Journal of Robotics, Networking and Artificial Life, Vol. 2, No. 2 (September 2015), 79-84

Facial Expression Recognition Using Thermal Image Processing and Efficient Preparation of Training-data

Yuu Nakanishi
Itoki Corporation,
1-4-12 Imafuku-higashi, Joto-ku, Osaka 536-0002, 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

Using our previously developed system, we investigated the influence of training data on the facial expression accuracy using the training data of "taro" for the intentional facial expressions of "angry," "sad," and "surprised," and the training data of respective pronunciation for the intentional facial expressions of "happy" and "neutral." Using the proposed method, the facial expressions were discriminable with average accuracy of 72.4% for "taro," "koji" and "tsubasa"**,** for the three facial expressions of "happy," "neutral," and "other".

Keywords: Efficient gathering of training data, Facial expression recognition, Thermal image processing, Speech recognition, Vowel judgment.

1. Introduction

The present study investigates the first stage of the development of a robot that has the ability to visually detect human feelings or mental states. 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 reason 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. In order to avoid this problem and develop a method for facial expression recognition that is applicable under widely varying lighting conditions, we used images produced by infrared rays (IR), which reveal the thermal distribution of the face.[1-12] We adopted an utterance as the key to expressing human feelings or mental states because humans tend to express feelings vocally.[3-12] Although several studies on facial expression recognition using thermal image processing have been reported (See Refs. 1-17), only our research[3-12] has focused on the speaker. The present study also focuses on the speaker. In addition, we added a judgment function of a front-view face to the proposed method for facial expression recognition.[7] We developed a method for efficiently updating the training data because frequent updates are time-consuming.[9] Through experiments, we concluded that updating the

Published by Atlantis Press

training data corresponding to the facial expressions of "happy" and "neutral" is practical. These two facial expressions are not only very common in our daily lives, but are also easier to express than other facial expressions. Furthermore, the classifications of "neutral," "happy," and "other" are efficient for facial expression recognition under the condition that updating the training data of facial expressions is not performed frequently.

In our previously developed method for facial expression recognition of a speaker, we prepared the training data for a pair of the first and last vowels pronounced by the speaker.[8] The number of pairs is 25 for the Japanese language. It is time-consuming and difficult for a subject to express facial expressions for each pair of vowels.

In the present study, we investigate the influence of training data on facial expression accuracy using the training data of "taro," the first and last vowels of which are /a/ and /o/, for the three intentional facial expressions of "angry," "sad," and "surprised," and the training data of 25 pairs of vowels for the two intentional facial expressions of "happy" and "neutral."

2. Image Acquisition

The principle behind thermal image generation is the Stefan-Boltzmann law, which is expressed as $W = \varepsilon \sigma T^4$, where ε is the 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 to range from 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.[18] Consequently, the human face region is easily extracted from an image using the value of 1 for ε.[1-12] In principle, IR temperature measurements 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. Method for Facial Expression Recognition

Figure 1 shows a flowchart of the proposed method, which consists of two modules. 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.[7] The proposed method is described in detail in our book.[10]

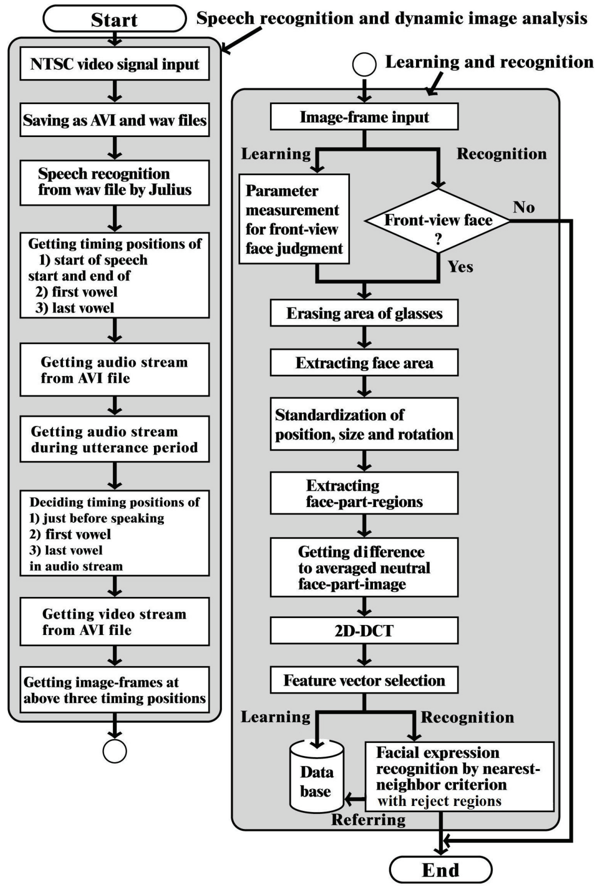
Start
Speech recognition and dynamic image analysis
NTSC video signal input
Saving as AVI and wav files
Speech recognition from wav file by Julius
Getting timing positions of 1) start of speech start and end of 2) first vowel 3) last vowel
Getting audio stream from AVI file
Getting audio stream during utterance period
Deciding timing positions of 1) just before speaking 2) first vowel 3) last vowel in audio stream
Getting video stream from AVI file
Getting image-frames at above three timing positions
Learning and recognition
Image-frame input
Learning
Recognition
Parameter measurement for front-view face judgment
Front-view face ?
No
Yes
Erasing area of glasses
Extracting face area
Standardization of position, size and rotation
Extracting face-part-regions
Getting difference to averaged neutral face-part-image
2D-DCT
Feature vector selection
Learning
Recognition
Data base
Facial expression recognition by nearest-neighbor criterion with reject regions
Referring
End

Fig. 1. Flowchart of the proposed method.[12]

3.1. *Speech recognition and dynamic image analysis*

We use a speech recognition system called Julius[20] to obtain the timing positions of the start of speech and the first and last vowels in a WAV file.[6-12] Figure 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. 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 study.[5] For the timing position of the first vowel, we use the position at which the absolute value of the amplitude of the waveform is the maximum while

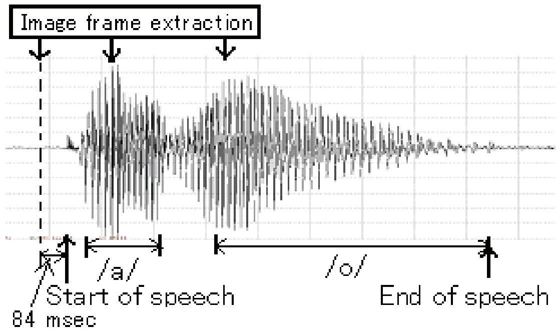

Fig. 2. Speech waveform of "taro" and timing positions for image frame extraction.[6]

speaking the vowel. For the timing position of the last vowel, we apply the 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 previous study.[5, 10] Next, 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 discrete cosine transform (2D-DCT). The feature vector is generated from the 2D-DCT coefficients according to a heuristic rule.[4, 5] The facial expression is recognized by the nearest-neighbor criterion in the feature vector space with the rejection domain using the training data just before speaking and that while speaking the phonemes of the first and last vowels.[12]

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 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) as the programming language. In order to generate a thermal image, we set the condition such that the thermal image had 256 gray levels for the detected temperature range. The temperature range for generating a thermal image was decided so as 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 with glasses, performed in alphabetic 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 listed in Table 1.

Table 1. Japanese first names used in the experiment.[8]

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

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 20 samples as training data and 10 or less samples as test data, in which all facial expressions of the subject were judged as front-view faces by our reported method.[7] The number of test data was decided as a result of the front-view face judgment. From one sample, we obtained three images at the timing positions of just before speaking and while speaking the phonemes of the first and last vowels.

In the present study, we have investigated the influence of training data on the facial expression recognition accuracy. The method for facial expression recognition using the training data of "taro" (the first and last vowels of which are /a/ and /o/), for the three intentional facial expressions of "angry," "sad," and "surprised," and the training data of 25 pairs of vowels for the two intentional facial expressions of "happy" and "neutral" were selected. The method is hereinafter referred to as the efficient method. Then, as

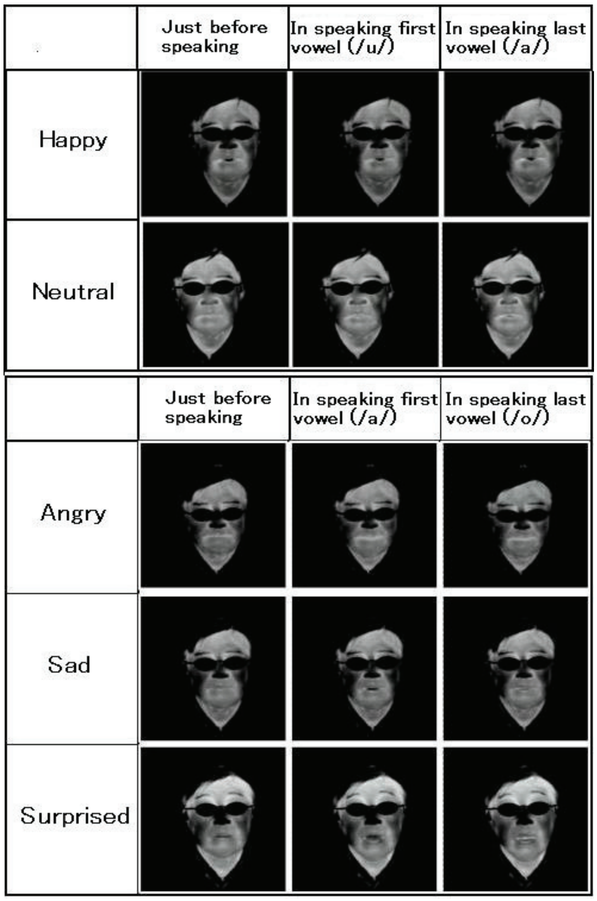
Just before speaking
In speaking first vowel (/u/)
In speaking last vowel (/a/)
Happy
Neutral
Just before speaking
In speaking first vowel (/a/)
In speaking last vowel (/o/)
Angry
Sad
Surprised

Fig. 3. Examples of thermal training images for speaking "tsubasa" in the efficient method.

pronunciations, "taro," "koji" (the first and last vowels of which are /o/ and /i/), and "tsubasa" (the first and last vowels of which are /u/ and /a/), were selected when applying the efficient method.

For comparison with the efficient method, the method for facial expression recognition using the training data of the same pronunciations as those of test data was also applied to the data for "koji" and "tsubasa." This method is hereinafter referred to as the reference method. Figures 3 and 4 show examples of thermal images of the subject used for training by the efficient method and in the test, respectively.

4.2. *Results and discussion*

Tables 2 and 3 show the values of the recognition accuracy of the facial expressions obtained using the efficient method and the reference method, respectively.

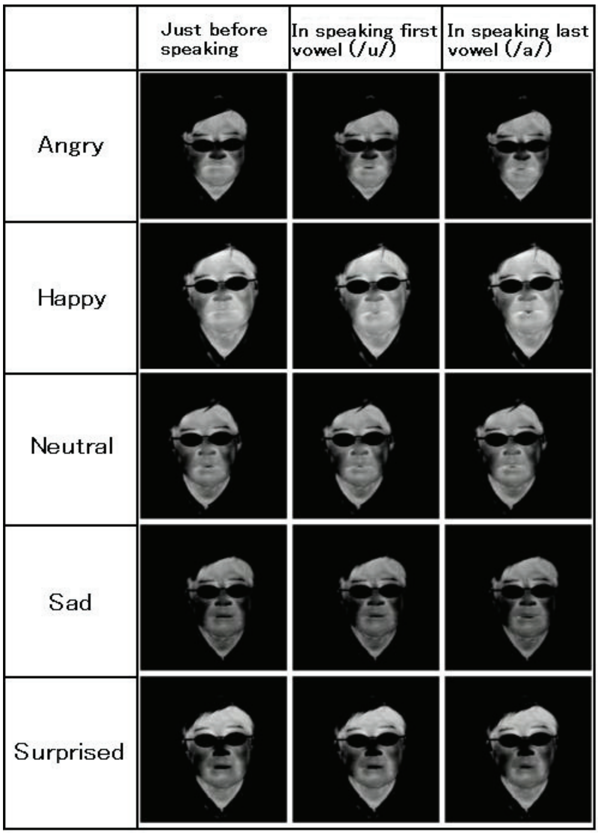

Fig. 4. Examples of thermal test images while speaking "tsubasa".

Using the efficient method, the average facial expression accuracies were 100%, 70.0%, and 47.2%, respectively, for "taro," "koji," and "tsubasa," for the three facial expressions of "happy," "neutral," and "other" when the speaker exhibited one of the five intentional facial expressions of "angry," "happy," "neutral," "sad," and "surprised." On the other hand, using the reference method, the average facial expression recognition accuracies were 90.6% and 85.7% for "koji" and "tsubasa," respectively. Accordingly, the influence of training data on the facial expression recognition accuracy might not be small.

5. Conclusion

We have investigated the influence of training data on facial expression accuracy using training data for "taro," the first and last vowels of which are /a/ and /o/, respectively, for the three intentional facial expressions of "angry," "sad," and "surprised," and the training data of 25 pairs of vowels for the two intentional facial expressions of "happy" and "neutral." Using the efficient

Table 2. Recognition accuracies for “taro,” “koji,” and “tsubasa” as obtained when the efficient method was used.

taro		Input facial expression					
		Correct speech recognition			Poor speech recognition		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	9/9			1/1		
	Neutral		10/10				
	Others			9/9			14/14
Rejected							7
Accuracy		28/28			15/15		
Total accuracy		43/43					

koji		Input facial expression					
		Correct speech recognition			Poor speech recognition		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	0/3			1/4		3/14
	Neutral		9/9				
	Others	3/3			3/4		11/14
Rejected				12	3		1
Accuracy		9/12			12/18		
Total accuracy		21/30					

tsubasa		Input facial expression					
		Correct speech recognition			Poor speech recognition		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	0/9		7/15			
	Neutral	9/9	6/9				
	Others		3/9	8/15			3/3
Rejected			1	8			4
Accuracy		14/33			3/3		
Total accuracy		17/36					

method, the facial expressions of one subject were discriminable with 100%, 70.0%, and 47.2% accuracy for “taro,” “koji” (the first and last vowels of which are/o/ and /i/), and “tsubasa” (the first and last vowels of which are /u/ and /a/), respectively, for the three facial expressions of “happy,” “neutral,” and “other” when the speaker exhibited one of the five intentional facial expressions of “angry,” “happy,” “neutral,” “sad,” and “surprised.” On the other hand, the facial expressions of one subject were discriminable with accuracies of 90.6% and 85.7% for “koji” and “tsubasa,” respectively, when the reference method was used. In order to develop a practical method for facial expression recognition, we intend to investigate a method for reducing the influence of training data on facial expression accuracy.

Acknowledgements

We would like to thank Ms. C. Shiraishi of Kyoto Prefectural University for her valuable support during the data analysis. The present study was supported by KAKENHI (22300077).

Table 3. Recognition accuracies for “koji” and “tsubasa” as obtained when the reference method was used.

koji		Input facial expression					
		Correct speech recognition			Poor speech recognition		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	2/3			4/5		1/7
	Neutral		9/9				
	Others	1/3		8/8	1/5		6/7
Rejected		1	1	2			6
Accuracy		19/20			10/12		
Total accuracy		29/32					

tsubasa		Input facial expression					
		Correct speech recognition			Poor speech recognition		
		Happy	Neutral	Others	Happy	Neutral	Others
Output	Happy	0/4		1/18			
	Neutral	4/4	9/9				
	Others			17/18			4/4
Rejected			1	5			3
Accuracy		26/31			4/4		
Total accuracy		30/35					

References

1. Y. Yoshitomi, S. Kimura, E. Hira, et al, Facial expression recognition using infrared rays image processing, in *Proc. the Annual Convention IPS Japan*, (Osaka, Japan, 1996), **2**, pp.339–340.
2. Y. Yoshitomi, N. Miyawaki, S. Tomita, et al, Facial expression recognition using thermal image processing and neural network, in *Proc. 6th IEEE Int. Workshop on Robot and Human Communication*, (Sendai, Japan, 1997), pp. 380–385.
3. Y. Yoshitomi, S. Kim, T. Kawano, et al, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in *Proc. 9th IEEE Int. Workshop on Robot and Human Interactive Communication*, (Osaka, Japan, 2000), pp. 178–183.
4. F. Ikezoe, R. Ko, T. Tanijiri, et al, Facial expression recognition for speaker using thermal image processing (in Japanese), *Trans. Human Interface Soc.* **6**(1) (2004) 19–27.
5. M. Nakano, F. Ikezoe, M. Tabuse, et al, 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) (2009) 156–163.
6. Y. Koda, Y. Yoshitomi, M. Nakano, et al, Facial expression recognition for a speaker of a phoneme of vowel using thermal image processing and a speech recognition system, in Proc. *18th IEEE Int. Symp. on Robot and Human Interactive Communication,* (Toyama, Japan, 2009), pp. 955–960.
7. T. Fujimura, Y. Yoshitomi, T. Asada, et al, Facial expression recognition of a speaker using front-view face judgment, vowel judgment, and thermal image processing, *J. Artif. Life and Robotics* **16**(3) (2011) 411–417.

8. Y. Yoshitomi, T. Asada, K. Shimada, et al, Facial expression recognition of a speaker using vowel judgment and thermal image processing, *J. Artif. Life and Robotics* **16**(3) (2011) 318–323.
9. Y. Nakanishi, Y. Yoshitomi, T. Asada, et al. Robust facial expression recognition of a speaker using thermal image processing and updating of fundamental training-data, *J. Artif. Life and Robotics* **17**(3) (2013) 342–349.
10. Y. Yoshitomi, M. Tabuse, and T. Asada, Facial expression recognition using thermal image processing, in V. H. Carvalho (ed) *Image processing: methods, applications and challenges* (Nova Science Publisher, New York, 2012), pp. 57–85.
11. Y. Yoshitomi, M. Tabuse, and T. Asada, Vowel judgment for facial expression recognition of a speaker, in Ipšić I (ed) *Speech Technologies* (InTech, Rijeka, 2011), pp. 405–424.
12. Y. Nakanishi, Y. Yoshitomi, T. Asada, et al, Facial expression recognition of a speaker using thermal image processing and reject criteria in feature vector space, *J. Artif. Life and Robotics* **19**(1) (2014), 76–88.
13. D. A. Socolinsky and A. Selinger, A comparative analysis of face recognition performance with visible and thermal infrared imagery, in *Proc. 16th ICPR*, (Quebec City, Canada, 2002), pp. 217–222.
14. M. M. Khan, R. D. Ward, and M. Ingleby, Automated classification and recognition of facial expressions using infrared thermal imaging, in *Proc. IEEE Conf. on Cybernetics and Intelligent Systems*, (Singapore, 2004), pp. 202–206.
15. M. M. Khan, M. Ingleby, and R. D. Ward, Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations, *ACM Trans. on Autonomous and Adaptive Systems* **1**(1) (2006), 91–113.
16. G. Jiang, X. Song, F. Zheng, et al, Facial expression recognition using thermal image, in *Proc. 27th Annual Int. Conf. of the Engineering in Medicine and Biology Society*, (Shanghai, China, 2006), pp. 631–633.
17. B. Hernández, G. Olague, R. Hammoud, et al, Visual learning of texture descriptors for facial expression recognition in thermal imagery, *Computer Vis. and Image Underst.* **106**(2-3) (2007), 258–269.
18. H. Kuno, Infrared rays engineering (in Japanese), *IEICE*, (Tokyo, 1994), pp. 22.
19. H. Kuno, Infrared rays engineering (in Japanese), *IEICE*, (Tokyo, 1994), pp. 45.
20. T. Kawahara, et al, Open-Source Large Vocabulary CSR Engine Julius, *Julius rev.4.1.5.1.* (2010) http://julius.sourceforge.jp/ Accessed 12 March 2013