The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th '13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

Facial expression recognition of a speaker using thermal image processing and reject criteria in feature vector space

Y. Nakanishi, Y. Yoshitomi, T. Asada, and M. Tabuse

Graduate School of Life and Environmental Sciences Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: y_nakanishi@mei.kpu.ac.jp, yoshitomi@kpu.ac.jp, t_asada@mei.kpu.ac.jp, tabuse@kpu.ac.jp

Abstract: In our previously developed method for the facial expression recognition of a speaker, the positions of feature vectors in the feature vector space in image processing were generated with imperfections. The imperfections, which caused misrecognition of the facial expression, tended to be far from the center of gravity of the class to which the feature vectors belonged. In the present study, to omit the feature vectors generated with imperfections, a method using reject criteria in the feature vector space was applied to facial expression recognition. By using the proposed method, the facial expressions of two subjects were discriminable with 90.0% accuracy for the three facial expressions of "happy," "neutral," and "others" when they exhibited one of the five intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised," whereas these expressions were discriminable with 78.0% accuracy by the conventional method.

Keywords: Facial expression recognition, Feature vector space, Reject criteria, Speech recognition, Thermal image processing.

1 INTRODUCTION

Although the mechanism for recognizing facial expressions has received considerable attention in the field of computer vision research, it still falls far short of human capability, especially from the viewpoint of robustness under widely varying lighting conditions. For example, nuances of shade, reflection, and local darkness influence the accuracy of facial expression recognition through the inevitable change of gray levels. To develop a method for facial expression recognition applicable under widely varied lighting conditions, we used images produced by infrared rays (IR), which show the thermal distribution of the face [1]-[13].

In experiments, imperfections in image processing in our method for facial expression recognition inevitably occurred and resulted in misrecognition of facial expressions. To overcome this difficulty, in the present study, a method using reject criteria in the feature vector space is applied to the recognition of facial expression of two male subjects when exhibiting the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised."

2 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 [14], [15]. In the present 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 [14]. Consequently, the human face region is easily extracted from an image by using the value of 1 for ε [1]-[13]. In principle, temperature measurements by IR also do not depend on skin color [15], darkness, or lighting condition, and so the face region and its characteristics are easily extracted from a thermal image.

3 PROPOSED METHOD

Fig. 1 illustrates the flowchart 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 embedded the module for front-view face judgment [10]. Some details of our method are explained in our book [16].

3.1 Speech recognition and dynamic image analysis

We use a speech recognition system named Julius [17] to obtain the timing positions of the start of speech and the first and last vowels in a WAV file [8]-[10]. Fig. 2 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/) are 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. For the timing position just before speaking, we use the timing position of

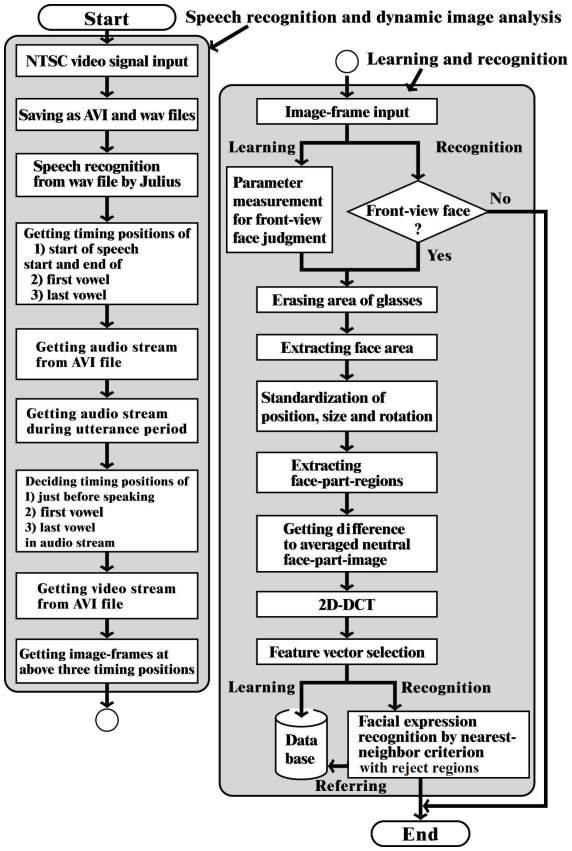

Fig. 1. Flowchart of our method

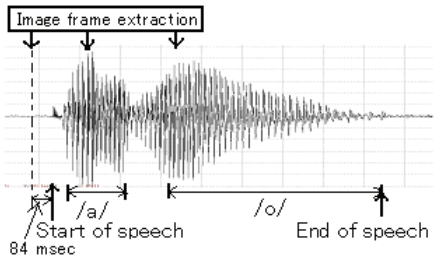

Fig. 2. Speech waveform of "Taro" and timing positions for image frame extraction [8]

84 ms before the start of speech, as determined in our previously reported study [7]. For the timing position of the first vowel, we use the position where the absolute value of the amplitude of the waveform 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.

3.2 Learning and recognition

For 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], [16]. Fig. 3 shows the blocks for extracting the face areas in a thermal image having 720×480 pixels. In the next step, we

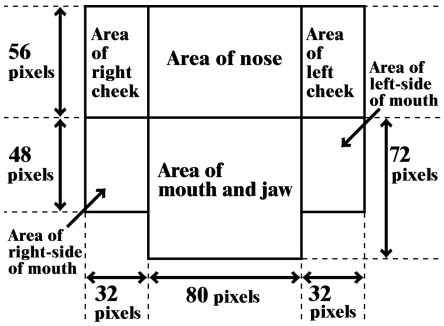

Fig. 3. Blocks for extracting face areas in the thermal image [18]

generate difference images between the averaged neutral face image and the target face image in the extracted face areas 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]. We refer to the method used in our previously reported research [8]-[13], [16], [18], [19] as the conventional method, where the facial expression is recognized by the nearest-neighbor criterion in the feature vector space by using the training data just before speaking and when speaking the phonemes of the first and last vowels.

3.3 Proposed method for facial expression recognition

Imperfections in image processing in our method for facial expression recognition inevitably occurred and resulted in misrecognition of facial expressions. To overcome this difficulty, in the present study, a method using reject criteria in the feature vector space of the training data just before speaking and when speaking the phonemes of the first and last vowels is proposed.

The new algorithm of pattern recognition in the feature vector space of facial expression recognition is as follows:

***Step 1*:** For all classes in the initial condition, the following procedure is performed. Vector $\boldsymbol{g}_i$ of the center of gravity of the class $i\,(i=1,2,\cdots,n)$, where n is the number of classes, is obtained. Then, go to Step 2.

***Step 2*:** For each class i, the following procedure is performed. The elements of class i are narrowed using the way that the feature vector having the largest distance from vector $\boldsymbol{g}_i$ among those in class i is omitted from the elements in class i. Then, if the number of feature vectors of class i at that moment is bigger than half the number N_i of feature vectors in the class defined in Step 1, go to Step 3. Otherwise, go to Step 4.

***Step 3*:** For class i, vector $\boldsymbol{g}_i$ is updated. Then, return to Step 2.

***Step 4*:** Except the class of neutral facial expression, the threshold T_j of class j is decided as (a) the value of the

maximum distance between vector $\boldsymbol{g}_j$ and the feature vector of class j in the initial condition, under the constraint that vector $\boldsymbol{g}_j$ among all vectors $\boldsymbol{g}_i$ $(i=1,2,\cdots,n)$ is the nearest to the feature vector. Similarly, for the class of neutral facial expression, the threshold T_j of class j is decided as the mean value of (a) and (b), where (b) is the minimum value among the values of the distance between vector $\boldsymbol{g}_j$ and the feature vector of class j in the initial condition, under the constraint that vector $\boldsymbol{g}_j$ among all vectors $\boldsymbol{g}_i$ $(i=1,2,\cdots,n)$ is not the nearest to the feature vector. Go to Step 5.

***Step 5*:** The elements of each class i are narrowed using the constraint that the distance between the feature vectors and vector $\boldsymbol{g}_i$ should not be bigger than threshold T_i. Go to Step 6.

***Step 6*:** The facial expression is recognized by the nearest-neighbor criterion in the feature vector space using (1) the feature vectors selected by the procedure up to Step 5, (2) the vector $\boldsymbol{g}_i$. Then, the recognized result showing class i is rejected if the distance between the test feature vector and the selected vector by the vector nearest-neighbor criterion is bigger than the threshold T_i. Otherwise the recognized result showing class i is accepted.

In the present study, the Euclidean distance is used as the distance in the feature vector space. The class of neutral facial expression receives special treatment because in many cases the value of (a) tends to be too small to accept the recognized result showing class j.

4 EXPERIMENTS

4.1 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), as 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 780, CPU: Intel Core 2 Duo E8400 3.00 GHz, main memory: 3.21 GB, and OS: Windows 7 Professional (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 set the condition that the thermal image had 256 gray levels for the detected temperature range. As a result, 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 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 video frame had a spatial resolution of 720×480 pixels and 8-bit gray levels, and the sound was saved in a stereo PCM format, 48 kHz and 16-bit levels.

Two subjects exhibited in alphabetic order each of the five intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised," while speaking the semantically neutral utterance "Taro." In the present study, we categorized all of the "angry," "sad," and "surprise" expressions as "others" in making the recognition results. Subjects A and B were males with glasses. Figs. 4 and 5 show examples of thermal images of the neutral expression captured in each period for Subjects A and B, respectively. We captured the thermal images for subject A on June 20, 2006 (hereinafter referred to as First_period_A) (Fig. 6); at approximately 3:00 pm on January 22, 2010 (hereinafter referred to as Second_period_A); and at approximately 5:00 pm on January 22, 2010 (hereinafter referred to as Third_period_A). For subject B, we captured the same thermal images on June 20, 2006 (hereinafter referred to as First_period_B) (Fig. 7); and January 25, 2010 (hereinafter referred to as Second_period_B). In the present study, we investigated several combinations of training and test data for each subject, as listed in Tables 1, 2, and 3 in Section 4.2. Case-A-1-2 for subject A consisted of the following cases: (1) thermal images for both training and test data captured on First_period_A, (2) thermal images for both

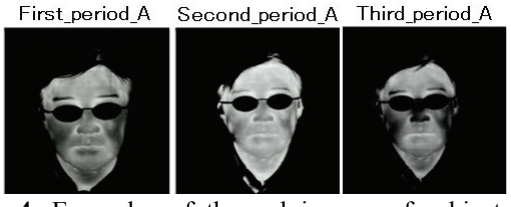

Fig. 4. Examples of thermal images of subject A having a neutral facial expression captured just before speaking at each period

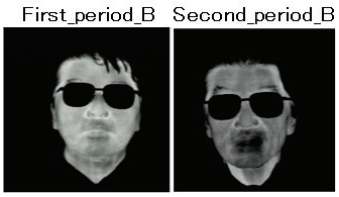

Fig. 5. Examples of thermal images of subject B having a neutral facial expression captured just before speaking at each period

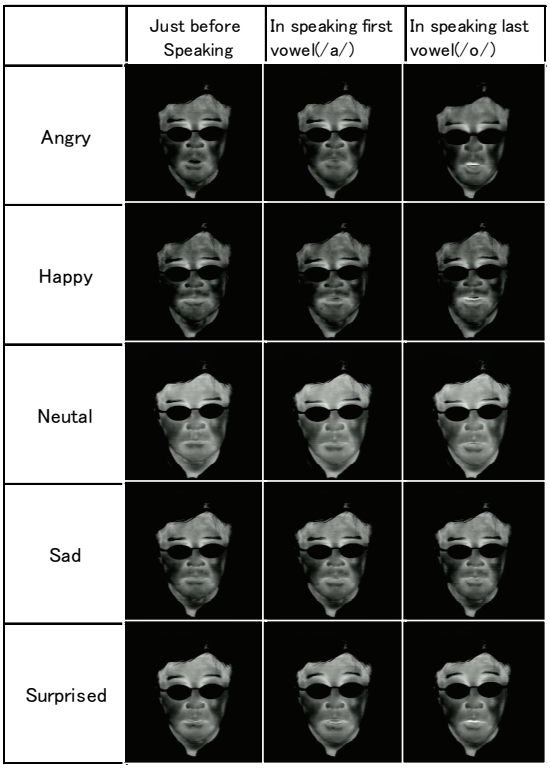

Fig. 6. Examples of thermal images of subject A captured on First_period_A and having each facial expression [12]

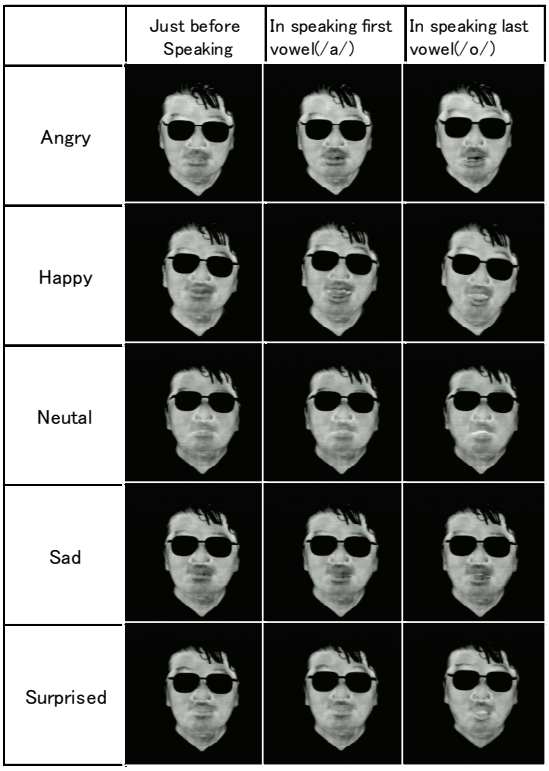

Fig. 7. Examples of thermal images of subject B captured on First_period_B and having each facial expression [12]

training and test data captured on Second_period_A, (3) thermal images for training and test data captured on First_period_A and Second_period_A, respectively, and (4) thermal images of training data of the facial expressions of "happy" and "neutral" captured on Second_period_A, thermal images of training data of the facial expressions of "others" captured on First_period_A, and the test data were captured on Second_period_A. In addition, Case-A-2-3 for subject A consisted of four cases decided by changing the data of First_period_A and Second_period_A in Case-A-1-2 to those of Second_period_A and Third_period_A, respectively. Case-B-1-2 for subject B consisted of the four cases defined in the same manner described above as Case-A-1-2 for subject A.

For Case-A-1-2, Case-A-2-3, and Case-B-1-2, we compared (3) and (4) for investigating the effects of updating the training data of the facial expressions of "happy" and "neutral", whereas (1) and (2) were added for reference of thermal images for the training and the test data captured without intentional intervals.

Subjects A and B freely changed their face direction during the capturing of thermal images for the test data on Second_period_A and Second_period_B, respectively. We assembled twenty samples as training data and ten samples as test data for all five facial expressions in all cases of Case-A-1-2, Case-A-2-3, and Case-B-1-2. All facial expressions of the test data for both subjects were judged as front-view faces by the method mentioned in our previously reported paper [10]. As one sample, we obtained one image each at the timing positions of just before speaking and while speaking the phonemes of the first and last vowels. If Julius misrecognized the vowel of the test sample, the corresponding image was not used for facial expression recognition. We had four cases of misrecognition for the vowel(s): (i) no misrecognition for the first and last vowels, (ii) misrecognition only for the first vowel, (iii) misrecognition only for the last vowel, and (iv) misrecognition for both the first and last vowels. We prepared feature vectors of the training data in each of the four cases.

4.2 Results and discussion

Tables 1, 2, and 3 show the facial expression recognition accuracy for the case of Case-A-1-2, Case-A-2-3, and Case-B-1-2, respectively. The total facial recognition accuracy by the conventional method was (429/550=78.0%), whereas that by the proposed method was (376/418=90.0%). Because the time period_ subject of Table 1 (2) is the same as that of Table 2 (1), the data in Table 2 (1) are not taken

Table 1. Recognition accuracy of Case-A-1-2

(1) Training and test data: First_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**		7/30	**10/10**		7/29
	Neutral		**8/10**			**8/10**	
	Others		2/10	**23/30**		2/10	**22/29**
Rejected					0/10	0/10	1/30
Accuracy		**41/50**			**40/49**		

(2) Training and test data: Second_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**9/10**			**7/8**		
	Neutral		**9/10**			**9/10**	
	Others	1/10	1/10	**30/30**	1/8	1/10	**28/28**
Rejected					2/10	0/10	2/30
Accuracy		**48/50**			**44/46**		

(3) Training data: First_period_A,
Test data: Second_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**4/10**	1/10	19/30	**4/10**	1/7	18/29
	Neutral		**9/10**			**6/7**	
	Others	6/10		**11/30**	6/10		**11/29**
Rejected					0/10	3/10	1/30
Accuracy		**24/50**			**21/46**		

(4) Training data: "happy" and "neutral"; Second_period_A, "others"; First_period_A,
Test data: Second_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**		2/30	**7/7**		3/25
	Neutral		**10/10**			**10/10**	
	Others			**28/30**			**22/25**
Rejected					3/10	0/10	5/30
Accuracy		**48/50**			**39/42**		

into account in calculating the total recognition accuracy. The proposed method improved the total facial recognition accuracy by 12.0%. The reject criteria explained in Section 3.3 were effective for improving facial recognition accuracy, as shown in Table 3 (3) in particular. The proposed method effectively judged whether the training data were acceptable for facial expression recognition at the moment; as shown in Table 3 (3), the training data should be updated.

The total facial recognition accuracy by the proposed method under condition (3) in Tables 1 and 2 was (43/71=60.6%), whereas that under condition (4) in Tables 1 and 2 was (81/84=96.4%), which was better than (134/145=92.4%), obtained under the conditions of Table 1 (1) and (2) and Table 2 (2). The total facial recognition accuracy by the conventional method under condition (3) in Tables 1, 2, and 3 was (76/150=50.7%), whereas that under condition (4) in Tables 1, 2, and 3 was (119/150=79.3%), which was lower than (234/250=93.6%), obtained under the conditions of Table 1 (1) and (2), Table 2 (2), and Table 3 (1) and (2). In both the proposed method and the conventional method, updating the training data for the facial expressions of "happy" and "neutral" remarkably improved the accuracy of facial expression recognition, as demonstrated in our previously reported paper [12].

Table 2. Recognition accuracy of Case-A-2-3

(1) Training and test data: Second_period_A
The same as Table 1 (2)

(2) Training and test data: Third_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**			**10/10**		
	Neutral		**10/10**			**10/10**	
	Others			**30/30**			**30/30**
Rejected					0/10	0/10	0/10
Accuracy		**50/50**			**50/50**		

(3) Training data: Second_period_A,
Test data: Third_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**0/10**			**0/3**		
	Neutral		**10/10**			**10/10**	
	Others	10/10		**30/30**	3/3		**12/12**
Rejected					7/10	0/10	18/30
Accuracy		**40/50**			**22/25**		

(4) Training data: "happy" and "neutral"; Third_period_A, "others"; Second_period_A,
Test data: Third_period_A

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**			**7/7**		
	Neutral		**10/10**			**10/10**	
	Others			**30/30**			**25/25**
Rejected					3/10		5/30
Accuracy		**50/50**			**42/42**		

Table 3. Recognition accuracy of Case-B-1-2

(1) Training and test data: First_period_B

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**			**10/10**		
	Neutral		**10/10**			**10/10**	
	Others			**30/30**			**30/30**
Rejected					0/10	0/10	0/30
Accuracy		**50/50**			**50/50**		

(2) Training and test data: Second_period_B

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**		5/30	**1/1**		
	Neutral		**10/10**			**10/10**	
	Others			**25/30**			**24/24**
Rejected					9/10	0/10	6/30
Accuracy		**45/50**			**35/35**		

(3) Training data: First_period_B,
Test data: Second_period_B

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**	8/10	30/30			
	Neutral		**2/10**				
	Others			**0/30**			
Rejected					10/10	10/10	30/30
Accuracy		**12/50**			**(All were rejected.)**		

(4) Training data: "happy" and "neutral"; Second_period_B, "others"; First_period_B,
Test data: Second_period_B

		Input facial expression					
		Conventional			Proposed method		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	**10/10**		15/30	**8/8**		13/15
	Neutral		**10/10**	14/30		**10/10**	
	Others			**1/30**			**2/15**
Rejected					2/10	0/10	15/30
Accuracy		**21/50**			**33/33**		

Though several studies on facial expression recognition using thermal image processing have been reported (see references [1]-[13], [16], [18]-[20]), only our research [5]-[13], [16], [18], [19] has focused on a speaker. In this paper,

we propose a method for using reject criteria in the feature vector space. The advantage of our method over the conventional method is demonstrated in Tables 1 to 3.

After preparing the training data for all combinations of the first and last vowels, we could apply our method to a speaker for any utterance [11], [16].

5 CONCLUSION

We previously developed a method for the facial expression recognition of a speaker. In the present study, a method using reject criteria in the feature vector space was applied. With the proposed method, the facial expressions of two male subjects were discriminable with 90.0% accuracy for the facial expressions of "happy," "neutral," and "others" when they exhibited one of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised." The proposed method effectively judged whether the training data were acceptable for the facial expression recognition at that moment.

Acknowledgment

This work was supported by KAKENHI (22300077).

REFERENCES

[1] Yoshitomi Y, Kimura S, Hira E, et al (1996), Facial expression recognition using infrared rays image processing. Proceedings of the Annual Convention IPS Japan, Osaka, Japan, Sep 4-6, 1996, 2:339-340

[2] Yoshitomi Y, Kimura S, Hira E, et al (1997), Facial expression recognition using thermal image processing. IPSJ SIG Notes, CVIM103-3, Kyoto, Japan, Jan 23-24, 1997, pp. 17-24

[3] Yoshitomi Y, Miyawaki N, Tomita S, et al (1997), Facial expression recognition using thermal image processing and neural network. Proceedings of 6th IEEE International Workshop on Robot and Human Communication, Sendai, Japan, Sep 29-Oct 1, 1997, pp. 380-385

[4] Sugimoto Y, Yoshitomi Y, 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, et al (2000), Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. Proceedings of 6th IEEE International Workshop on Robot and Human Interactive Communication, Osaka, Japan, Sep 27-29, 2000, pp. 178-183

[6] Ikezoe F, Ko R, Tanijiri T, et al (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, et al (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, et al (2009), Facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system. Proceedings of 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, Sep 29-Oct 1, 2009, pp. 955-960

[9] Yoshitomi Y (2010), Facial expression recognition for speaker using thermal image processing and speech recognition system. Proceedings of 10th WSEAS International Conference on Applied Computer Science, Appi Kogen, Iwate, Japan, Oct 4-6, 2010, pp. 182-186

[10] Fujimura T, Yoshitomi Y, Asada T, et al (2011), Facial expression recognition of a speaker using front-view face judgment, vowel judgment, and thermal image processing. J. Artificial Life and Robotics 16(3):411-417

[11] Yoshitomi Y, Asada T, Shimada K, et al (2011), Facial expression recognition of a speaker using vowel judgment and thermal image processing. J. Artificial Life and Robotics 16(3):318–323

[12] Nakanishi Y, Yoshitomi Y, Asada T, et al (2012), Robust facial expression recognition of a speaker using thermal image processing and updating of fundamental training-data. J. Artificial Life and Robotics, 17(2):263-269

[13] Yoshitomi Y (2012), Facial expression recognition of speaker using vowel judgment and features of thermal face image", Proceedings of 1st WSEAS International Conference on Information Technology and Computer Networks, Vienna, Austria, Nov 10-12, 2012, pp.139-145

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

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

[16] Yoshitomi Y, Tabuse M, Asada T (2012), Image processing: methods, applications and challenges. Nova Science Publisher, pp.57-85

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

[18] Asada T, Yoshitomi Y, Tabuse M (2012), A system for facial expression recognition of a speaker using front‐view face judgment, vowel judgment and thermal image processing. J. Artificial Life and Robotics, 17, in press.

[19] Yoshitomi Y, Tabuse M, Asada T (2011), Speech Technologies. InTech, pp.405-424

[20] Hernández B, Olague G, Hammoud R, et al (2007), Visual learning of texture descriptors for facial expression recognition in thermal imagery. Computer Vision and Image Understanding 16 (2-3): 258–269