Journal of Robotics, Networking and Artificial Life, Vol. 2, No. 4 (March 2016), 258-262

Facial Expression Analysis while Using Video Phone

Taro Asada[1], Yasunari Yoshitomi[1], Airi Tsuji[2], Ryota Kato[1], Masayoshi Tabuse[1], Noriaki Kuwahara[2], and Jin Narumoto[3]
1: Graduate School of Life and Environmental Sciences Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan, {t_asada, r_kato}@mei.kpu.ac.jp, {yoshitomi, tabuse}@kpu.ac.jp,
2: Graduate School of Science and Technology, Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto 606-8585, Japan, diff.dim0505@gmail.com, nkuwahar@kit.ac.jp
3: Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kajii-cho, Kawaramachi-Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan, jnaru@koto.kpu-m.ac.jp

Abstract

We have been developing a method for analyzing the facial expressions of a person using a video phone system (Skype) to talk to another person. The video is recorded and analyzed by image processing software (OpenCV) and the newly proposed feature vector of facial expression. The newly proposed facial expression intensity can be used to analyze a change of facial expression. The judgment of speaking is performed in our study. The experimental results show the usefulness of the proposed method.

Keywords: Facial expression analysis, Mouth area extraction, OpenCV, Skype, Video phone.

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 elderly 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.

According to the experiences of psychiatrists in the medical treatment, people with dementia and/or depression have a poor change of facial expressions. On the other hands, the number of elderly people with dementia and/or depression is increasing rapidly because of the recent growing trend of aging. Therefore, a method for assessing emotional state of an elderly person with dementia and/or depression by analyzing facial expressions is missing.

Information and communication technology 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.

To improve the quality of life of elderly people living in a home or a healthcare facility, we propose 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, the phone video is recorded and analyzed by image processing software (OpenCV) and the newly proposed feature vector of facial expression, which is extracted in the mouth area. Moreover, the judgment of speaking is performed by using the intensity of the sound wave.

Published by Atlantis Press

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.)[2] is introduced for recording the audio and video dialogue. Tapur[3] is also introduced for recording the audio data. Conversations are recorded for the analysis of facial expression. The recorded data are analyzed by image processing software, Open Source Computer Vision Library (Open CV, Intel), for real-time computer vision[4] and the newly proposed feature vector of facial expression described in this paper. The Y component obtained from each frame in the dynamic image is used for measuring the facial expression intensity. The proposed method consists of (1) extraction of the mouth area, (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 mouth area from a dynamic image*

First, the face area is extracted from each frame in the dynamic image by the classifiers for a front-view face included in OpenCV. The Haar-like feature parameters and Adaboost algorithm for learning are used as the classifiers.[5] It is assumed that the distance between a subject and the camera generally remains constant during a Skype conversation. However, we observed that the size of the face area tends to increase when the face deviates from a front-view image. In this case, it is difficult for OpenCV to extract the face area. Therefore, the minimum size of dynamic image within a specified period is assumed to be the most likely front view. In the present study, we set one second as the period. Next, by using OpenCV, the mouth area is extracted for the frame selected by the face-area size criterion described above. The mouth area is selected because the difference between the facial expressions of neutral and happy distinctly appears in this area.

It is essential for assessing emotional state of a person with dementia and/or depression to judge whether he or she can react with a facial expression of "happy" during the talk. Therefore, we think that the mouth area is enough to be analyzed at the present stage of our study.

2.3. *Measurement of facial expression intensity*

For the Y component of the frame selected by the processing described above, the newly proposed feature vector of facial expression is extracted in the mouth area by applying 2-dimensional Discrete Cosine Transform (2D-DCT) for each domain of 8×8 pixels.

The mouth area used for measuring the facial expression intensity has n blocks of 8×8 pixels, where n is obtained as $n = [a/8] \times [b/8]$, a and b denote the number of pixels of the mouth area in the face area obtained by OpenCV in the vertical and horizontal directions, respectively, and $[x]$ equals the maximum integer that does not exceed x.

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. 1). This selection of 2D-DCT coefficients is popular among researchers in facial expression recognition.[6]

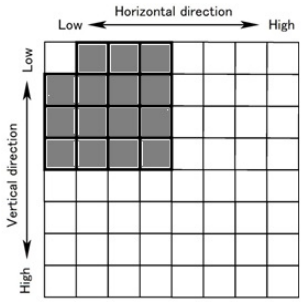

Fig. 1. Special frequency bands used for the analysis.

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. 2). 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

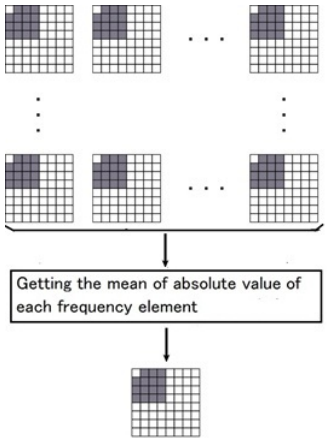

Fig. 2. Schematic diagram of the DCT feature parameter calculation in the mouth area.[7]

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. 8–10, 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, every sampled data that falls 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*

Two males (subjects A and B) in their 20s participated in the experiment. Using Skype, the two subjects held a conversation for approximately 80 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 facial expression intensity. The WAV files were used for judgment of an utterance.

3.2. *Results and discussion*

Facial expression intensity changes of subjects A and B during their conversation were recorded (Fig. 3). The timing of utterances (Fig. 4), and the timing of no utterances (Fig. 5) are shown. In both Figs. 4 and 5, face images and mouth images show the characteristic timing positions for the facial expression intensity.

Subject A expressed four local peaks of facial expression intensity at approximately 20, 37, 47, and 71 seconds from the start point, while subject B expressed six local peaks of facial expression intensity at approximately 40, 45, 55, 65, 71, and 75 seconds from the start point (Fig. 3). For the timing of utterances during their conversation, subject A expressed four local peaks of facial expression intensity at approximately 18, 37, 47, and 71 seconds from the start point, while subject B expressed five local peaks of facial expression intensity at approximately 45, 55, 65, 71, and 75 seconds from the start point (Fig. 4). For the timing of no utterances during their conversation, subject A expressed three local peaks of facial expression intensity at approximately 20, 46, and 70 seconds from the start point, while subject B expressed one local peak of facial expression intensity at approximately 40 seconds from the start point (Fig. 5).

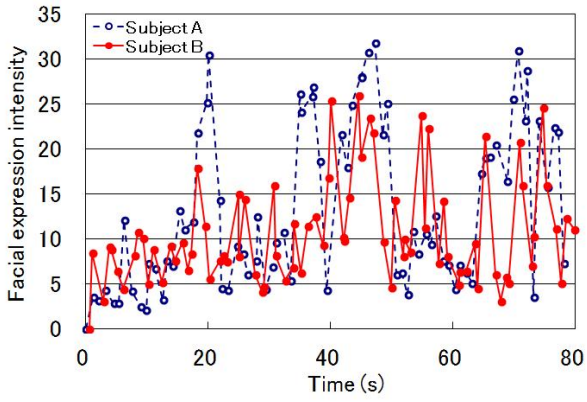

Fig. 3. Facial expression intensity change of subjects A and B during their conversation.

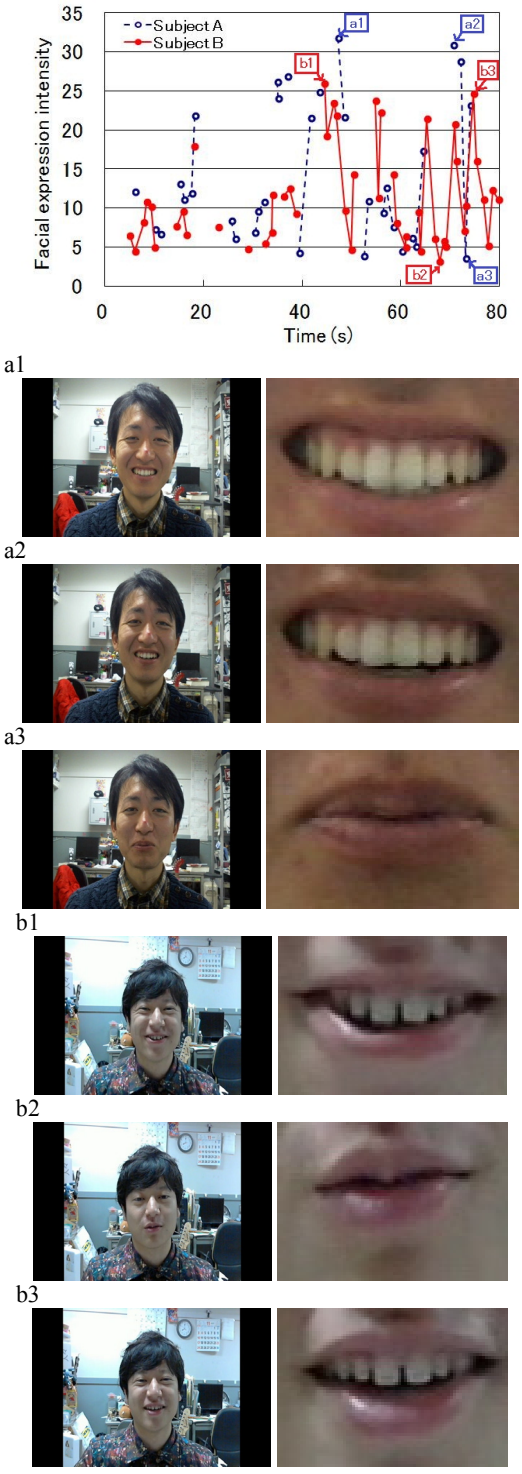

Fig. 4. Facial expression intensity changes (upper graph), face images (lower left side), and mouth image (lower right side) of subjects A and B at the timing of utterances during their conversation.

Fig. 5. Facial expression intensity changes (upper graph), face images (lower left side), and mouth images (lower right side) of subjects A and B at the timing of no utterances during their conversation.

Subject B mainly made an utterance during the second half of the conversation (Fig. 4), while the facial expression intensity increased gradually for the timing of no utterances during the first half of the conversation (Fig. 5). Subject A expressed two local peaks of facial expression intensity at approximately 18 and 37 seconds from the start point, and subject B expressed no local peaks of facial expression intensity when the data were limited to the timing of utterances during the first half of the conversation (Fig. 4). Just after the only local peak of facial expression intensity having no utterance at approximately 40 seconds from the start point, subject B expressed continual peaks of facial expression intensity with utterances.

The images of the face and mouth areas at the characteristic timing points show that the proposed method can quantitatively express the facial expression (Figs. 4 and 5).

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 image processing software (OpenCV) 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 the change of facial expression. The judgment of speaking is performed by using the intensity of the sound wave. The experimental results show the usefulness of the proposed method.

Acknowledgements

We would like to thank the subjects who participated in the experiments. This research is supported by SCOPE (122307003) of the Ministry of Internal Affairs of Communications of Japan and COI STREAM of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Skype Web page, http://www.skype.com/ Accessed 5 November 2013.
2. VodBurner Web page, http://www.vodburner.com/ Accessed 1 December 2013.
3. Tapur Web page, http://www.tapur.com/jp/ Accessed 1 December 2013.
4. OpenCV Web page, http://opencv.willowgarage.com/ Accessed 1 December 2013.
5. P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (USA, Kauai, 2001), Vol.1, pp.511-518.
6. T. Sakaguchi and S. Morishige, Real-time facial expression recognition based on the 2-dimensional DCT, *Trans IEICE* **J80-D-Ⅱ** (6) (1997) 1547-1554.
7. 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.
8. 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.
9. 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.
10. M. Nakano, Robust facial expression recognition for various speakers (in Japanese), *Master's thesis,* Dept. of Environmental Information Graduate School of Human Environmental Sciences Kyoto Prefectural University, (2008), pp 1-70.