Emotion Recognition of a Speaker Using Facial Expression Intensity of Thermal Image and Utterance Time

Yuuki Oka
NTT DATA Financial Solutions Corp.
Kandanishikicho, Chiyoda-ku, Tokyo 101-0054, Japan

Yasunari Yoshitomi, Taro Asada, 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@mei.kpu.ac.jp}
http://www2.kpu.ac.jp/ningen/infsys/English_index.html

Abstract

Herein, we propose a method for recognizing human emotions that utilizes the standardized mean value of facial expression intensity obtained from a thermal image and the standardized mean value of the time at utterance. In this study, the emotions of one subject could be discerned with 76.5% accuracy when speaking 23 kinds of utterances while intentionally displaying the five emotions of "anger," "happiness," "neutrality," "sadness," and "surprise."

Keywords: Emotion recognition, Mouth and jaw area, Thermal image, Utterance judgment.

1. Introduction

This research is intended to facilitate the development of robots that can perceive human feelings and mental states. While various mechanisms for recognizing human feelings from facial expressions have received considerable attention in the field of computer vision research in recent years, the results achieved to date fall far short of human capabilities. This is due to the limited accuracy of facial expression recognition, which is influenced by inevitable gray level changes that result from local variations in lighting, shade, reflection, and darkness.

To avoid this problem and to facilitate the development of a robust facial expression recognition method that would be applicable to widely varied lighting conditions, we use imagery obtained from infrared rays to describe the thermal distribution of a subject's face.[1-3] The timing used by a robot when attempting to recognize facial expressions is also important because the required processing can be time-consuming. Accordingly, we adopted utterances as the key for expressing human feelings because humans tend to speak aloud when expressing their feelings.[2, 3]

In the present study, using 23 different utterance combinations based on the first and last vowels in each utterance, we investigated the performance of our previously proposed method[4] of recognizing emotions by utilizing facial expression intensity[5] and the utterance time.

2. Evaluation Method

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

2.1. *Extraction of mouth and jaw areas*

Six facial areas extracted at 0.1 frames per second via the thermal image processing reported in our previous study[3] are used in the processing of dynamic thermal imagery, as shown in Fig. 1. The mouth and jaw area was selected because facial expression differences between neutrality and happiness are most distinctive in this area.[5] Fig. 2 shows thermal image examples of whole face and the mouth and jaw areas.

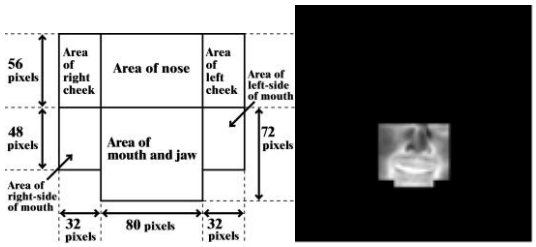

Fig. 1. Blocks for extracting partial face areas (left), and the thermal image after partial face extraction (right).[2]

Fig. 2. Thermal face (left) and mouth and jaw area imagery (right).[5]

2.2. *Measurement of facial expression intensity*

For the extracted frames, the expression feature vectors in the mouth and jaw area are extracted by applying a two-dimensional discrete cosine transform (2D-DCT) for each 8×8 pixel domain.[5] To accomplish this, we begin by selecting 15 low-frequency components of the 2D-DCT coefficients, excluding the direct current component, as the facial expression feature parameters.[5]

Next, we obtain the mean of the absolute value for each 2D-DCT coefficient component in the mouth and jaw area.[5] A total of 15 values are obtained and used as feature vector elements. Facial expression intensity, which is defined as the norm of the differences between the neutral facial expression feature and the observed expression vectors, can then be used for analyzing facial expression changes.[5]

2.3. *Utterance judgment*

After sample collection, the sound data are smoothed to erase noise. Then, all sampled data that fall within $\left[\bar{x}_s - 14\sigma_s, \bar{x}_s + 14\sigma_s\right]$, where $\bar{x}_s$ and σ_s express the average and the standard deviation of the sound data value, respectively, for one second under the no utterance condition, are considered to be the range for no utterance.[5] When at least one sampled datum has a value outside $\left[\bar{x}_s - 14\sigma_s, \bar{x}_s + 14\sigma_s\right]$, our system judges that the sound data contains an utterance.[5]

2.4. *Feature parameters*

In our proposed method,[4] two parameters are used as feature vector elements. One is the mean of the standardized facial expression intensity at the time of each utterance and for 0.3 s before and after the utterance, while the other is the standardized time at each utterance. The standardization used for creating feature parameters is expressed by Eq. (1).

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

where $x^*_{i,j}$, $x_{i,j}$, $\bar{x}_i$, 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, using the Ward method, clustering is performed separately for the training and test data to determine major and minor clusters for each class of the emotions in the feature vector space. Next, to recognize emotions from the test data, the center-of-gravity coordinates of the major cluster for each emotion class of the training and test data is used.

3. Experiments

3.1. *Conditions*

Thermal imagery produced by a TVS-700 thermal video system (Nippon Avionics, Tokyo, Japan) and sounds captured by an ECM-23F5 Electret condenser microphone (Sony, Tokyo, Japan), as amplified by an AT-PMX5P mixer (Audio-Technica, Tokyo, Japan), were transformed into a digital signal via an ADVC-300 analog/digital (A/D) converter (Thomson Canopus, Kobe, Japan).

The converted signals were input into an Optiplex 780 personal computer (PC) (DELL, Round Rock, TX) equipped with an E8400 3.00 GHz Core 2 Duo central processing unit (CPU) (Intel, Santa Clara, CA) and 3.21

GB of main memory, as well as a 1394-PCI3/DV6 IEEE1394 interface board (I-O Data Device, Ishikawa, Japan).

As for software, the PC was equipped with the Windows 7 Professional operating system (OS), while Visual C++ 6.0, and Visual C++ 2008 Express Edition, (Microsoft, Redmond, WA) were used as the programming languages.

To generate a thermal image that would allow the face area of a subject to be easily extracted, 256 gray levels were set to cover the detectable temperature range of our experimental device setup. The visual and audio information was saved in the PC as a Type 2 digital video-audio video interleave (DV-AVI) file, in which the video frame had a spatial resolution of 720 × 480 pixels and 8-bit gray levels, while 48 kHz and 16-bit level sound was saved in a stereo pulse-code modulation (PCM) format.

Subject A, a male wearing eyeglasses, deliberately performed each of the emotions of "anger," "happiness," "neutrality," "sadness," and "surprise," while speaking the semantically neutral utterance of each of the Japanese first names listed in Table 1. However, all utterances of the Japanese first name "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/) were excluded because they had been used in the previous study.[4]

Table 1. Japanese first names used in the experiment.

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

In this study's experiment, Subject A intentionally maintained a front view in the AVI files, which were saved as both training and test data. In total, 15 training data and 15 test data samples were assembled. The AVI files were used for measuring the facial expression intensity. WAVform Audio format (WAV) files obtained from the AVI files were used for measuring the utterance times.

3.2. *Results and discussion*

Our results show that the subject's emotion was reflected in his thermal face image even at 0.3 s before he began to speak.[4] From a comparison with the data analyzed at the time of utterance, we found that the time series facial expression intensity differences for the five targeted emotions became more distinct when the analysis included the time range beginning 0.3 s before to the subject started to speak to 0.3 s after he finishing speaking.[4]

Table 2 shows the emotion recognition accuracy values obtained using our method. As can be seen in this table, the average emotion recognition accuracy was 76.5%, which was near 80% for "taro" and "tsubasa" used in the previous experiment.[4] Table 3 shows the emotion recognition accuracy values for each utterance obtained using our method. The emotion recognition accuracy for each utterance was from 60% to 100%.

In the case of "sadness," the accuracy was 100%, primarily because the utterance time was remarkably longer than for other utterances, which made it easy to distinguish. In contrast, the emotions of "anger," and "surprise," were confused with each other much more often, primarily because the utterance times when expressing these two emotions were quite similar. Fig. 3 shows two-dimensional center-of-gravity distributions for the dominant cluster of each emotion class for both the training and test data samples. When the nearest neighbor rule is used, it can be seen that emotion recognition accuracy was 100% and 60% for the utterance of "kaede" and "fuyu," respectively.

Table 2. Emotion recognition accuracy.

		Input emotion				
		Angry	Happy	Neutral	Sad	Surprised
Recognized emotion	Angry	**86.96**	4.35			26.09
	Happy		**56.51**	13.04		8.7
	Neutral		26.09	**82.61**		8.7
	Sad		8.7		**100**	
	Surprised	13.04	4.35	4.35		**56.51**

(%)

Table 3. Emotion recognition accuracy for each utterance.

	Utterance	Input emotion					Accuracy (%)
		A	H	N	Sa	Su	
Recognized emotion for each utterance	ayaka	Su	H	Su	Sa	Su	60
	kazuki	A	Sa	N	Sa	Su	80
	takeru	Su	N	N	Sa	Su	60
	kaede	A	H	N	Sa	Su	100
	shinnya	A	N	N	Sa	A	60
	hikari	A	Sa	N	Sa	Su	80
	shigeru	A	H	N	Sa	A	80
	misae	A	N	N	Sa	Su	80
	hiroko	A	H	N	Sa	Su	100
	yuki	A	H	N	Sa	A	80
	fuyu	Su	H	N	Sa	A	60
	yusuke	A	H	H	Sa	Su	80
	yuto	A	Su	N	Sa	A	60
	keita	A	N	N	Sa	N	60
	megumi	A	H	H	Sa	A	60
	megu	A	H	N	Sa	Su	100
	keisuke	A	H	H	Sa	Su	80
	keiko	A	N	N	Sa	Su	80
	tomoya	A	N	N	Sa	Su	80
	koji	A	A	N	Sa	N	60
	noboru	A	H	N	Sa	H	80
	kozue	A	H	N	Sa	H	80
	tomoko	A	H	N	Sa	Su	100

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

4. Conclusion

In a previous study, we proposed a method for recognizing human emotions. Using that method, the emotions of one subject were discernable with 76.5% accuracy in speaking each of 23 utterance types while intentionally exhibiting each of the five emotions of "anger," "happiness," "neutrality," "sadness," and "surprise."

The recorded accuracy level was near 80% for "taro" and "tsubasa" in the previous experiment.[4] These results show that the method is capable of producing good results.

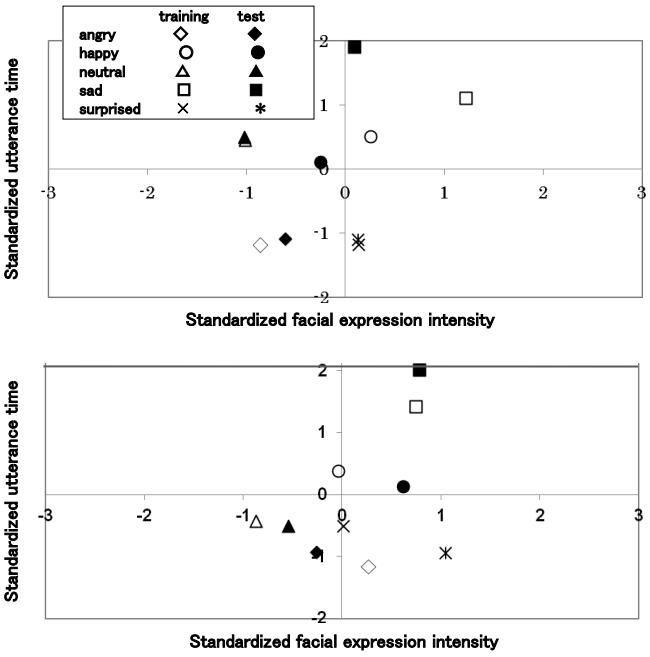

Fig. 3. Two-dimensional center-of-gravity distribution of major cluster of training and test data; upper: "kaede", lower: "fuyu".

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, Facial Expression Recognition Using Facial Expression Intensity Characteristics of Thermal Image, *J. Robotics, Networking and Artif. Life* **2**(1) (2015) 5-8.
5. 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.