Journal of Robotics, Networking and Artificial Life, Vol. 1, No. 1 (June 2014), 7-11

Method of Facial Expression Analysis Using Video Phone and Thermal Image

Yasunari Yoshitomi*
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan

Taro Asada
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan

Ryota Kato
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan

Masayoshi Tabuse
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan

*Corresponding author. E-mail: yoshitomi@kpu.ac.jp
Tel&Fax: +81-75-703-5432, http://seika.kpu.ac.jp/~yoshitomi/*

Abstract

To improve the quality of life of elderly people living in a home or healthcare facility, especially in a rural area, we have been developing a method for analyzing the facial expressions of a person using a video phone system (Skype) to speak with another person. In the present study, we proposed a method for analyzing facial expressions of a person using the video phone system to talk to another person. The recorded video is analyzed by thermal image processing and the newly proposed feature vector of facial expression, which is extracted in the mouth area by applying 2D-DCT. 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 to analyze a change of facial expression. The judgment of utterance is performed by using the intensity of the sound wave. The experimental results show the usefulness of the proposed method. We intend to use the proposed method of facial expression analysis to develop a method for estimating the emotions or mental state of people, especially elderly patients.

Keywords: Facial expression analysis, Video phone, Area of mouth and jaw, Thermal image, and Skype.

Published by Atlantis Press

1. Introduction

In Japan, the average age of the population has been increasing, and this trend is expected to continue. Because of this trend, the number of older people with dementia and/or depression, especially those living in rural areas, is increasing very rapidly. Due to the mismatch between the number of patients and the number of healthcare professionals, it is difficult to provide adequate psychological assessments and support for all patients.

Information and communication technology (ICT) is a promising method for overcoming the difficulty caused by the lack of adequate healthcare. In Japan, the first inexpensive connection to the Internet became available only recently in rural areas and high-quality free software, such as Skype[1], is being distributed.

Although the mechanism for recognizing a facial expression 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. To avoid this problem and develop a method for facial expression recognition that is applicable under widely varying lighting conditions, we use images produced by infrared rays (IR), which reveal the thermal distribution of the face.[2-13]

To improve the quality of life (QOL) of elderly people living in a home or healthcare facility, we have been developing a method for analyzing the facial expressions of a person using a video phone system to speak with another person. In the present study, we proposed a method for analyzing the facial expressions of a person using the Skype video phone system. Instead of a visible ray image, we use an IR image that is applicable under widely varying lighting conditions. Moreover, the judgment of utterance is performed by using the intensity of the sound wave.

2. Proposed Method

2.1. *System overview and outline of the method*

As already mentioned, the video phone is Skype[1]. VodBurner (Netralia Pty Ltd.)[14] is introduced for recording the audio and video dialogue. Tapur[15] is also introduced for recording the audio data. Conversations are recorded for the analysis of facial expression. The recorded data are analyzed by thermal image processing and the newly proposed feature vector of facial expression described in this paper. The proposed method consists of (1) extraction of the area of the mouth and jaw, (2) measurement of facial expression intensity, and (3) judgment of utterance. In the following subsections, these three are explained in detail.

2.2. *Extraction of area of mouth and jaw from a dynamic image*

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 study.[11] The area of the mouth and jaw is selected because the difference between the facial expressions of neutral and happy distinctly appears in this area.

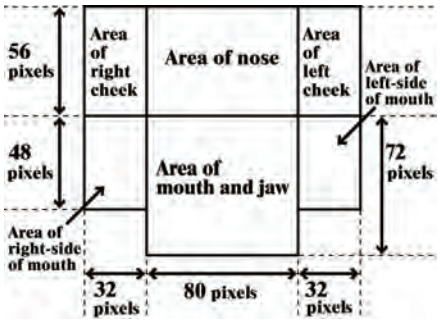

Fig. 1. Blocks for extracting face areas in the thermal image.[11]

2.3. *Measurement of facial expression intensity*

For the extracted frame, the newly proposed 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.

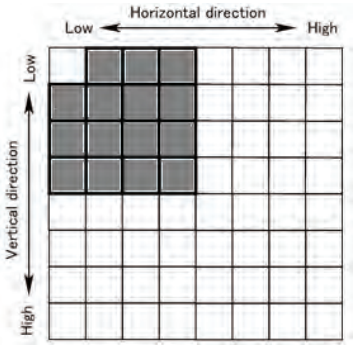

Fig. 2. Special frequency bands used for the analysis.

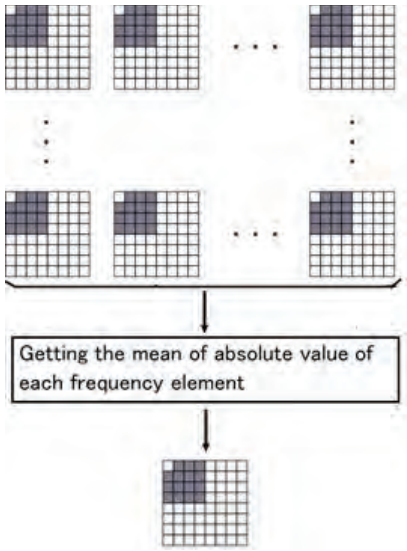

Fig. 3. Schematic diagram of the DCT feature parameter calculation in the mouth area.[11]

The high-frequency components of the 2D-DCT coefficients tend to express a minute change in the data, and thus result in the presence of noise. Therefore, we select 15 low-frequency components of the 2D-DCT coefficients, except for a direct current component, as the feature parameters for expressing facial expression (Fig. 2). Because we do not know the combination of the specific face location and the frequency component of the 2D-DCT coefficients to successfully recognize a facial expression, we adopt the strategy described below.

To gather useful information from the mouth area, we obtain the absolute value of 2D-DCT coefficients, then we obtain the mean of the absolute value for each 2D-DCT coefficient component in the mouth area (Fig. 3). The number of 2D-DCT coefficient components is 15. Therefore, we obtain 15 values as the 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.

2.4. *Judgment of utterance*

Combining the video signal obtained from Skype with the sound signal, we can distinguish the facial expression with speaking from that without speaking. Based on the method reported in Refs. 16–17, the sound data are smoothed and sampled to erase noise. The judgment of speaking is performed by using a threshold of the sound intensity. The threshold is determined by the average and the standard deviation of the sound intensity when the subject does not speak in the sound environment where Skype is used. The thresholds for the sound data values are set as $\bar{x}_s - 14\sigma_s$ and $\bar{x}_s + 14\sigma_s$, where $\bar{x}_s$ and σ_s express the average and the standard deviation, respectively, of the sound data value for one second under the condition of no utterance.

Then, all sampled data that fall within $[\bar{x}_s - 14\sigma_s, \bar{x}_s + 14\sigma_s]$ are considered to be the range of no utterance. When at least one sampled datum has a value outside $[\bar{x}_s - 14\sigma_s, \bar{x}_s + 14\sigma_s]$, our system judges that the sound data contain an utterance.

3. Experiment

3.1. *Condition*

The thermal image was produced by a thermal video system (Thermo Shot F30, NEC Avio Infrared Technologies Co.). Two males (subjects A in his 50s and B in his 20s) participated in the experiment. Using Skype, they held a conversation for approximately 100 seconds. The videos saved by VodBurner were transformed into AVI files, and WAV files were saved by Tapur. The AVI files were used for measuring the

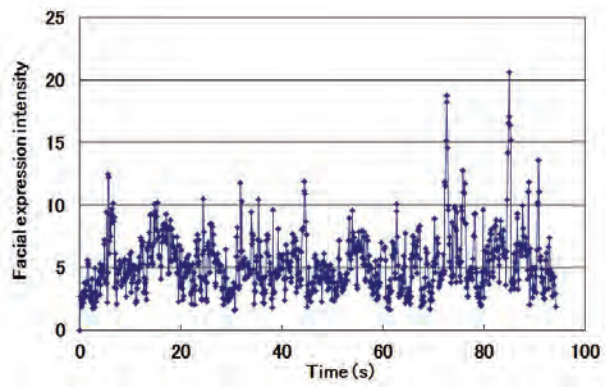

Fig. 4. Facial expression intensity change of Subject A during the conversation with Subject B.

facial expression intensity. The WAV files were used for judgment of the utterance of both subjects A and B.

3.2. *Results and discussion*

Facial expression intensity change of Subject A during the conversation with Subject B were recorded (Fig. 4), and the subjects' utterances were classified (Fig. 5). In Fig. 5, images of the face and images of the mouth and the jaw show the characteristic timing positions for the facial expression intensity.

Subject A expressed constant changes of facial expression intensity during utterances (Fig. 5 (a), (b)), while he expressed drastic changes of facial expression intensity during no utterances (Fig. 5 (c), (d)), which corresponded to his smile. Therefore, the big changes in facial expression intensity shown in Fig. 4 are mainly expressed in the natural smile of Subject A.

Several images of the face and images of the mouth and jaw at the characteristic timing points show that the proposed method can quantitatively express the facial expression (Fig. 5 lower images).

Based on the proposed method, we intend to develop a method for estimating the emotions or mental state of a patient.

4. Conclusion

We proposed a method for analyzing the facial expressions of a person while speaking with a video phone system (Skype). The recorded video is analyzed by thermal image processing and the newly proposed

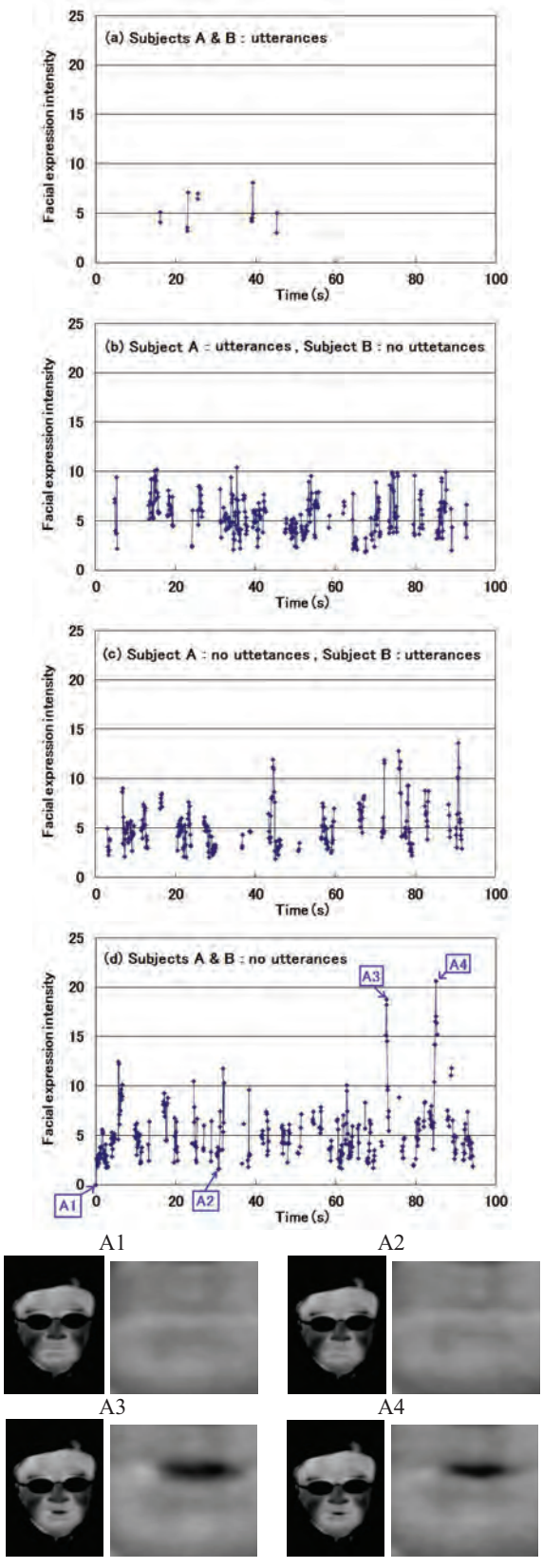

Fig. 5. Facial expression intensity changes (upper graphs), images of face and images of area of mouth and jaw (lower) of subject A during the conversation between subjects A and B.

feature vector of facial expression, which is extracted in the mouth area by applying 2D-DCT. 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 to analyze the change of facial expression. The judgment of utterance is performed by using the intensity of the sound wave. The experimental results show the usefulness of the proposed method.

Acknowledgements

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

References

1. Skype. http://www.skype.com/ Accessed 5 November 2013.
2. Y. Yoshitomi, S. Kimura, E. Hira, and S. Tomita, Facial expression recognition using infrared rays image processing, in *Proc. Ann. Conv. IPS Japan*, (Japan, Osaka, 1996), **2**, pp. 339–340.
3. 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.
4. Y. Yoshitomi, S. Kim, T. Kawano, and T. Kitazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, in *Proc. 6th IEEE Int. Workshop on Robot and Human Interactive Communication*, (Japan, Osaka, 2000), pp. 178–183.
5. F. Ikezoe, R. Ko, T. Tanijiri, and Y. Yoshitomi, Facial expression recognition for speaker using thermal image processing (in Japanese), *Trans. Human Interface Soc.* **6**(1) (2004) 19–27.
6. 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 (in Japanese), *J. IEEJ* **38**(2) (2009) 156–163.
7. 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, in *Proc. 18th IEEE Int. Symp. on Robot and Human Interactive Communication*, (Japan, Toyama, 2009), pp. 955–960.
8. 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, *J. Artif. Life and Robotics* **16**(3) (2011) 411–417.
9. 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.
10. Y. Nakanishi, Y. Yoshitomi, T. Asada, and M. Tabuse, 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.
11. 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.
12. Y. Yoshitomi, M. Tabuse, and T. Asada, Vowel judgment for facial expression recognition of a speaker, in *Speech Technologies*, ed. I. Ipšić (InTech, Rijeka, 2011), pp. 405–424.
13. Y. Nakanishi, Y. Yoshitomi, T. Asada, and M. Tabuse, 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.
14. VodBurner. http://www.vodburner.com/ Accessed 1 December 2013.
15. Tapur. http://www.tapur.com/jp/ Accessed 6 December 2013.
16. F. Ikezoe, M. Nakano, Y. Yoshitomi, and M. Tabuse, Facial expression recognition using thermal face image automatically acquired in speaking (in Japanese), in *Proc. Human Interface Symp. 2005,* (Japan, Fujisawa, 2005), pp. 7–12.
17. M. Nakano, Y. Yoshitomi, and M. Tabuse, Efficient facial expression recognition using thermal face image in speaking and its application to analysis of individual variations (in Japanese), in *Proc. of Human Interface Symp. 2006*, (Japan, Kurashiki, 2006), pp. 1151–1156.