Quantitative Evaluation of Facial Expressions and Movements of Persons While Using Video Phone

Taro Asada, Yasunari Yoshitomi, 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: {t_asada, r_kato}@mei.kpu.ac.jp, {yoshitomi, tabuse}@kpu.ac.jp

Jin Narumoto
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

A video is analyzed by image processing and the feature parameters of facial expressions and movements, which are extracted in the mouth area. The feature parameter for expressing facial expressions is defined as the average of the facial expression intensity. The parameter for expressing movement of a person is defined as the average of the absolute value of the vertical coordinate for the center of gravity of the mouth area in the relative coordinate system. The experimental result shows the usefulness of the proposed method.

Keywords: Facial expression analysis, Movement analysis, Mouth area, OpenCV, and Skype.

1. Introduction

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

To improve the quality of life (QOL) of elderly people living in a care facility or at home, we have been developing a method for analyzing the facial expressions of a person while speaking with another person on a video phone.[1-3] In the present study, we developed a method for analyzing facial expressions and movements of a person while speaking with another person by using our previously reported method[2], which is standardization of the size of face to be analyzed, and newly proposed feature parameters of facial expressions and movements of a person.

2. Proposed Method

2.1. *System overview and outline of the method*

The platform includes Skype[4] for the video phone. To record the audio and video dialogue, Netralia Pty Ltd's VodBurner[5] and Tapur[6] are introduced. Conversations are recorded for the analysis of facial expressions and movements of a person. The recorded data are analyzed by using the image processing software Open Source Computer Vision Library for real-time computer vision developed by Intel (Open CV)[7], standardization of the size of face to be analyzed, and the newly proposed

Published by Atlantis Press

feature parameters of facial expressions and movements of a person, as described in the following subsections. The Y component obtained from each frame in the dynamic image is used for analyzing facial expressions and movements of a person. The proposed method consists of (1) standardization of the size of the lower part of the face area, (2) extraction of the mouth area, (3) measurement of facial expression intensity, (4) judgment of utterance, (5) calculation of the feature parameter for facial expression strength, and (6) calculation of the feature parameter for movements of a person. In the following subsections, these six are explained in detail.

2.2. *Standardization of lower part of face area in size*

First, the face area obtained from each frame in the dynamic image is extracted by using the classifier for a front-view face included in OpenCV. In the classifier, the Haar-like feature parameter and the Adaboost algorithm for learning are used.[8] It is assumed that the distance between a subject and the camera is almost always the same during a conversation by using Skype. The face of the frame, where the face area is extracted using OpenCV and has a minimum size among those for a period in the dynamic imaging, is assumed to be the most likely front-view among those in the period[1], and that a frame is selected and used. In the present study, we set 1/3 seconds as the period. Then, the lower part of the face area extracted by the above method is standardized in length and width for extracting the mouth area. This standardization in size is performed with the aim of not only improving the performance of extracting the mouth area by OpenCV but also normalizing the feature parameter generation by using 2D-DCT (Discrete Cosine Transform) at the mouth area. Under the circumstance that the size of the face in a dynamic image cannot be kept constant every time, this standardization in size is indispensable for reliably measuring the feature parameters described in Sections 2.6 and 2.7.

2.3. *Extraction of mouth area*

Next, by using OpenCV, the mouth area is extracted as a rectangular shape. The mouth area is selected because the difference between the facial expressions of neutral and happy distinctly appears there. Fig. 1 shows an example of a face image: the lower part of the face image before and after standardization of size, and the image of the extracted mouth area.

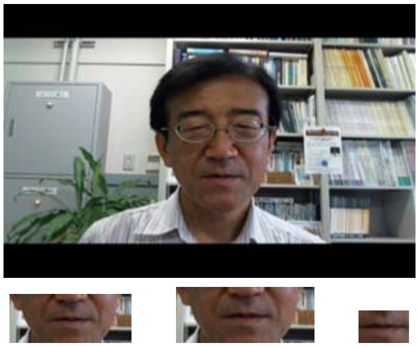

Fig. 1. Total face image (upper), lower part of face image before (lower left) and after (lower center) standardization of size, and mouth-area image extracted (lower right).

2.4. *Measurement of facial expression intensity*

For the Y component of the frame selected by the processing described above, the feature vector of facial expression is extracted in the mouth area by the use of 2D-DCT performed for each domain having 8×8 pixels.

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.[2] Then, we obtain the mean of the absolute value for each 2D-DCT coefficient component in the area of the mouth.[2] 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 the change of facial expression.[2]

2.5. *Judgment of utterance*

The sound data are smoothed and sampled 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, respectively, of the sound data value for one second under the condition of no utterance, are considered to be in the range of no utterance. 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 contain an utterance after erasing the noise.

2.6. *Feature parameter for facial expression strength*

In diagnosing a patient having dementia and/or depression, it might be useful for healthcare professionals to evaluate the strength of facial expressions by using a simple measure. Moreover, it might be more advantageous for a diagnosis of dementia and/or depression to separately evaluate the strength of the facial expression as a speaker and a listener. Therefore, we measure the feature parameter for facial expression strength as the average of facial expression intensity in the four cases of (1) both subjects A and B speak, (2) subject A speaks and subject B does not speak, (3) subject A does not speak and subject B speaks, and (4) both subjects A and B do not speak, by using the method for judging an utterance described in Section 2.5.

2.7. *Feature parameter for movements of a person*

Head movements such as a nodding during conversation might suggest a mental state and/or recognition ability of a patient having dementia and/or depression. Therefore, as the feature parameter for movements of a person, we measure the average of the absolute value of the vertical coordinate for the center of gravity of the mouth area in the relative coordinate system in the four cases described in Section 2.6. The relative coordinate system is defined by using the mouth area extracted at the starting point for measuring the feature parameter for the movements of a person. At the starting point, the height of the mouth area is set to be one and the vertical coordinate for the center of gravity of the mouth area is set to be 0 in the relative coordinate system.

3. Experiment

3.1. *Condition*

Two males (subject A in his 50s and subject B in his 20s) participated in an experiment where they had a conversation for about 70 seconds by using Skype. The videos saved by VodBurner were transformed into AVI files, and the audios saved by Tapur were transformed into WAV files. The AVI files were used for measuring feature parameters of facial expressions and movements of the subject. The WAV files were used for judgment of an utterance. The size of the image frame was 720×480 pixels, and the size of the standardized lower part of a face image was set to 240×96 pixels.

3.2. *Results and discussion*

Fig. 2 shows the mouth-area images at the starting point. The facial expression intensity changes of subjects A and B during their conversation (Fig. 3) and the changes of coordinates of the center of gravity on the mouth area (Fig. 4) are shown. In Fig. 5, face images are shown at the characteristic timing positions. Feature parameters of facial expressions and movements of the subjects are shown in Table 1. The definitions of these parameters are described in Sections 2.6 and 2.7.

As shown in Figs. 3 and 5, facial expression intensity was very sensitive for facial expression change. Both subjects A and B did not move vertically so much during conversation (Fig. 4 and Table 1). The number of mouth-area images extracted by OpenCV was increased for subject A from 196 to 197 by performing size standardization

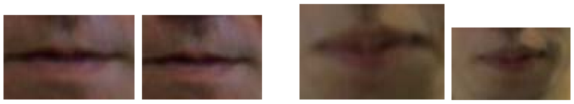

Fig. 2. Mouth-area images of subjects A (left two images) and B (right two images) at the starting point with (left side) and without (right side) size standardization before extracting the mouth area.

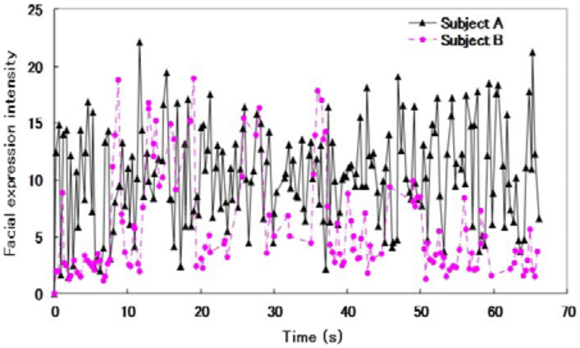

Fig. 3. Facial expression intensity change of subjects A and B during the conversation between the two subjects with size standardization before extracting the mouth area.

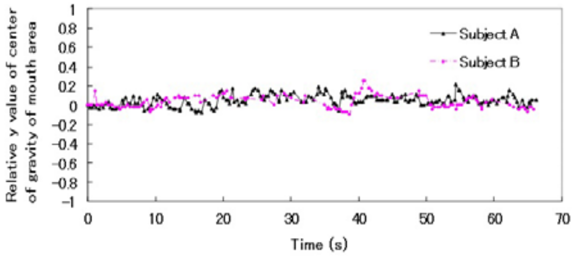

Fig. 4. Changes of coordinates of center of gravity on mouth area with size standardization before extracting mouth area.

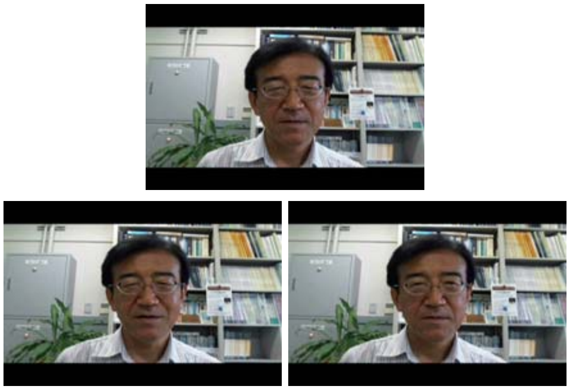

Fig. 5. Face images at characteristic timing positions for facial expression intensity value of subject A; upper: 0 (starting point), lower left: maximum, lower right: minimum except that at starting point.

Table 1. Feature parameters for facial expressions and movements.

(1) With size standardization before extracting mouth area

Subject	Utterance		Feature parameter	
	A	B	Facial expression	Movement of person
A	without	without	11.41	0.05
	with	without	9.47	0.08
	without	with	10.50	0.05
	B	A		
B	without	without	4.53	0.05
	with	without	12.51	0.07
	without	with	3.14	0.06

(2) Without size standardization before extracting mouth area

Subject	Utterance		Feature parameter	
	A	B	Facial expression	Movement of person
A	without	without	8.55	0.03
	with	without	8.50	0.04
	without	with	10.06	0.05
	B	A		
B	without	without	5.48	0.04
	with	without	10.88	0.05
	without	with	5.18	0.03

before extracting the mouth area, while it was increased for subject B from 27 to 138 by standardization. Though mouth-area images were influenced by size standardization before extracting the mouth area (Fig. 2), the feature parameter values of both facial expressions and movements of subjects were not influenced so much by size standardization before extracting the mouth area (Table 1). The value of the feature parameter for facial expression was relatively high for subject A in all three cases, whereas it was relatively high for subject B in the only case that subject B spoke and subject A did not speak (Table 1). Because the period in which both subjects A and B spoke was very short, the data are not described in Table 1.

4. Conclusion

A video is analyzed by image processing and the newly proposed feature parameters of facial expressions and movements. The experimental result shows the usefulness of the proposed method. In future work, we will develop the method for estimating the mental state and/or recognition ability of a patient by using the proposed method.

Acknowledgment

This research is partially supported by COI STREAM of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. T. Asada, Y. Yoshitomi, A. Tsuji, R. Kato, M. Tabuse, N. Kuwahara, and J. Narumoto, Method of facial expression analysis of person while using video phone, *in Proc. of Human Interface Symposium 2013* (Japan, Tokyo, 2013), pp.493-496.
2. T. Asada, Y. Yoshitomi, A. Tsuji, R. Kato, M. Tabuse, N. Kuwahara, and J. Narumoto, Facial expression analysis while using video phone, *in Proc. of Int. Conf. on Artif. Life and Robotics* (Japan, Oita, 2014), pp.230-234.
3. J. Narumoto, N. Kuwahara, Y. Yoshitomi, T. Asada, Y. Kato, H. Kamimura, K. Fukui, Development of support system for patients with dementia through teleconference system, *in Proc. of 4th World Conf. of Asian Psychiatry* (Thailand, Bangkok, 2014), pp.1-4.
4. Skype Web page, http://www.skype.com/ Accessed 5 November 2013.
5. VodBurner Web page, http://www.vodburner.com/ Accessed 11 July 2014.
6. Tapur Web page, http://www.tapur.com/jp/ Accessed 18 December 2014.
7. OpenCV Web page, http://opencv.willowgarage.com/ Accessed 11 July 2014.
8. 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.