Facial Expression Recognition for Speaker Using Thermal Image Processing and Speech Recognition System

YASUNARI YOSHITOMI
Graduate School of Life and Environmental Sciences
Kyoto Prefectural University
1-5 Nakaragi-cho Shimogamo Sakyo-ku Kyoto 606-8522
JAPAN
yoshitomi@kpu.ac.jp http://seika.kpu.ac.jp/~yoshitomi/

Abstract: - I investigated a method for facial expression recognition for a human speaker by using thermal image processing and a speech recognition system. In this study, we improved our speech recognition system to save thermal images at the three timing positions of just before speaking, and just when speaking the phonemes of the first and last vowels. With this method, intentional facial expressions of "angry", "happy", "neutral", "sad", and "surprised" were discriminable with good recognition accuracy.

Key-Words: Facial Expression Recognition, Thermal Image Processing, Speech Recognition System, Vowel, Speaker, Discrete Cosine Transformation

1 Introduction

To better integrate robots into our society, a robot should be able to interact in a friendly manner with humans. The goal of my research is to develop a robot that can perceive human feelings and mental states. For example, a robot could encourage a human who looks sad, advise a person to stop working and rest for a while when the individual looks tired, or take care of a person advanced in years.

The present investigation concerns the first stage of the development of a robot that acquires the ability to detect human feeling or inner mental states visually. Although the mechanism for recognizing facial expressions has received considerable attention in the field of computer vision research [1]–[6], its present stage still falls far short of human capability, especially from the viewpoint of robustness under widely varying lighting conditions. One of the reasons 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 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 describes the thermal distribution of the face [7]–[16]. Although a human cannot detect IR, it is possible for a robot to process the information around it using 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. 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, I introduce our method for the facial expression recognition of a speaker. For facial expression recognition, we select three images: (i) just before speaking; speaking (ii) the first and (iii) last vowels in an utterance. A frame of the front-view face in a dynamic image is selected by estimating the face direction [15]. A two-dimensional discrete cosine transformation (2D-DCT) is performed to transform the grayscale values of each block in the face portion of an image into their frequency components, which are used to generate feature vectors. Using our method, the facial expressions are discriminable with good recognition accuracy when he or she exhibits one of the intentional facial expressions of "angry", "happy", "neutral", "sad", and "surprised".

2 Image Acquisitions

The principle behind thermal image generation is the Stefan–Boltzmann law, which is 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, however, the approximate value of 1 was used as ε for human skin. The value of ε for almost all substances is lower than that for 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 male face images obtained using VR and IR. We can obtain a thermal image expressing the face without light, 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.

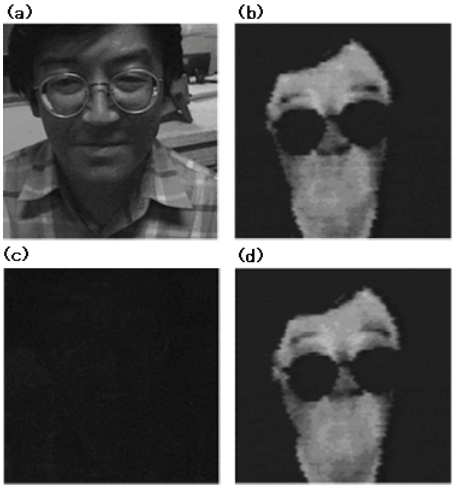

Fig. 1. Examples of face image at night; (a) VR with lighting, (b) IR with lighting, (c) VR without lighting, (d) IR without lighting [9].

3 Basic Processing Flow

As a pre-processing module, we added a judgment function [19] of a front-view face to our method for facial expression recognition [15]. 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 after the front-view face selection. 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. The procedure is explained in the following.

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 [14]–[16]. Fig. 3 shows an example of the wave form of "Taro"; the timing position of the start of speech

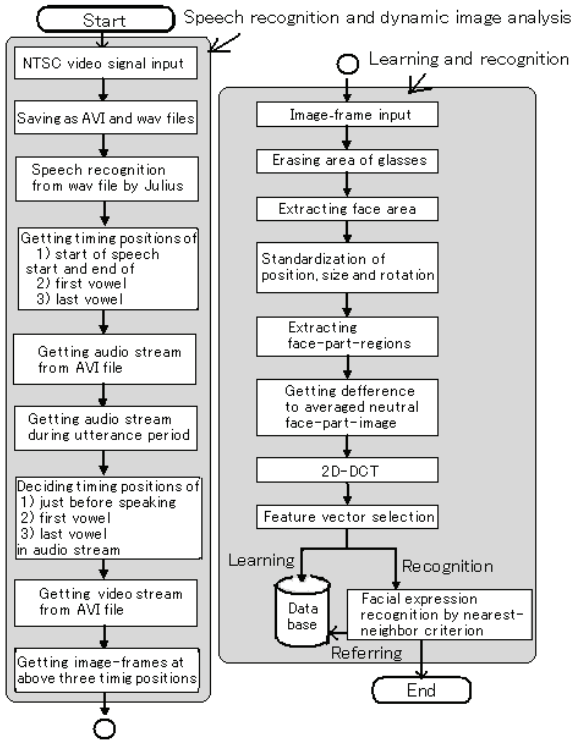

Fig. 2. Flow chart of our method [14].

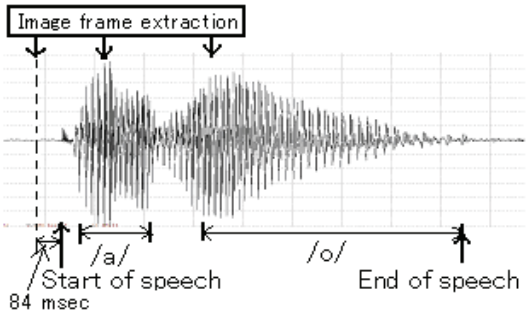

Fig. 3. Speech waveform of "Taro" and timing positions for image frame extraction [14].

and the timing ranges of the first vowel (/a/) and last vowel (/o/) were decided by Julius.

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 of just before speaking, we use the timing position of 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 also use the maximum amplitude.

3.2 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 [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-DCT. The feature vector is generated from the 2D-DCT coefficients according to a heuristic rule [12], [13].

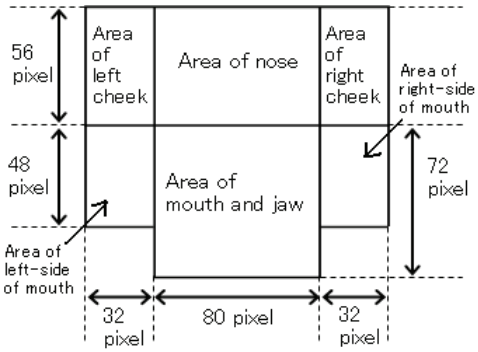

Fig. 4. Blocks for extracting face areas in the thermal image [14].

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, the correction cannot be performed for 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 those at the utterance of the first and last vowels.

4 Performance Evaluation

4.1 Experimental Environment

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) and 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 with an IEEE1394 interface board (I·O Data Device 1394-PCI3/DV6). We used Visual C++ 6.0 (Microsoft) as the programming language.

For generating a thermal image, we applied the condition that the thermal image had 256 gray levels for 5 K. Accordingly, one gray level corresponded to 1.95×10^{-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 into a 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 48 kHz, 16-bit levels, and stereo type.

4.2 Examples of Face Images

Fig. 5 shows examples of the thermal image of each subject. There were four subjects. Subjects A and B were males with glasses. Subject C was a male without glasses. Subject D was a female without glasses. Figs. 6 and 7 show examples of the thermal images of subjects A and D, who exhibited each of the intentional facial expressions, in alphabetical order, of "angry", "happy", "neutral", "sad", and "surprised", while speaking the semantically neutral utterance, the Japanese name "Taro".

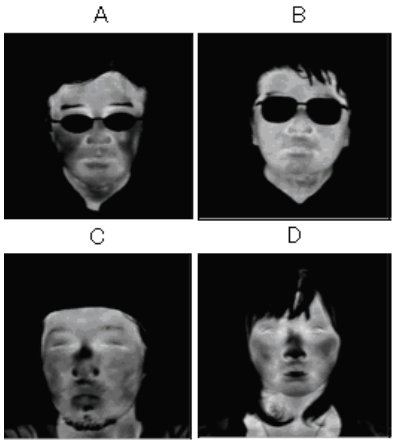

Fig. 5. Examples of thermal images having a neutral facial expression just before speaking [14].

4.3 Experimental Results and Discussion

Table 1 shows the average accuracy of facial expression recognition for the four subjects. The mean recognition accuracy was 80.5%. The proposed method can be applied to any word without the preparation needed for each word in our previously reported method [11]–[13]. In Table 1, the main reason for misrecognition of the facial expression was the misrecognition of vowel(s) [14].

Using the training data for each subject at each timing position, which is (i) just before speaking, and in speaking (ii) the first and (iii) last vowels, we made each averaged neutral face image. For applying the proposed method to any speaker, we need to prepare eleven averaged neutral face images for each subject: one image just before speaking, five images in

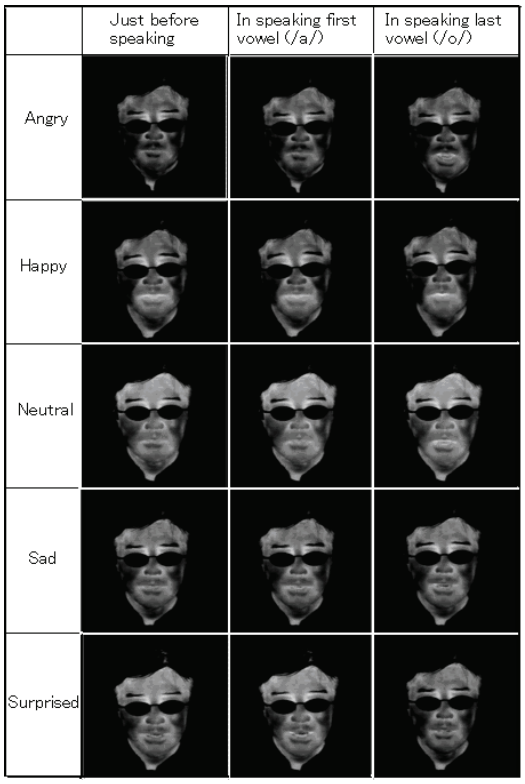

Fig. 6. Examples of thermal images of subject A with each facial expression of speaking [14].

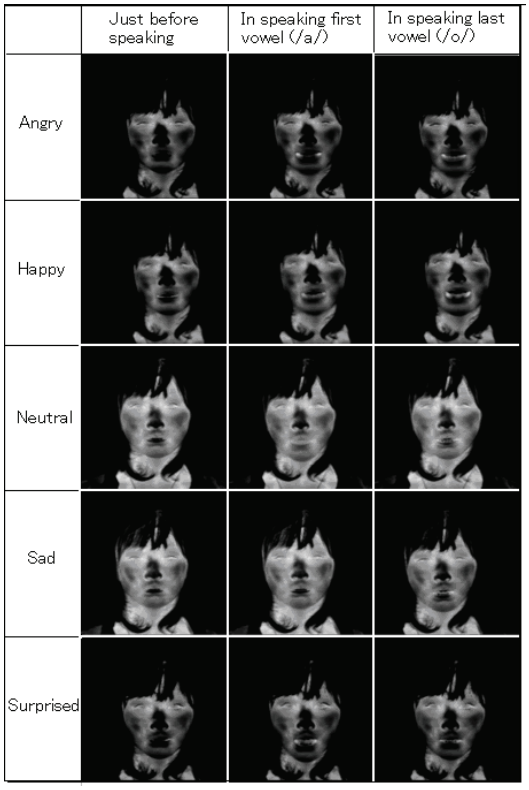

Fig. 7. Examples of thermal images of subject D with each facial expression of speaking [14].

speaking the first vowel, and five images in speaking the last vowel. If the difference of the averaged neutral face image in speaking the first vowel to that in speaking the last vowel is negligibly small, we use the averaged neutral face image for the vowel, resulting in six averaged neutral face images for each subject.

Table 1. Facial expression recognition accuracy [14].

		Input facial expression				
		Angry	Happy	Neutral	Sad	Surprise
Output	Angry	**75**	5		2.5	7.5
	Happy	22.5	**90**			2.5
	Neutral			**80**	5	
	Sad			15	**80**	12.5
	Surprised	2.5	5	5	12.5	**77.5**

We applied the proposed method for one subject who spoke 25 kinds of Japanese names, which provided all combinations of the first and last vowels, after preparing the learning data for all combinations of the first and last vowels. The mean accuracy for facial expression recognition in expressing one of the intentional emotions of “angry”, “happy”, “neutral”, “sad”, and “surprised” was 78.5% [16]. In the experiment, the mean accuracy of the speech recognition of vowels by Julius was 84% [16]. Because there was misjudgment of vowels caused by using the wrong learning data to recognize a facial expression, the facial expression recognition accuracy could be made better by improving the speech recognition accuracy of Julius. In the future, I will only use the face image when speaking a vowel recognized by Julius with word reliability [20] higher than a threshold, which will be decided experimentally beforehand.

It should be noted that the mean judgment accuracy of the front-view face was 99.3% for six subjects, who changed their face direction freely [15].

5 Conclusion

I introduced our method for facial expression recognition for a speaker by using thermal image processing and a speech recognition system. In this study, by using the speech recognition system, thermal static images were saved at the timing positions of just before speaking, and just speaking the phonemes of the first and last vowels. Using our method, five kinds of facial expressions were discriminable with good recognition accuracy. I expect that the proposed method will be applicable for recognizing facial expressions in daily conversation.

Acknowledgment

This work was supported by KAKENHI(22300077).

References:

[1] A. L. Yuille, D. S. Cohen, and P. W. Hallinan, Feature Extraction from Faces Using Deformable Templates, *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1989, pp. 104-109.

[2] H. Harashima, C. S. Choi, and T. Takebe, 3-D Model-based Synthesis of Facial Expressions and Shape Deformation, *Human Interface*, Vol. 4, 1989, pp. 157-166 (in Japanese).

[3] K. Mase, An Application of Optical Flow – Extraction of Facial Expression, *IAPR Workshop on Machine Vision and Application*, 1990, pp. 195-198.

[4] K. Mase, Recognition of Facial Expression from Optical Flow, *Trans. IEICE,* Vol.E74, No. 10, 1991, pp. 3474-3483.

[5] K. Matsuno, C. Lee, and S. Tsuji, Recognition of Facial Expressions Using Potential Net and KL Expansion, *Trans. IEICE*, Vol. J77-D-II, No.8, 1994, pp.1591-1600 (in Japanese).

[6] H. Kobayashi, and F. Hara, Analysis of Neural Network Recognition Characteristics of 6 Basic Facial Expressions, *Proc. of 3rd IEEE Int. Workshop on Robot and Human Communication*, 1994, pp. 222-227.

[7] Y. Yoshitomi, S. Kimura, E. Hira, and S. Tomita, Facial Expression Recognition Using Infrared Rays Image Processing, *Proc. of the Annual Convention IPS Japan*, Vol. 2, 1996, pp. 339-340

[8] Y. Yoshitomi, S. Kimura, E. Hira, and S. Tomita, Facial Expression Recognition Using Thermal Image Processing, *IPSJ SIG Notes*, Vol. CVIM103-3, 1997, pp. 17-24.

[9] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, Facial Expression Recognition Using Thermal Image Processing and Neural Network, *Proc. of 6th IEEE Int. Workshop on Robot and Human Communication*, 1997, pp. 380-385.

[10]Y. Sugimoto, Y. Yoshitomi, and S. Tomita, 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*, Vol. 31, 2000, pp. 147-160.

[11]Y. Yoshitomi, S.-Ill Kim, T. Kawano, and T. Kitazoe, 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*, 2000, pp. 178-183.

[12]F. Ikezoe, R. Ko, T. Tanijiri, and Y. Yoshitomi, Facial Expression Recognition for Speaker Using Thermal Image Processing, *Trans. Human Interface Society*, Vol. 6, No. 1, 2004, pp. 19-27 (in Japanese).

[13]M. Nakano, F. Ikezoe, M. Tabuse, and Y. Yoshitomi, A Study on the Efficient Facial Expression Using Thermal Face Image in Speaking and the Influence of Individual Variations on Its Performance, *J. IEEJ*, Vol. 38, No.2, 2009, pp. 156-163 (in Japanese).

[14]Y. Koda, Y. Yoshitomi, M. Nakano, and M. Tabuse, 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,* 2009, pp. 955-960.

[15]T. Fujimura, Y. Yoshitomi, T. Asada, and M. Tabuse, 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,* to submit.

[16]K. Shimada, Facial Expression Recognition of a Speaker Using Vowel Judgment and Thermal Image Processing, *Graduate Thesis, Kyoto Prefectural University*, 2010.

[17]H. Kuno, *Infrared Rays Engineering*. Tokyo, IEICE, 1994, pp. 22 (in Japanese).

[18]H. Kuno, *Infrared Rays Engineering*, Tokyo, IEICE, 1994, pp. 45 (in Japanese).

[19]Y. Yoshitomi, A. Tsuchiya, and S. Tomita, Face Recognition Using Dynamic Thermal Image Processing, *Proc. of 7th IEEE Int. Workshop on Robot and Human Communication*, 1998, pp.443-448.

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