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

A system for facial expression recognition of a speaker using thermal image processing and feature vector space characteristics

T. Asada, Y. Yoshitomi, Y. Nakanishi, 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: t_asada@mei.kpu.ac.jp, yoshitomi@kpu.ac.jp, y_nakanishi@mei.kpu.ac.jp, tabuse@kpu.ac.jp

Abstract: We developed an on-line system for the facial expression recognition of a speaker. In the feature vector space in image processing, the positions of feature vectors generated with imperfection, which caused misrecognition of 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 imperfection in image processing, a module using reject criteria in the feature vector space was added to the system for facial expression recognition. We adopted the utterance of the Japanese name "Taro," which is semantically neutral, to investigate the improved system. The facial expressions of one subject were analyzed when he exhibited one of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised." By using the facial expression strength, the position of the test feature vector in the feature vector space is shown.

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. One of the reasons for this lack of robustness is that 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 method for facial expression recognition applicable under widely varied lighting conditions, we do not use visible ray images. Instead, we use images produced by infrared rays (IR), which show the thermal distribution of the face [1]-[17]. We previously developed an on-line system for the facial expression recognition of a speaker by image processing [13]. In the feature vector space, the positions of feature vectors generated with imperfection, which caused misrecognition of 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 imperfection in image processing, a module using reject criteria in the feature vector space was added to the system [13] for facial expression recognition. As an initial investigation, we adopted the utterance of the Japanese name "Taro," which is semantically neutral.

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 ($=5.6705\times10^{-12}$ W/cm^{2}K^4), and T is the temperature (K). For human skin, ε is estimated as 0.98 to 0.99 [18], [19]. 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]–[17]. In principle, the temperature measurements by IR do not depend on skin color [19], darkness, or lighting condition, and so the face region and its characteristics are easily extracted from a thermal image.

3 PROPOSED SYSTEM

Fig. 1 illustrates the flowchart of our method [17]. 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 [14].

3.1 Speech recognition and dynamic image analysis

We use a speech recognition system named Julius [20] 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

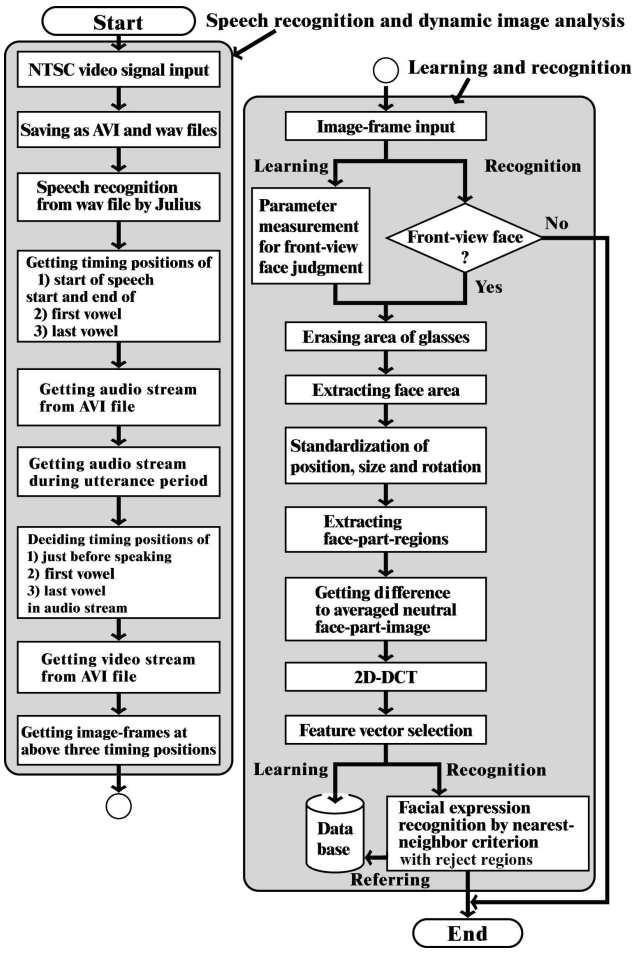

Fig. 1. Flowchart of our method [17]

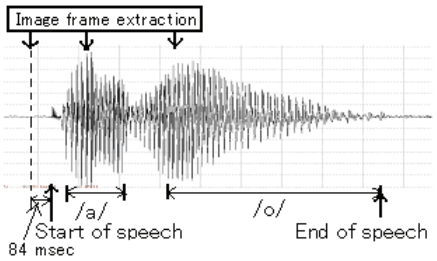

Fig. 2. Speech waveform of "Taro" and timing positions for image frame extraction [8]

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 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 the static images obtained from 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]. Fig. 3 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 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 call the method used in our previously reported research [8]-[16] the "conventional" method, in which the facial expression was 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.

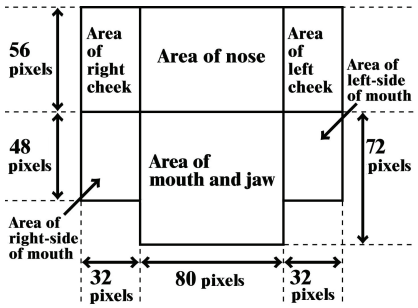

Fig. 3. Blocks for extracting face areas in the thermal image [13]

3.3 Proposed method for facial expression recognition

Imperfection in image processing with our method for facial expression recognition inevitably occurs, and results in the misrecognition of facial expression. To overcome this difficulty, a method is proposed that uses reject criteria in the feature vector space by using the training data just before speaking and when speaking the phonemes of the first and last vowels.

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. Then, go to Step 7.

***Step 7*:** The element $\boldsymbol{v}_i$ having the shortest distance d_i to the test feature vector is found among (1) those selected by the procedure up to Step 5 for each class i and (2) the vector $\boldsymbol{g}_i$. The facial expression strength FES_i for each class i is defined as (d_s/d_i), where $d_s = \min_{i \in I} d_i$ under the condition that set I consists of all classes.

In the present study, the Euclidean distance is used as the distance in the feature vector space. Special treatment for the class of neutral facial expression is used because in many cases the value of (a) tends to be too small to accept the recognized result showing class j. Steps 1 to 6 are the same as those in [17]. The facial expression strength FES_i described in Step 7 is used for expressing the position of the test feature vector in the feature vector space.

3.4 Proposed system

Fig. 4 shows the structure of the proposed system. Figs. 5 and 6 show flow charts of the system during facial expression learning and recognition, respectively. Figs. 5 and 6 demonstrate the processing shared by three computers (PC1, PC2, and PC3) connected by cables to form a local area network for performing the flow chart shown in Fig. 1. In Figs. 5 and 6, a line with an arrow attached denotes a process order, while a dotted line with an arrow attached denotes a communication from a PC to another PC in the local area network. The only difference between Figs. 5 and 6 is the processing in PC3. An NTSC video signal from a thermal video system is successively inputted to either PC1 or PC2 to obtain image frames: (i) just before speaking, and speaking (ii) the first vowel and (iii) the last vowel in an utterance. We use the proposed system with our previously reported method [10] for front-view face judgment.

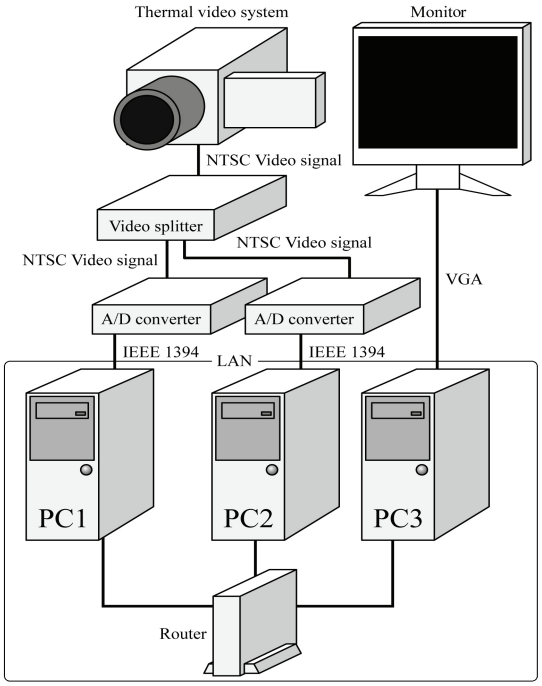

Fig. 4. Structure of the proposed system [13]

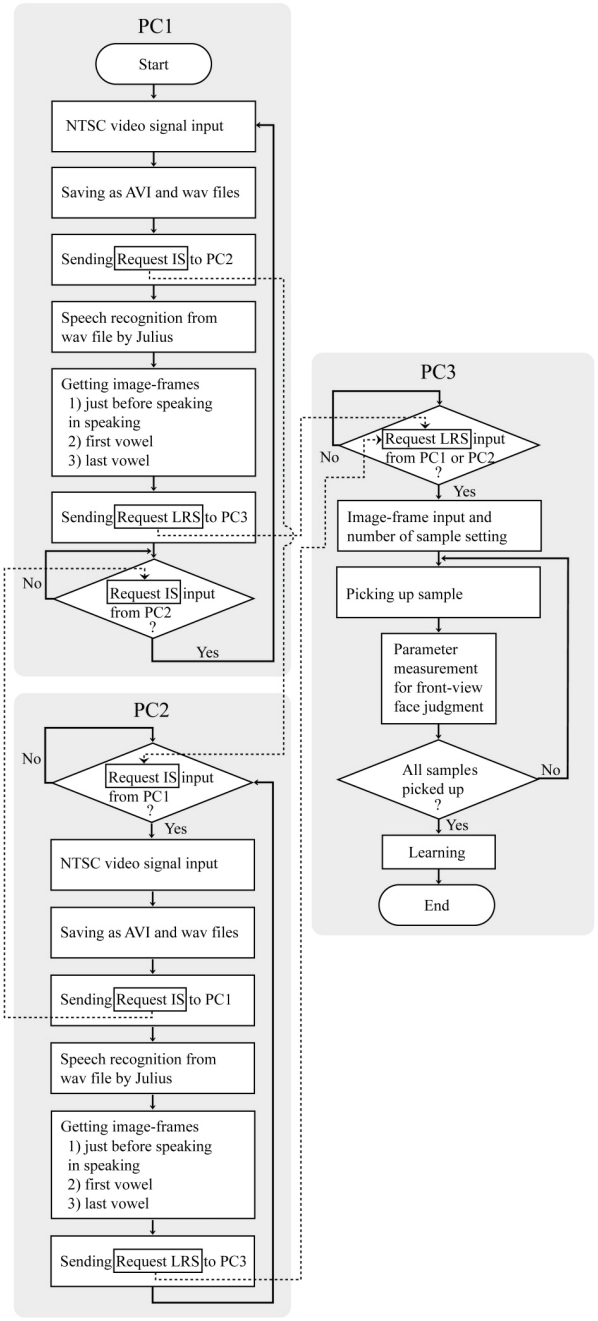

Fig. 5. Flow chart of the proposed system in learning for facial expression recognition [13]

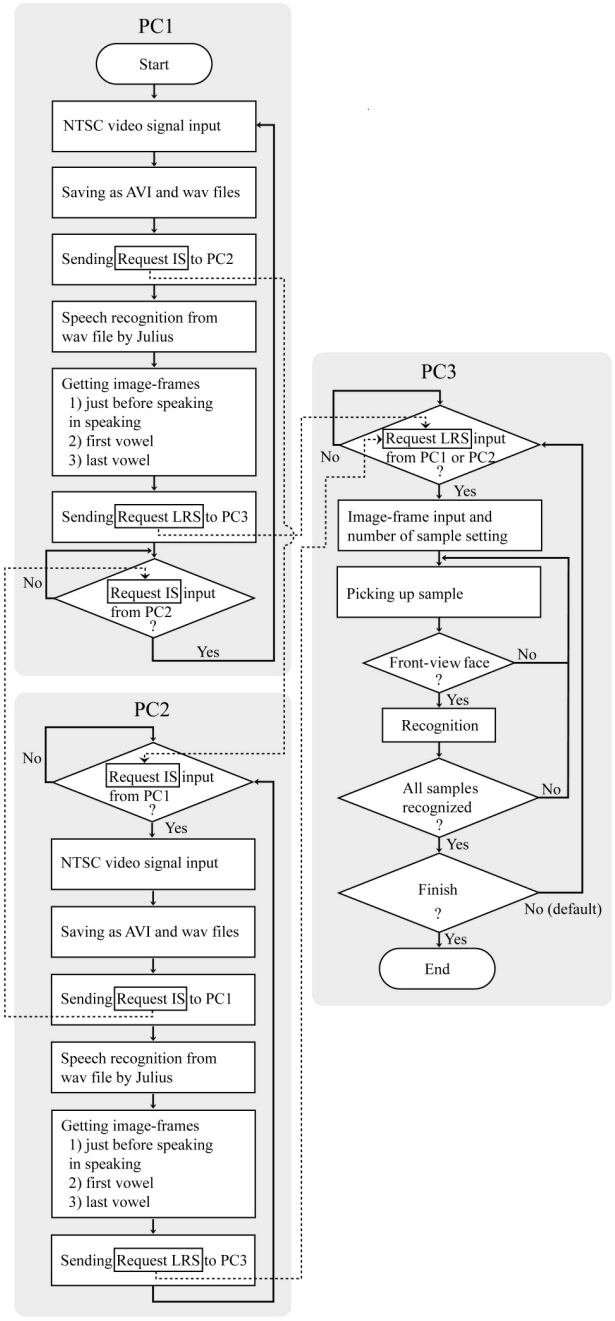

Fig. 6. Flow chart of the proposed system during facial expression recognition [13]

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 through an audio and video distribution amplifier (Maspro Denkoh VSP4) into a digital signal by two A/D converters (Thomson Canopus ADVC-100 for PC1 in Fig. 4 and Thomson Canopus ADVC-300 for PC2 in Fig. 4) and input into two computers (PC1 and PC2 in Fig. 4) with the same specification (Dell Optiplex 780, CPU: Intel Core 2 Duo E8400 3.00 GHz, main memory: 4.00 GB, OS: Windows 7 Professional (Microsoft) with an IEEE1394 interface board (I·O data device 1394-PCI3/DV6)). As PC3 in Fig. 4, we used a computer (Dell Precision T1600, CPU: Intel Xeon CPU E31225 3.10 GHz, main memory: 8.00 GB, OS: Windows 7 Professional (Microsoft)). The three computers (PC1, PC2, and PC3) were connected through a router (Buffalo WZR-HP-AG300H) with cables. 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 of 304 to 309 K. Therefore, one gray level corresponded to 1.95 10^{-2} K. The temperature range for generating a thermal image was decided 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.

Subject A, a male wearing glasses, 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. 7 shows examples of the thermal images of subject A.

In the experiment, subject A intentionally kept front-view faces in the AVI files saved as both the training and 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 the last vowels.

We assembled twenty samples as training data and ten samples as test data for each facial expression in each case, in which all facial expressions of test data for all subjects were judged as front-view faces by the method mentioned in our previously reported paper [10]. For each sample, we obtained three images at the timing positions of just before speaking and while speaking the phonemes of the first and the last vowels.

4.2 Results and discussion

Table 1 shows the facial expression recognition accuracy and rejection ratio for each facial expression. The mean recognition accuracy was 77.2%. The facial expressions of "others" were more difficult to recognize than those of "happy," and "neutral," which were perfectly recognized with 100% accuracy (Table 1).

Table 2 shows examples of the results of facial expression recognition. By using the facial expression strength described in Step 7, the position of the test feature vector in the feature vector space is quantitatively shown.

In the present experiment, a subject spoke one word that was the semantically neutral utterance "Taro." The system recognized the facial expression in speaking each word. When we are ready to apply the proposed system for recognizing facial expressions in daily conversation, we should be able to recognize the facial expression during speaking for a certain interval, such as sentence by sentence. This is because we focus on human feeling through facial expression and it is difficult for humans to change their feeling on a word-by-word basis. When we use the training data for all combinations of the first and the last vowels [11], we can apply the proposed system to a speaker for any utterance. Though several studies on facial expression recognition using thermal image processing have been reported (see references [5]–[17], [21]), only our research [5]–[17] has focused on a speaker. In this paper, we propose an on-line system for recognizing the facial expression of a speaker by adding a module using reject criteria in the feature vector space to our previously reported system [13], which is

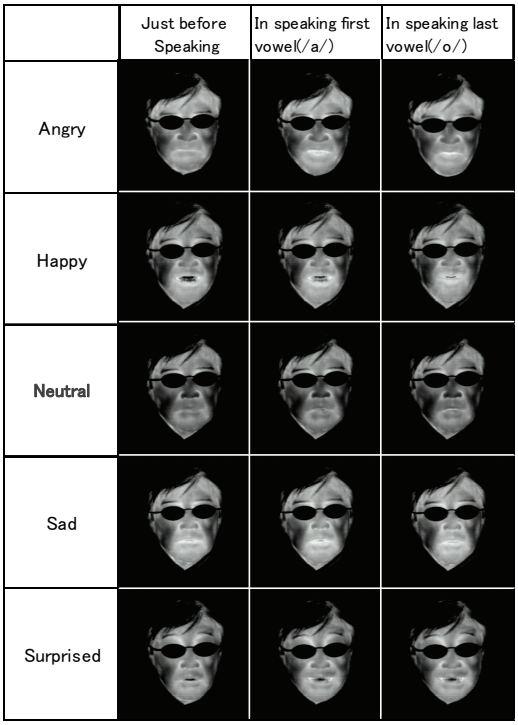

Fig. 7. Examples of thermal images of subject A having each facial expression in speaking [13]

Table 1. Recognition accuracy and rejection ratio

		Input facial expression		
		Happy	Neutral	Others
Output	Happy	**100**		10.5
	Neutral		**100**	57.9
	Others			**31.6**
Rejected		76.0	0.0	59.2

(%)

Table 2. Results of facial expression recognition

Input facial expression	Facial expression strength					Nearest	Second nearest
	A	H	N	Sa	Su		
Angry	**0.98**	0.57	0.30	0.51	1.0	Su	A
Happy	0.21	**1.0**	0.26	0.61	0.19	H	Sa
Neutral	0.13	0.23	**1.0**	0.27	0.12	N	Sa
Sad	0.24	0.56	1.0	**0.71**	0.22	N	Sa
Surprised	0.68	0.45	0.24	0.38	**1.0**	Su	A

A=Angry, H=Happy, N=Neutral, Sa=Sad, Su=Surprised

based on our previously reported method [10], [11], which is superior to our earlier reported method [5]–[8]. Compared with the conventional method, the effect of adding reject criteria in the feature vector space is shown in detail in [17].

5 CONCLUSION

We propose an on-line system for recognizing the facial expression of a speaker by adding a module using reject criteria in the feature vector space to our previously reported system [13]. Using the proposed system, the facial expressions of a subject were discriminable with 77.2% accuracy for the facial expressions of "happy," "neutral," and "others" when the subject exhibited one of the intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised." By using the facial expression strength, the position of the test feature vector in the feature vector space is quantitatively shown. We expect the proposed system to be applicable for recognizing facial expressions in daily conversation.

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

[13] 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(2):263-269

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

[15] 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

[16] 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, in press

[17] Nakanishi Y, Yoshitomi Y, Asada T, et al (2013), Facial expression recognition of a speaker using thermal image processing and reject criteria in feature vector space, In:Sugisaka M (ed), Proceedings of the International Symposium on Artificial Life and Robotics (AROB 18th), Daejeon, Korea, Jan 30 - Feb 1, 2013, in press

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

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

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

[21] 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