nod* is the value of the vertical Feret's diameter divided by the horizontal Feret's diameter on the segmented face image. This parameter is used to evaluate the downward rotation around the X axis (Fig. 4). Here, to eliminate the influence of the hair style and the throat region on the segmented face image as much as possible, the area having horizontal pixels less than an experimentally decided threshold is ignored in the measurement of the vertical Feret's diameter on the segmented face image, and the area having vertical pixels less than another experimentally decided threshold is also ignored in the measurement of the horizontal Feret's diameter on the segmented face image (Fig. 4). When a subject looks downward, the feature parameter *nod* tends to be smaller than that of the front-view face of the subject.

B. Parameter up

The feature parameter *up* is the area ratio of the region having a gray level of "0" in the rectangle defined by the straight lines determined by the measurements of the horizontal and vertical Feret's diameters to the rectangular area on the segmented face image. This parameter is used to evaluate the upward rotation around the X axis (Fig. 5). When a subject looks upward, the feature parameter *up* tends to be smaller than that of the front-view face of the subject.

C. Parameter rotate

The feature parameter *rotate* is the area of the region surrounded by the first and the second standard lines, on which are explained in the first paragraph in Section III-1 ("Front-view Face Judgment"). This parameter is used to evaluate the rotation around the Y axis (Fig. 6). Here, to eliminate the influence of the hair style and the throat region on the segmented face image as much as possible, the top and bottom of both the first and second standard lines are set inside the face region by each constant pixel, which is also decided experimentally (Fig. 6). The feature parameter *rotate* tends to increase when the rotation from the front-view face around the Y axis increases.

D. Parameter lean

The feature parameter *lean*, which is the angle between the second standard straight line and the horizontal line, is used to estimate the rotation around the Z axis. The value of *lean* is never greater than 90 degrees. The rotation is estimated to be small when the deviation of the value of *lean* from 90 degrees is small.

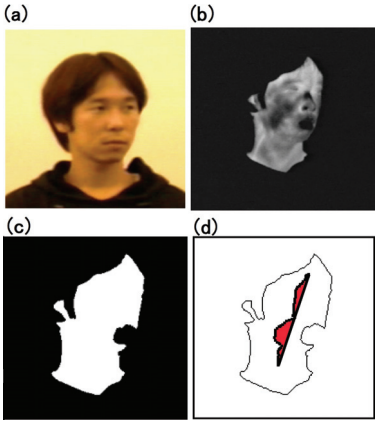

Fig. 6. Images expressing the face rotation around the *Y* axis; (a) visible image, (b) thermal image, (c) binary image, (d) schematic diagram for the parameter *rotate*

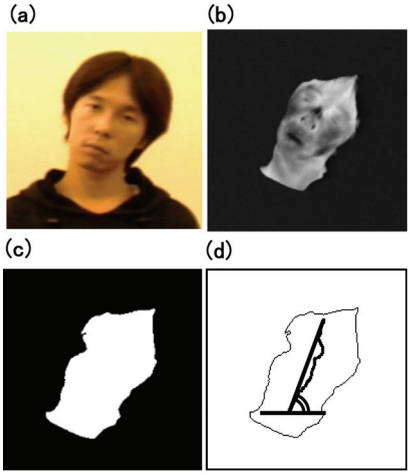

Fig. 7. Images expressing face rotation around the *Z* axis; (a) visible image, (b) thermal image, (c) binary image, (d) schematic diagram for the parameter *lean*

2. Speech Recognition and Dynamic Image Analysis

We use a speech recognition system named Julius [14] to obtain the timing positions of the start of speech, and the first and last vowels in a WAV file [8], [9]. Fig. 8 shows an example of the waveform of the Japanese name "Taro"; the timing position of the start of speech and the timing ranges of the first vowel (/a/) and the last vowel (/o/) were decided by Julius. By using the timing position of the start of speech and the timing ranges of the first and last vowels obtained from the WAV file, three image frames are extracted from an AVI file at the three timing positions. As the timing position just before speaking, we use the timing position of 84 ms before the start of speech, as determined in our previously reported study [7]. As the timing position of the first vowel, we use the position where the absolute value of the amplitude of the wave form is the maximum while speaking the vowel. For the timing position of the last vowel, we apply the same procedure used for the first vowel.

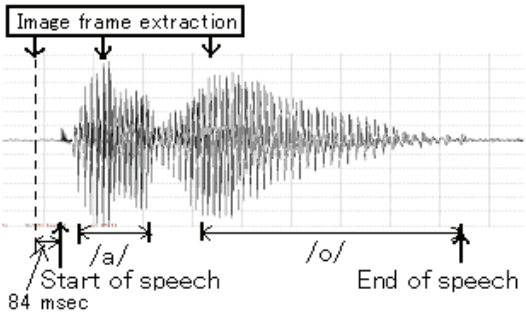

Fig. 8. Speech waveform of "Taro" and timing positions for image frame extraction [8]

3. Learning and Recognition

For the static images obtained from the extracted image frames, the process of erasing the area of the glasses, extracting the face area, and standardizing the position, size, and rotation of the face are performed according to the method described in our previously reported study [7]. In the next step, we generate difference images between the averaged neutral face image and the target face image in the extracted face areas in order to perform a 2D discrete cosine transform (2D-DCT). The feature vector is generated from the 2D-DCT coefficients according to a heuristic rule [6], [7].

As stated above, we use the speech recognition system named Julius. Julius sometimes makes a mistake in recognizing the first and/or last vowel(s). For example, /a/ for the first vowel might be misrecognized as /i/. For the training data, we correct the misrecognition. However, a correction cannot be made on the test data. The facial expression is recognized by the nearest-neighbor criterion in the feature vector space by using the training data just before speaking and just speaking the phonemes of the first and last vowels.

IV. EXPERIMENTS

1. Condition

The thermal image produced by the thermal video system (Nippon Avionics TVS-700) and the sound captured from an Electret condenser microphone (Sony ECM-23F5), amplified by a mixer (Audio-Technica AT-PMX5P), were transformed into a digital signal by an A/D converter (Thomson Canopus ADVC-300) and

input into a computer (DELL Optiplex 760, CPU: Intel Core 2 Duo E7400 2.80 GHz, main memory: 2.0 GB, and OS: Windows XP (Microsoft) with an IEEE1394 interface board (I·O Data Device 1394-PCI3/DV6). We used Visual C++ 6.0 (Microsoft) as the programming language. To generate a thermal image, we applied the condition that the thermal image had 256 gray levels for the range 5 to 12.9 K. Accordingly, one gray level corresponded to 1.95×10^{-2} to 5.04×10^{-2} K. The temperature range for generating a thermal image was decided for each subject in order to easily extract the face area on the image. We saved the visual and audio information in the computer as a Type 2 DV-AVI file, in which the frame had a spatial resolution of 720×480 pixels and 8-bit gray levels and the sound was saved in a PCM format of a stereo type, 48 kHz, and 16-bit levels.

Six subjects exhibited in alphabetic order each of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised," while speaking the semantically neutral utterance "Taro." Fig. 9 shows examples of the thermal image of each subject. Subjects A and B were males with glasses. Subjects C and D were females without glasses. Subject E was a male without glasses. Subject F was a male with a cap. Fig. 10 shows examples of the images of subject A.

In the experiment, all subjects kept intentionally front-view faces in the AVI files saved as the training data, and they freely changed their face direction in the AVI files saved as the test data. We obtained the feature parameter ranges for judging the front-view face with the use of the training data and the proposed method mentioned in Section III-1 ("Front-view Face Judgment"). The range was calculated as $a_i - \sigma_i < x_i < a_i + \sigma_i$ for the *nod*, *up*, and *rotate* parameters of subject A, and $a_i - 2\sigma_i < x_i < a_i + 2\sigma_i$ for the *lean* parameter of subject A and all feature parameters of other subjects, where x_1, x_2, x_3, x_4 were *nod*, *up, rotate,* and *lean*, and a_1, a_2, a_3, a_4 were their mean values and $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ were their standard deviation values, respectively. In the case of subject A, the difference of *nod*, *up*, and *rotate* between the front-view and the non-front-view faces was relatively small, and so the feature parameter ranges for judging the front-view face were set to be narrower than those of the other subjects. We assembled twenty samples as training data and ten or less samples as test data, in which all facial expressions of all subjects were judged as front-view faces by the proposed method. The number of test data was decided as a result of the front-view face judgment. From one sample, we obtained three images at the timing positions of just before speaking and just speaking the phonemes of the first and last vowels. We had the training data of only /a/ for the first vowel and only /o/ for the last vowel. If Julius misrecognized the vowel of the test sample, the corresponding image was not used for facial expression recognition. We had four cases on misrecognition for vowel(s); (1) no misrecognition for the first and last vowels, (2) misrecognition only for the first vowel, (3) misrecognition only for the last vowel, (4) misrecognition for both of the first and last vowels. We prepared feature vectors of the training data in each of the four cases.

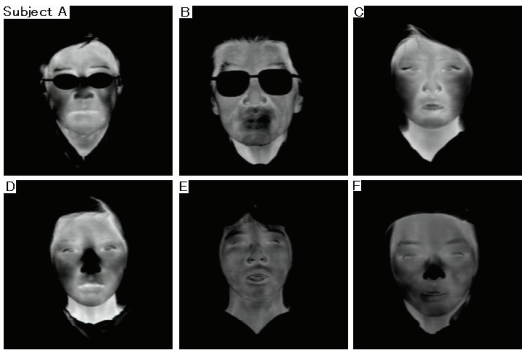

Fig. 9. Examples of thermal images having a neutral facial expression just before speaking

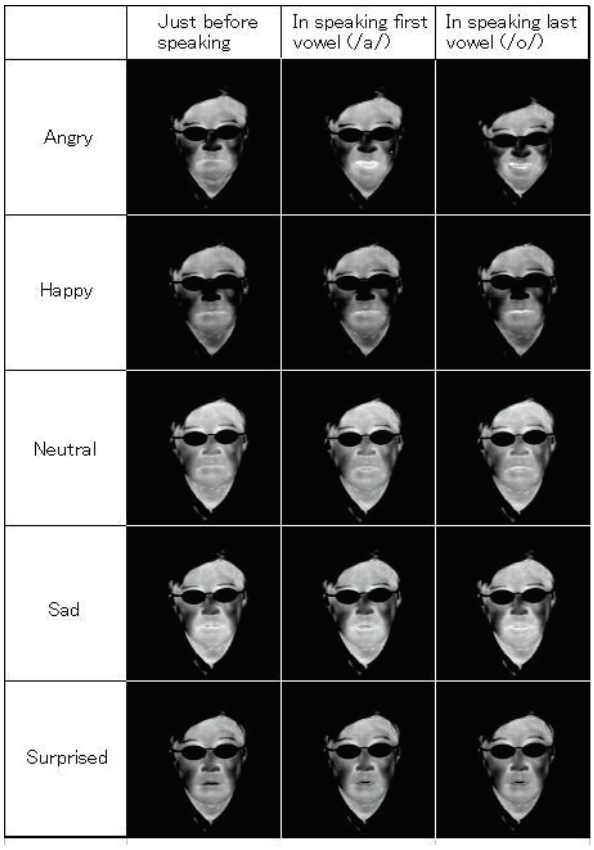

Fig. 10. Examples of thermal images of subject A having each facial expression in speaking

2. Results and Discussion

The mean value of the front-view face judgment accuracy for all subjects was 99.3%. The mean value of the recognition accuracy of vowels decided by Julius for all subjects was 93.6% for the first vowel (/a/) and 88.0% for the last vowel (/o/). Table 1 shows the facial expression recognition accuracy for all subjects. The mean recognition accuracy was 87.7%. The mean recognition accuracy for each subject was 88% for Subject A, 90% for Subject B, 88% for Subject C, 82% for Subject D, 84% for Subject E, and 94% for Subject F. We have applied the present method to a speaker of any utterance to prepare the training data for all combinations of the first and last vowels [15]. We expect the proposed method to be applicable for recognizing facial expressions in daily conversation.

Table 1. Recognition accuracy for all subjects

		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	**90**	1.7		10	3.3
	Happy	3.3	**93.3**			1.7
	Neutral			**80**	3.3	1.7
	Sad	5		5	**86.7**	5
	Surprised	1.7	5	15		**88.3**

V. CONCLUSION

We developed a method for facial expression recognition for a speaker by using thermal image processing and a speech recognition system. In this paper, we propose a pre-processing module that has a judgment function for discriminating the front-view face. Using the module, a frame of the front-view face in a dynamic image is selected by estimating the face direction. The mean judgment accuracy of the front-view face was 99.3% for six subjects, who changed their face direction freely. Using the proposed method, five kinds of facial expressions of six subjects were discriminable with 87.7% accuracy when the subject changed his or her face direction freely.

Acknowledgment

We would like to thank all the subjects who cooperated with us in the experiments. This work was supported by KAKENHI(22300077).

REFERENCES

[1] Yoshitomi Y, Kimura S, Hira E, and Tomita S (1996), Facial expression recognition using infrared rays image processing. Proc. of the Annual Convention IPS Japan, 2:339-340

[2] Yoshitomi Y, Kimura S, Hira E, and Tomita S (1997), Facial expression recognition using thermal image processing. IPSJ SIG Notes, CVIM103-3:17-24

[3] Yoshitomi Y, Miyawaki N, Tomita S, and Kimura S (1997), Facial expression recognition using thermal image processing and neural network. Proc. of 6th IEEE Int. Workshop on Robot and Human Communication 380-385

[4] Sugimoto Y, Yoshitomi Y, and Tomita S (2000), A method for detecting transitions of emotional states using a thermal face image based on a synthesis of facial expressions. J. Robotics and Autonomous Systems 31:147-160

[5] Yoshitomi Y, Kim SIll, Kawano T, and Kitazoe T (2000), Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. Proc. of 6th IEEE Int. Workshop on Robot and Human Interactive Communication 178-183

[6] Ikezoe F, Ko R, Tanijiri T, and Yoshitomi Y (2004), Facial expression recognition for speaker using thermal image processing (in Japanese). Trans. Human Interface Society 6(1):19-27

[7] Nakano M, Ikezoe F, Tabuse M, and Yoshitomi Y (2009), A study on the efficient facial expression using thermal face image in speaking and the influence of individual variations on its performance (in Japanese). J. IEEJ 38(2):156-163

[8] Koda Y, Yoshitomi Y, Nakano M, and Tabuse M (2009), Facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system. Proc. of 18th IEEE Int. Symp. on Robot and Human Interactive Communication 955-960

[9] Yoshitomi Y (2010), Facial expression recognition for speaker using thermal image processing and speech recognition system. Proc.of 10th WSEAS Int. Conf. on Applied Computer Science 182-186

[10]Kuno H (1994), Infrared rays engineering (in Japanese). Tokyo, IEICE 22

[11]Kuno H (1994), Infrared rays engineering (in Japanese). Tokyo, IEICE 45

[12]Yoshitomi Y, Tsuchiya A, and Tomita S (1998), Face recognition using dynamic thermal image processing. Proc. of 7th IEEE Int. Workshop on Robot and Human Communication 443-448

[13]Yamazaki S, Kamakura H, Tanijiri T, and Yoshitomi Y (2004), Three-dimensional CG expression of face rotation using fuzzy algorithm and thermal face image (in Japanese). Trans. Human Interface Society, 6(3):321-331

[14]http://julius.sourceforge.jp/

[15]Yoshitomi Y, Asada T, Shimada K, and Tabuse M (2011), Facial expression recognition of a speaker using vowel judgment and thermal image processing. Proc. of 16th Int. Symp. on Artificial Life and Robotics in press