Facial Expression Recognition Using Facial Expression Intensity Characteristics of Thermal Image

Yasunari Yoshitomi, Taro Asada, Ryota Kato, and Masayoshi Tabuse
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: {yoshitomi, tabuse}@kpu.ac.jp, {t_asada, r_kato}@mei.kpu.ac.jp
http://www2.kpu.ac.jp/ningen/infsys/English_index.html

Abstract

To develop a robot that understands human feeling, we propose a method for recognizing facial expressions. A video was analyzed by thermal image processing and the feature parameter of facial expression, which was extracted in the area of the mouth and jaw by a two-dimensional discrete cosine transform. The facial expression intensity, defined as the norm of the difference vector between the feature vector of the neutral facial expression and that of the observed one, was measured. The feature vector made by facial expression intensity and time at utterance was used for recognizing facial expression.

Keywords: Facial expression recognition, Area of mouth and jaw, Thermal image, and Utterance judgment.

1. Introduction

The goal of our research is to develop a robot that can perceive human feelings and mental states. Although the mechanism for recognizing facial expressions of human feelings has received considerable attention in computer vision research, it currently falls far short of human capability. This is due to the decreased accuracy of facial expression recognition, which is influenced by the inevitable change of gray levels due to nuances of shade, reflection, and local darkness. To avoid this problem and to develop a robust method for facial expression recognition applicable under widely varied lighting conditions, we use an image registered by infrared rays to describe the thermal distribution of the face.[1-3] The timing of recognizing facial expressions is also important for a robot because the processing can be time-consuming. We adopted an utterance as the key of expressing human feelings because humans tend to say something when expressing their feelings.[2, 3]

In the present study, we propose a method for recognizing facial expressions by using the facial expression intensity[4] and the time at utterance.

2. Proposed Method

The proposed method consists of (1) extraction of the area of the mouth and jaw, (2) measurement of facial expression intensity, (3) judgment of utterance, and (4) calculation of feature parameters for facial expression and voice.

2.1. *Extraction of area of mouth and jaw*

The frame extracted every 0.1 second in the dynamic image is used for thermal image processing. Six face areas (Fig. 1) are extracted by the thermal image processing reported in our previous study.[3] The area of the mouth and jaw is selected because the difference between the facial

Published by Atlantis Press

expressions of neutral and happy appears distinctly in this area.[4] Fig. 2 shows an example of the thermal face image and the image of the area of a mouth and jaw.

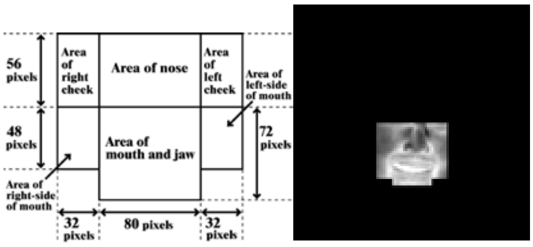

Fig. 1. Blocks for extracting face-part areas (left), and the thermal image after face part extraction (right).[2, 6]

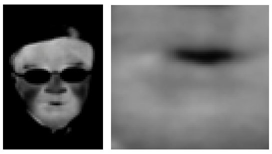

Fig. 2. Thermal face image (left) and image of the area of the mouth and jaw (right).[4]

2.2. *Measurement of facial expression intensity*

For the extracted frame, the feature vector of facial expression is extracted in the area of the mouth and jaw by applying a two-dimensional discrete cosine transform (2D-DCT) for each domain of 8×8 pixels.[4] We select 15 low-frequency components of the 2D-DCT coefficients, except for the direct current component, as the feature parameters for expressing facial expression.[4] Then, we obtain the mean of the absolute value for each 2D-DCT coefficient component in the area of the mouth and jaw.[4] In total, we obtain 15 values as elements of the feature vector. The facial expression intensity, defined as the norm of the difference vector between the feature vector of the neutral facial expression and that of the observed expression, can be used for analyzing a change of facial expression.[4]

2.3. *Judgment of utterance*

The sound data are smoothed and sampled to erase noise. Then, all sampled data that fall within $\left[\overline{x}_s-14\sigma_s,\overline{x}_s+14\sigma_s\right]$, where $\overline{x}_s$ and σ_s express the average and the standard deviation of the sound data value, respectively, for one second under the condition of no utterance, are considered to be the range of no utterance.[4] When at least one sampled datum has a value outside $\left[\overline{x}_s-14\sigma_s,\overline{x}_s+14\sigma_s\right]$, our system judges that the sound data contain an utterance.[4]

2.4. *Feature parameters*

We use two feature parameters as the elements of the feature vector. One is the mean of the standardized facial expression intensity at each utterance and for 0.3 seconds before and after the utterance. The other is the standardized time at each utterance. The standardization used for making feature parameters is expressed by Eq. (1).

$$x^*_{i,j}=\frac{x_{i,j}-\overline{x}_i}{\sigma_i}\ ,\qquad (1)$$

where $x^*_{i,j}$, $x_{i,j}$, $\overline{x}_i$, and σ_i express the standardized feature parameter, the measured feature parameter, the average, and the standard deviations of the measured feature parameters of the training data, respectively, and i, j denote the number (1 or 2) of the feature parameter and number $(1,2,\cdots,m)$ of the utterance, respectively. Then, clustering by using the Ward method is performed for each of the training and the test data to decide major and minor clusters for each class of facial expressions in the feature vector space. Then, for recognizing facial expressions of the test data, we use the coordinates of the center of gravity of the major cluster for each class of facial expressions for the training and the test data.

3. Experiments

3.1. *Conditions*

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 were 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) and Visual C++ 2008 Express Edition (Microsoft) as the programming language. To generate a thermal image, we set the condition so that the thermal image had 256 gray levels for the detected temperature range. The temperature range for generating a thermal image was set 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 of 48 kHz and 16-bit levels was saved in a stereo PCM format.

Subject A, a man with glasses, performed in alphabetical 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 of “taro” (the first and last vowels of which are /a/ and /o/) and “tsubasa” (the first and last vowels of which are /u/ and /a/). In the experiment, Subject A intentionally maintained a front view in the AVI files, which were saved as both training and test data. We assembled 15 samples as the training data and 15 samples as the test data. The AVI files were used for measuring the facial expression intensity. The WAV files obtained from the AVI files were used for measuring time at utterance.

3.2. *Results and discussion*

The thermal face image depended on the emotion of the subject even at 0.3 seconds before starting to speak (Fig. 3). In comparison with the data analyzing at the time of utterance, the difference of the time series of facial expression intensity for the five kinds of emotion became more distinct when analyzing data in the time range from 0.3 seconds before starting to speak to 0.3 seconds after finishing speaking.

Table 1 shows the number of utterances belonging to each cluster. Fig. 4 shows the facial expression intensity corresponding to each cluster in the time range from 0.3 seconds before starting to speak to 0.3 seconds after finishing speaking. Fig. 5 shows the waveform at utterance corresponding to each cluster. Fig. 6 shows the two-dimensional distribution of the center of gravity of the major cluster for each class of facial expressions for the training and the test data. The recognition accuracy of facial expressions was 100% and 60% for the utterance of “taro” and “tsubasa”, respectively, by

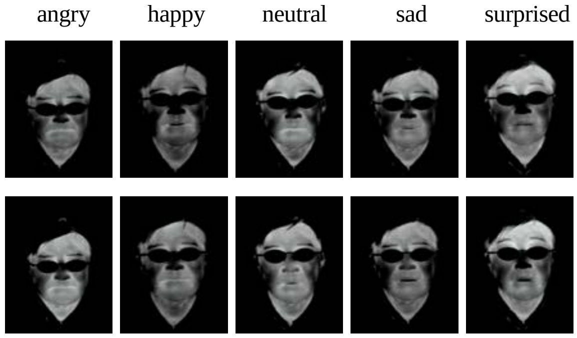

Fig. 3. Thermal face images belonging to major clusters of training data at 0.3 seconds before starting to speak “taro” (upper) and “tsubasa” (lower).

Table 1. Number of utterances belonging to each cluster.

(1) taro

		angry	happy	neutral	sad	surprised
training	major	13	11	10	14	11
	minor	2	4	5	1	4
test	major	13	13	13	8	8
	minor	2	2	2	7	7

(2)tsubasa

		angry	happy	neutral	sad	surprised
training	major	12	10	9	11	14
	minor	3	5	6	4	1
test	major	14	12	14	13	10
	minor	1	3	1	2	5

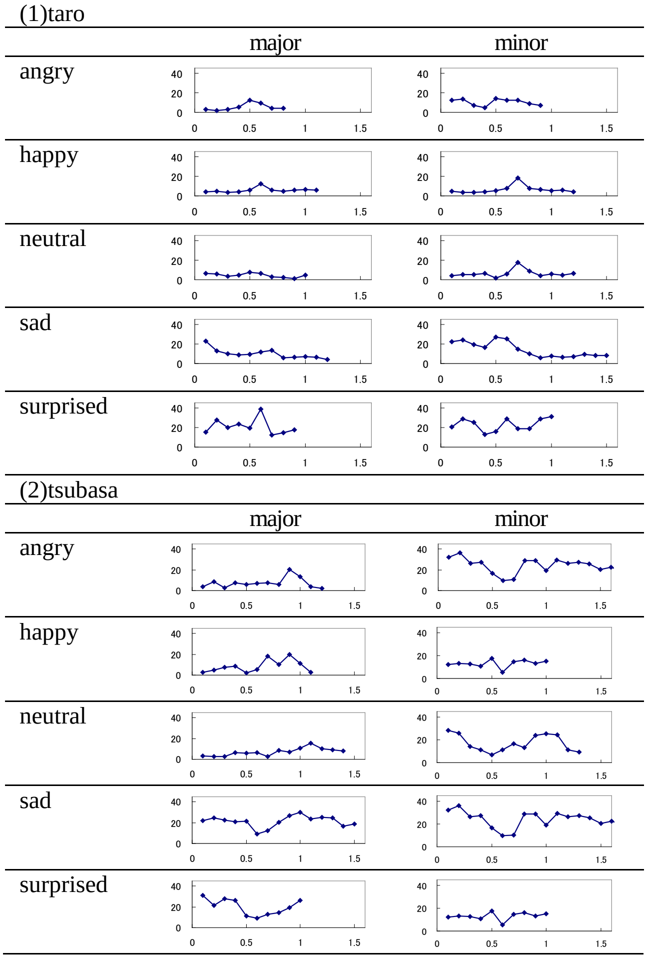

Fig. 4. Example of time series of facial expression intensity corresponding to each cluster of training data; vertical axis: facial expression intensity, horizontal axis: time in seconds.

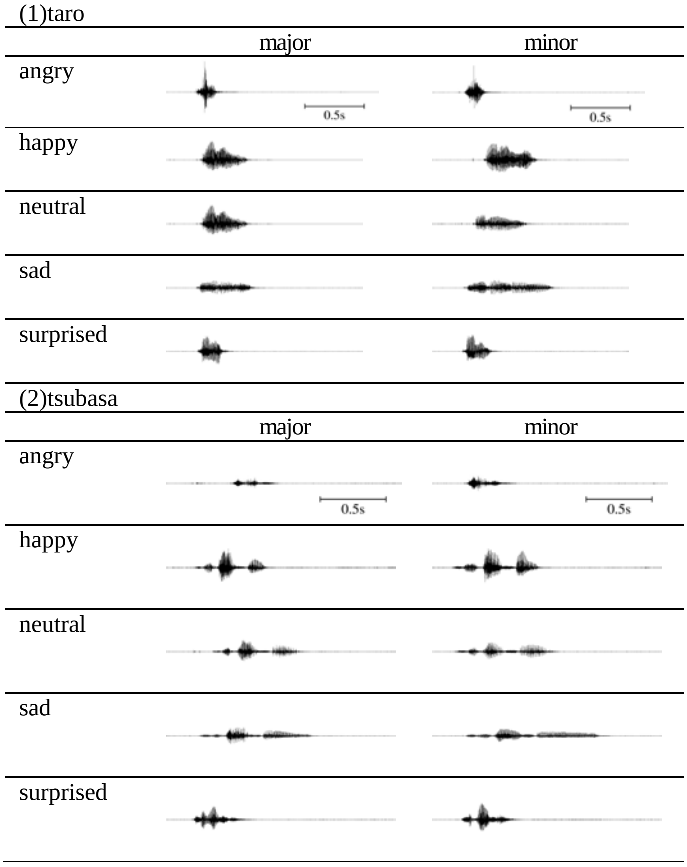

Fig. 5. Example of waveform at utterance corresponding to each cluster of training data.

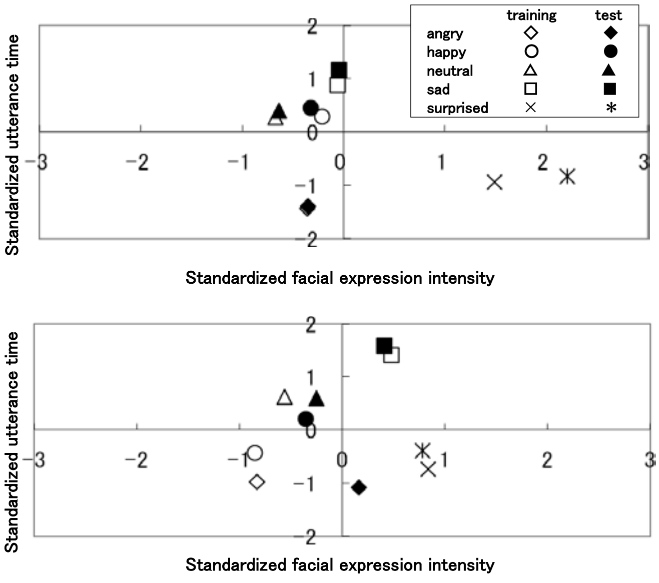

Fig. 6. Two-dimensional distribution of center of gravity of major cluster of training and test data; upper: “taro”, lower: “tsubasa”.

using the nearest neighbor rule. In the case of the angry expression for “taro” in Fig. 6, the symbol for the training data is almost covered by that for the test data.

4. Conclusion

We proposed a method for recognizing facial expressions by thermal image processing and facial expression intensity. The standardized mean value of facial expression intensity for a major cluster, and the standardized mean value of time at utterance for a major cluster are used for recognizing facial expression. The experimental results show the usefulness of the proposed method.

Acknowledgements

The present study was partially supported by KAKENHI (22300077).

References

1. Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, Facial expression recognition using thermal image processing and neural network, in *Proc. 6th IEEE Int. Workshop on Robot and Human Communication*, (Japan, Sendai, 1997), pp. 380–385.
2. Y. Yoshitomi, T. Asada, K. Shimada, and M. Tabuse, Facial expression recognition of a speaker using vowel judgment and thermal image processing, *J. Artif. Life and Robotics* **16**(3) (2011) 318–323.
3. Y. Yoshitomi, M. Tabuse, and T. Asada, Facial expression recognition using thermal image processing, in *Image processing: methods, applications and challenges* ed. V. H. Carvalho (Nova Science Publisher, New York, 2012), pp. 57–85.
4. Y. Yoshitomi, T. Asada, R. Kato, M. Tabuse, Method of facial expression analysis using video phone and thermal image, *J. Robotics, Networking and Artif. Life* **1**(1) (2014) 7-11.