The Sixteenth International Symposium on Artificial Life and Robotics 2011 (AROB 16th '11),
B-Con Plaza, Beppu,Oita, Japan, January 27-29, 2011

Facial Expression Recognition of a Speaker Using Vowel Judgment and Thermal Image Processing

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

[1]*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, t_asada@mei.kpu.ac.jp, tabuse@kpu.ac.jp
[2]*Nova System Co., Ltd., 3-3-3 Nakanoshima, Kita-ku, Osaka 530-0005, Japan*

Abstract: We previously developed a method for the facial expression recognition of a speaker. 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. By using the speech recognition system named Julius, thermal static images are saved at the timing positions of just before speaking, and just speaking the phonemes of the first and last vowels. To implement our method, three subjects, who spoke 25 Japanese first names providing all combinations of the first and last vowels, were used to prepare first the training data and then the test data. Julius sometimes makes a mistake in recognizing the first and/or last vowel(s). For example, /a/ for the first vowel is sometimes misrecognized as /i/. In the training data, we correct this misrecognition. However, the correction cannot be performed in the test data. In the implementation of our method, the facial expressions of three subjects were discriminable with the mean accuracy of 80.1% when the subjects exhibited one of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised," and the mean accuracy of the speech recognition of vowels by Julius was 71.0%.

Keywords: Facial expression recognition, 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 this research is to develop a robot that can perceive human feelings and mental states. For example, a robot could encourage a person who looks sad, advise an individual to stop working and rest when he or she looks tired, or take care of an elderly person.

The present investigation concerns the first stage of the development of a robot that has the ability to detect visually human feeling or inner mental states. Although the mechanism for recognizing facial expressions has received considerable attention in the field of computer vision research [1]–[6], at the present stage, it still falls far short of human capability, especially from the viewpoint of robustness under widely varying lighting conditions. One of the reasons for this is that nuances of shade, reflection, and local darkness influence the accuracy of facial expression recognition through the inevitable change of gray levels.

To develop a robust method for facial expression recognition applicable under widely varied lighting conditions, we do not use a visible ray (VR) image. Instead, we used an image produced by infrared rays (IR), which show the thermal distribution of the face [7]–[16]. Although a human cannot detect IR, it is possible for a robot to process the information in 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.

In addition, the timing of recognizing facial expressions is important for a robot because the processing can be time-consuming. We adopted an utterance as the key to expressing human feelings or mental states because humans tend to say something to express feeling [11]–[16].

In this paper, we briefly introduce our method [14] for the facial expression recognition of a speaker. For facial expression recognition, we select three images: (i) just before speaking, and speaking (ii) the first vowel and (iii) the last vowel in an utterance. A frame of the front-view face in a dynamic image is selected by estimating the face direction [16]. To apply our method [14], three subjects, who spoke 25 Japanese first names providing all combinations of the first and last vowels, were used to prepare first the training data and then the test data.

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 [17], [18]. 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 [17]. 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 [7]–[16], [19]. Fig. 1 shows examples of face images obtained by VR and IR of a male. Thermal images of a face can be obtained without light, that is, even at night. In principle, the temperature measurement by IR does not depend on skin color [18], darkness, or lighting condition, and so the face region and its characteristics are easily extracted from a thermal image.

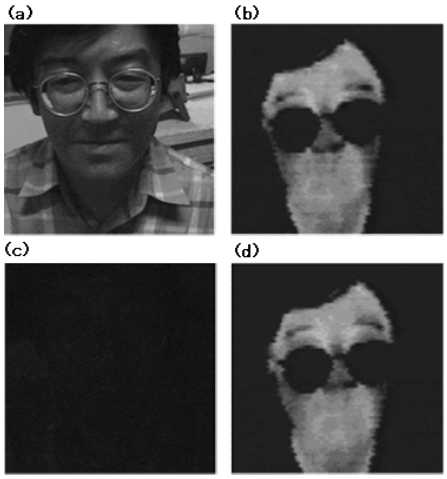

Fig. 1. Examples of a face image at night:
(a) VR with lighting, (b) IR with lighting, (c) VR without lighting, (d) IR without lighting [9]

III. PROPOSED METHOD

As a pre-processing module, we added a judgment function [19] of a front-view face to our method [14] for facial expression recognition [16]. Therefore, we can choose a front-view face as the target for recognizing facial expressions in daily conversation.

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

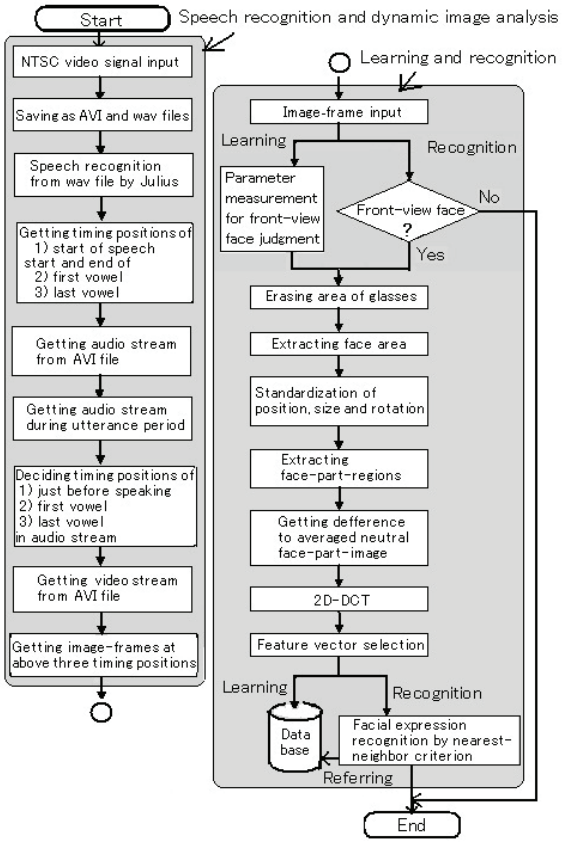

Fig. 2. Flow chart of our method [16]

the module for learning and recognition, we embedded the module for front-view face judgment. The procedure, except the pre-processing module for front-view face judgment [16], is explained in the following.

1. Speech Recognition and Dynamic Image Analysis

We use a speech recognition system named Julius [20] to save the timing positions of the start of speech, and the first and last vowels in a WAV file [14]–[16]. Fig. 3 shows an example of the waveform of the Japanese first 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 these three timing positions obtained from the WAV file, three thermal image frames are extracted from an AVI file. As the timing position just before speaking, we use 84 ms before the start of speech, as determined in our previously reported study [13]. 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.

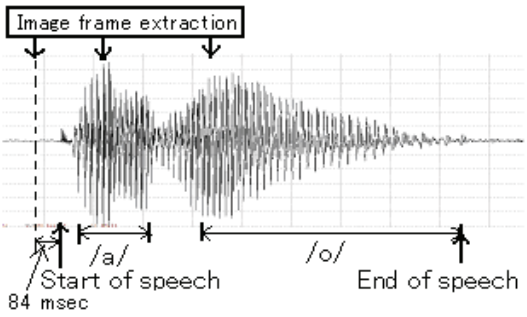

Fig. 3. Speech waveform of "Taro" and timing positions for image frame extraction [14]

2. Learning and Recognition

For the static thermal 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 [13]. Fig. 4 shows the blocks for extracting the face areas in a thermal image having 720 × 480 pixels. 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 [12], [13].

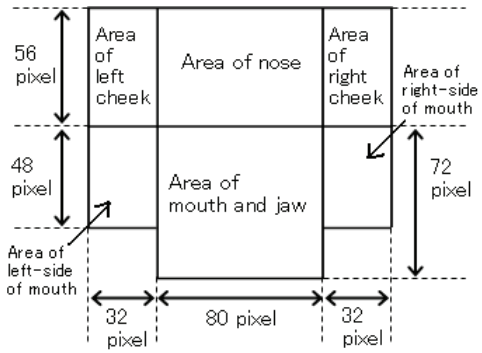

Fig. 4. Blocks for extracting face areas in the thermal image [14]

As stated above, we use the speech recognition system named Julius, which sometimes makes a mistake in recognizing the first and/or last vowel(s). For example, /a/ for the first vowel is sometimes misrecognized as /i/. For the training data, we correct the misrecognition. However, a correction cannot be made on the test data. For example, when Julius correctly judges the first vowel at the utterance of "Taro" but misjudges the last vowel as /a/, the training data in speaking "Ayaka" are used for recognition instead of those for speaking "Taro." The facial expression is recognized by the nearest-neighbor criterion in the feature vector space by using the training data just before speaking and while 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 GX620, CPU: Pentium IV 3.4 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 detected temperature range. This range was decided independently for each subject in order to best extract the face area. We saved the visual and audio information in the computer as a Type 2 DV-AVI file, in which a frame had a spatial resolution of 720 × 480 pixels and 8-bit gray levels, and the sound was saved in a PCM format of stereo type, 48 kHz, and 16-bit levels.

All subjects exhibited in alphabetic order each of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised," while speaking the semantically neutral utterance of each of the Japanese first names listed in Table 1. Fig. 5 shows examples of the thermal image of each subject. There were three subjects. Subject A was a male without glasses. Subject B was a male with glasses. Subject C was a female without glasses. Fig. 6 shows examples of the images of Subject A.

Table 1. Japanese first names used in the experiment

		First vowel				
		a	i	u	e	o
Last vowel	a	ayaka	shinnya	tsubasa	keita	tomoya
	i	kazuki	hikari	yuki	megumi	koji
	u	takeru	shigeru	fuyu	megu	noboru
	e	kaede	misae	yusuke	keisuke	kozue
	o	taro	hiroko	yuto	keiko	tomoko

In the experiment, all subjects kept intentionally front-view faces in the AVI files saved as both the training and test data. Accordingly, the pre-processing

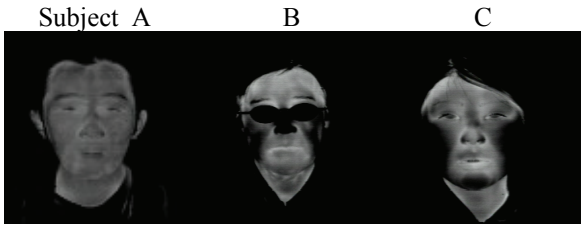

Fig. 5. Examples of thermal images having a neutral facial expression just before speaking "Ayaka"

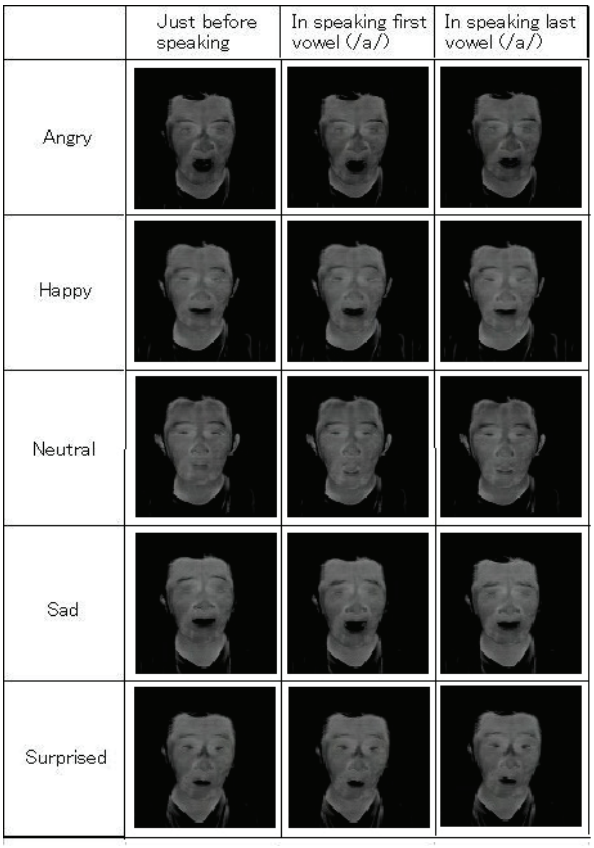

Fig. 6. Examples of thermal images of Subject A having each facial expression in speaking "Ayaka"

module for judging the front-view face [16] was not used in the experiment. We assembled 20 samples as training data and 10 samples as test data. 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 obtained training data for all combinations of vowel type of the first and last vowels.

2. Results and Discussion

The mean value of the recognition accuracy for the first vowels and last vowels of all subjects was 71.0%. The mean value of the recognition accuracy for the first vowels and last vowels was 84% for Subject A, 70% for Subject B, and 59% for Subject C. Table 2 shows the recognition accuracy for the first and last vowels of

Table 2. Accuracy (%) of speech recognition for Subject A

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	56	70	80	90	0	59
a-i	90	100	100	90	90	94
a-u	80	40	100	100	80	80
a-e	20	20	30	70	100	48
a-o	100	80	100	100	100	96
i-a	100	100	100	100	70	94
i-i	90	90	100	100	80	92
i-u	90	90	90	100	90	92
i-e	40	90	100	88	0	64
i-o	100	100	90	60	100	90
u-a	90	100	50	100	40	76
u-i	90	100	40	10	100	68
u-u	100	100	90	100	100	98
u-e	100	100	100	100	100	100
u-o	100	100	90	100	100	98
e-a	63	50	60	40	0	43
e-i	100	100	100	100	100	100
e-u	40	60	100	100	90	78
e-e	22	90	80	90	50	66
e-o	90	100	100	100	100	98
o-a	100	90	100	100	50	88
o-i	100	100	100	100	100	100
o-u	100	100	100	100	70	94
o-e	90	100	100	100	80	94
o-o	100	100	100	100	100	100
Mean	82	87	88	90	76	84

Table 3. Accuracy (%) of facial expression recognition

Subject A		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	90.8	0.4	0.0	4.8	3.6
	Happy	2.8	94.8	2.4	4.4	8.0
	Neutral	0.0	0.0	94	0.0	0.0
	Sad	3.2	3.2	0.8	79.2	15.2
	Surprised	1.6	1.6	2.8	11.6	73.2

Subject B		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	64.1	5.2	5.2	6.7	9.4
	Happy	10.8	77.6	0.8	6.9	12.6
	Neutral	0.0	3.6	83.6	0.4	0.0
	Sad	7.2	7.6	2.0	78	10.9
	Surprised	17.9	6	8.4	8	67.3

Subject C		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprised
Output	Angry	78.8	4.4	7.2	5.2	0.8
	Happy	2.4	77.1	3.6	2.4	0.0
	Neutral	10.8	2.0	71.9	3.2	2.0
	Sad	2.8	11.3	12.9	78.8	5.2
	Surprised	5.2	5.2	4.4	10.4	92

Table 4. Accuracy (%) of facial expression recognition of Subject A

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	100	100	100	100	0	80.0
a-i	100	100	70	100	90	92.0
a-u	90	100	100	100	90	96.0
a-e	100	100	100	100	100	100.0
a-o	100	100	100	60	90	90.0
i-a	60	80	100	60	60	72.0
i-i	100	100	70	100	100	94.0
i-u	100	100	100	100	40	88.0
i-e	90	100	100	90	90	94.0
i-o	90	100	90	90	100	94.0
u-a	90	100	60	40	70	72.0
u-i	100	100	100	0	100	80.0
u-u	100	100	90	90	100	96.0
u-e	100	100	100	100	100	100.0
u-o	100	100	90	100	90	96.0
e-a	100	80	100	50	10	68.0
e-i	80	100	100	70	100	90.0
e-u	100	80	100	90	100	94.0
e-e	100	90	80	80	0	70.0
e-o	10	100	100	50	30	58.0
o-a	100	60	100	100	10	74.0
o-i	100	90	100	90	100	96.0
o-u	100	100	100	30	90	84.0
o-e	90	100	100	90	90	94.0
o-o	70	90	100	100	80	88.0
Mean	90.8	94.8	94.0	79.2	73.2	86.4

Table 5. Accuracy (%) of facial expression recognition of Subject B

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	10	70	10	90	90	54.0
a-i	50	50	100	40	44	56.8
a-u	80	90	100	89	89	89.6
a-e	22	100	100	78	80	76.0
a-o	30	100	100	60	50	68.0
i-a	100	30	90	70	80	74.0
i-i	40	80	100	90	10	64.0
i-u	70	100	100	100	100	94.0
i-e	90	100	90	90	78	89.6
i-o	60	100	100	90	80	86.0
u-a	60	10	70	90	30	52.0
u-i	0	90	90	40	70	58.0
u-u	100	40	40	90	70	68.0
u-e	100	100	100	83	100	96.6
u-o	80	40	100	80	90	78.0
e-a	90	90	90	90	90	90.0
e-i	90	90	100	90	0	74.0
e-u	10	80	100	90	20	60.0
e-e	100	100	10	100	70	76.0
e-o	90	100	100	100	80	94.0
o-a	50	90	100	70	90	80.0
o-i	50	60	100	10	50	54.0
o-u	50	80	100	80	80	78.0
o-e	80	90	0	100	70	68.0
o-o	100	60	100	40	70	74.0
Mean	64.1	77.6	83.6	78.0	67.2	74.1

Subject A. Table 3 shows the facial expression recognition accuracy as mean values over all combinations of first and last vowels. The mean recognition accuracy of the facial expressions of all subjects was 80.1%. The mean recognition accuracy of the facial expressions was 86.4% for Subject A, 74.1% for Subject B, and 79.7% for Subject C. Tables 4 to 6 show all the values of the recognition accuracy of the facial expressions of each subject. We have not yet found a relationship between the recognition accuracy of vowels and the recognition accuracy of facial expressions. As a continuation of our work, we will apply our method for facial expression recognition by using other subjects.

V. CONCLUSION

We previously developed a method for facial expression recognition for a speaker by using thermal image processing and a speech recognition system. To implement our method, three subjects spoke 25 Japanese first names, which provided all combinations of the first and last vowels. These subjects were used to prepare first the training data and then the test data for all combinations of the first and last vowels. The mean accuracy of the recognition of vowels by Julius was 71.0% for all subjects. Using our method, the facial expressions of three subjects were discriminable with 80.1% accuracy when he or she exhibited one of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised." We expect our method to be applicable for recognizing facial expressions in daily conversation.

Acknowledgment

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

Table 6. Accuracy (%) of facial expression recognition of Subject C

First-last vowel	Angry	Happy	Neutral	Sad	Surprised	Mean
a-a	100	67	30	40	80	63.4
a-i	80	50	100	30	60	64.0
a-u	70	100	80	70	100	84.0
a-e	60	70	10	60	60	52.0
a-o	100	50	78	60	100	77.6
i-a	80	100	100	100	90	94.0
i-i	80	100	100	60	90	86.0
i-u	40	70	90	60	60	64.0
i-e	70	40	70	30	100	62.0
i-o	60	80	90	80	90	80.0
u-a	80	100	90	90	90	90.0
u-i	90	10	50	90	100	68.0
u-u	60	90	90	100	100	88.0
u-e	80	100	100	80	100	92.0
u-o	90	10	10	90	100	60.0
e-a	100	100	70	80	100	90.0
e-i	90	100	100	100	100	98.0
e-u	100	80	40	60	100	76.0
e-e	90	100	100	100	100	98.0
e-o	90	100	30	100	100	84.0
o-a	40	80	80	100	100	80.0
o-i	90	60	80	90	80	80.0
o-u	70	90	20	100	100	76.0
o-e	90	100	100	100	100	98.0
o-o	70	80	90	100	100	88.0
Mean	78.8	77.1	71.9	78.8	92.0	79.7

REFERENCES

[1] Yuille AL, Cohen DS, and Hallinan PW (1989), Feature extraction from faces using deformable templates. Proc. of IEEE Conf. on Computer Vision and Pattern Recognition 104-109

[2] Harashima H, Choi CS, and Takebe T (1989), 3-D model-based synthesis of facial expressions and shape deformation (in Japanese). Human Interface 4:157-166

[3] Mase K (1990), An application of optical flow – extraction of facial expression. IAPR Workshop on Machine Vision and Application 195-198

[4] Mase K (1991), Recognition of facial expression from optical flow. Trans. IEICE E74(10):3474-3483

[5] Matsuno K, Lee C, and Tsuji S (1994), Recognition of facial expressions using potential net and KL expansion (in Japanese). Trans. IEICE J77-D-II(8):1591-1600

[6] Kobayashi H and Hara F (1994), Analysis of neural network recognition characteristics of 6 basic facial expressions. Proc. of 3rd IEEE Int. Workshop on Robot and Human Communication 222-227

[7] 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

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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

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

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

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

[19]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

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