Human-Computer Communication Using Facial Expression

Yasunari Yoshitomi
Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: yoshitomi@kpu.ac.jp
http://www2.kpu.ac.jp/ningen/infsys/English_index.html

Abstract

To develop a complex computer system such as a robot that can communicate smoothly with humans, it is necessary to equip the system with a function for both understanding human emotions and expressing emotional signals. From both perspectives, facial expression is a promising research area. In our research, we have explored both aspects of facial expression using infrared-ray images and visible-ray images and have developed a personified agent for expressing emotional signals to humans.

Keywords: Emotion, Facial expression recognition, Infrared-ray image, Facial expression synthesis, Personified agent.

1. Introduction

The goal of our study is to present a paradigm whereby a complex computer system such as a robot can cooperate smoothly with humans. To do this, the computer system must have the ability to communicate with humans using some form(s) of information transmission. Such a system must be equipped with a function for both understanding human emotions and expressing emotional signals to its human counterparts. In this regard, facial expression is a promising target for research. Accordingly, we have been investigating both aspects of facial expression.

In this paper, we describe the challenges of reaching our goal. The remainder of the paper is organized as follows: Section 2 summarizes our studies on facial expression recognition; Section 3 briefly describes our studies on human-computer-human communication via the Internet; Section 4 outlines our studies on human-computer communication; Section 5 discusses our work on integration with speech; Section 6 concludes the paper.

2. Facial Expression Recognition

2.1. *Infrared-ray image utilization*

We have developed a method for recognizing facial expressions using thermal image processing.[1] In this study, infrared-ray was used. Figure 1 shows the influence of lighting at night on a facial image using both visible-ray ((a),(c)) and infrared-ray ((b),(d)). As is evident in the figure, the visible-ray image is strongly influenced by lighting conditions, while the thermal image is unaffected. With our method, neutral, happy, surprised, and sad facial expressions were recognized with 90% accuracy.[1] Figure 2 shows examples.

2.2. *Sensor fusion*

Sensor fusion is a promising way to improve the recognition accuracy of facial expression or emotion recognition. Several studies[2,3] to improve accuracy using sensor fusion have produced promising results.

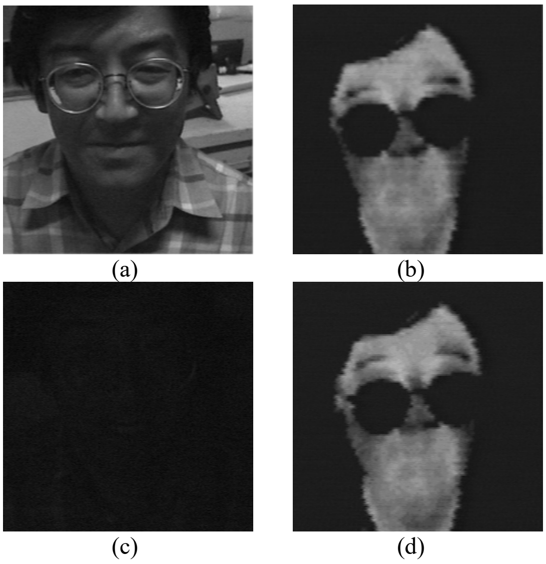


Fig. 1. Examples of face-image at night; (a) visible-ray image with lighting, (b) infrared-ray image with lighting, (c) visible-ray image without lighting, (b) infrared-ray image without lighting.[1]

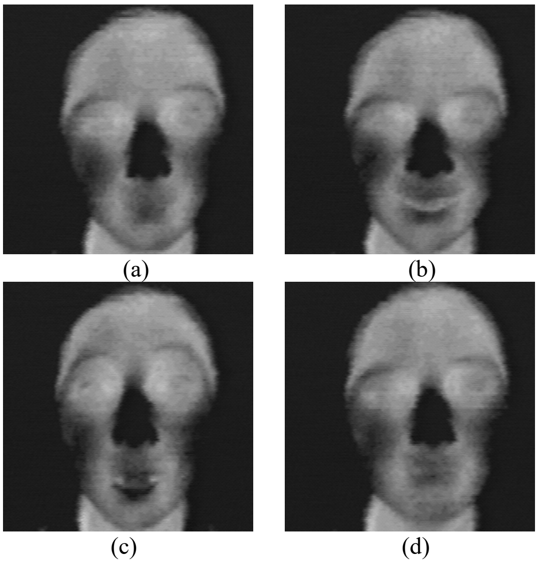


Fig.2. Examples of facial expressions; (a) neutral, (b) happy, (c) surprised, (d) sad.[1]

2.3. *Analysis for medical use*

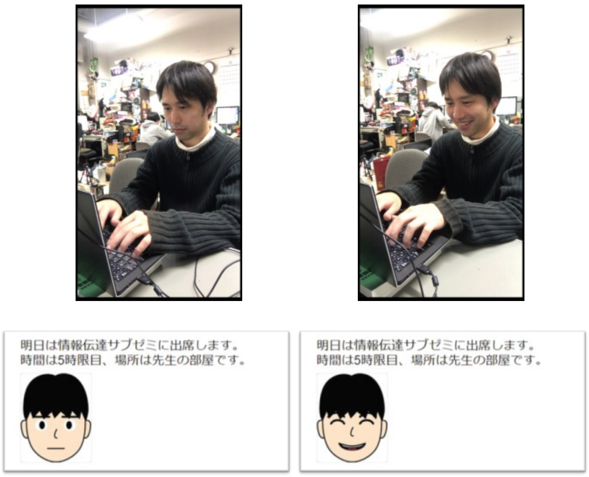
Facial expression analysis using infrared-ray images[4] or visible-images[5-7] has been successfully conducted with the goal of identifying subjects suffering from pre-stage dementia or other medical problems.

3. Human-computer-human Communication

Social network services (SNSs) have become extremely popular as communication tools on the Internet. However, while it is possible to post a message, a static image, or a moving image on a platform such as Twitter, it is difficult to communicate the actual emotions felt when writing a message or posting an image. We believe that a support system is needed to facilitate smoother communication between humans in their use of SNSs. Not having immediate and direct contact with one another risks misunderstanding, especially from an emotional point of view.

One of our studies is aimed at expressing the real emotions of individuals writing messages for posting on an SNS site by analyzing their facial expressions and visualizing them as pictographs. To this end, we have developed a real-time system for expressing emotion as a pictograph selected according to the writer's facial expression while writing a message.[8,9] We applied the system to the posting on Twitter of both a message and a pictograph.[8,9]

The lower panels in Figure 3 show the output of our system in a situation where the subject was asked to intentionally show two types of emotions—neutral or smiling—when writing the message, '明日は情報伝達システム学サブゼミに参加します。時間は5時限目、場所は先生の部屋です。' (in Japanese), which means, "I will attend the discussion section held at the professor's room in the information communication system lab in fifth period tomorrow." [8]



明日は情報伝達サブゼミに出席します。
時間は5時限目、場所は先生の部屋です。

明日は情報伝達サブゼミに出席します。
時間は5時限目、場所は先生の部屋です。

Fig. 3. Snap-shots (upper) of posting on Twitter; messages and pictographs (lower) posted on Twitter (left: neutral; right: smiling).[8]

4. Human-computer Communication

4.1. *Personified agent*

The process of agent generation in our system[10] consists of six steps: (1) creating facial expression data, (2) recording vocal utterances, (3) automatic WAVE file division, (4) speech recognition by Julius[11], (5) insertion of expressionless data, and (6) the creation of facial expression motion.

Expressive motions are generated by combining the expression data of each vowel for each utterance motion. Then, the utterance contents are input as text and used by the MikuMikuDanceAgent (MMDAgent),[12] which is a freeware animation program that allows users to create and animate movies with agents, to output synthesized voice that is then recorded by a stereo mixer inside a PC and saved as a WAVE file. Speech is recognized using a speech recognition system called Julius,[11] followed by facial expression synthesis of the agent using preset parameters depending on each vowel. Facial expression data were created with MikuMikuDance[13]. In this study, in order to generate more human-like agent facial expressions, facial expression data were created for the vowels / a /, / i /, / u /, / e /, and / o / (Fig. 4).[10] In order to create more natural agent facial expressions, processing is then performed to insert a neutral facial expression when the same vowel, for example / a /, is continuous.[10]

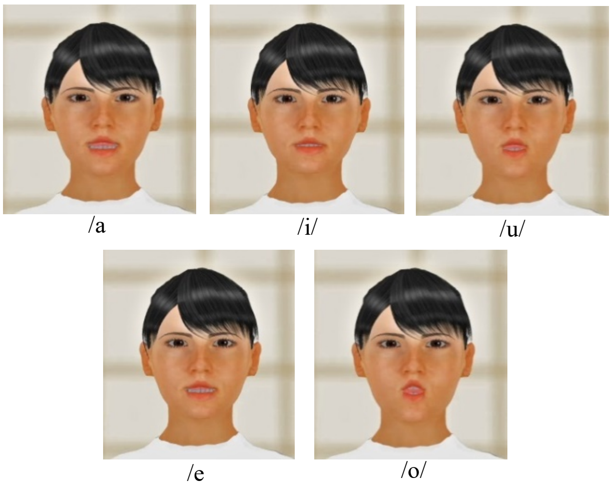


Fig. 4. Facial expression of the agent when uttering each vowel.[10]

4.2. *Human-computer communication in music recommendation*

The music recommendation module of the proposed system[14] is based on a previously proposed system[15] that uses collaborative filtering and impression words (see the paper[15] for details of the music recommendation module).

In the music-recommendation process, all user navigations are performed by the synthetic voice of the agent appearing on the PC screen facing the user. All dialogue spoken by the agent is situationally selected by the proposed system[14]. The user's answers to the questions generated by the agent are recognized using the voice recognition function of the system, and the agent motions, including facial expressions, are then generated.

Figure 5 shows two snapshots of the reaction of the agent after recognizing (a) a positive answer, i.e., the user wishes to listen to the recommended song again in the future, and (b) a negative answer, i.e., the user does not wish to listen to the recommended song in the future. In the case of (a), the agent nods twice and raises the corners of the mouth slightly, while in the case of (b), the agent also nods twice, but lowers the corners of the mouth slightly. Figure 6 shows a snapshot of the music recommendation being performed by the proposed system[14].

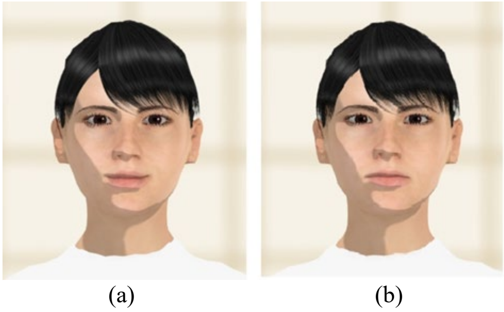


Fig. 5. Snapshots of the reaction of the agent after recognizing (a) a positive answer, and (b) a negative answer. [14]

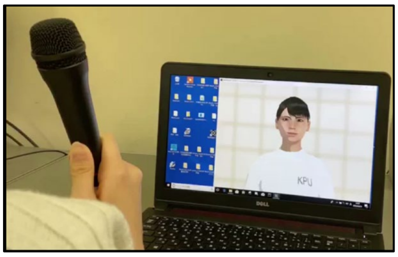


Fig. 6. Snapshot of performing song-recommendation by the proposed system.[14]

5. Integration with Speech

Utterance judgment is necessary for deciding the timing of facial expression recognition. Moreover, the mouth shape with or without an utterance influences facial expression. In our studies, the first and last vowels in an utterance such as a name were recognized for deciding the timing of the facial expression recognition.[16,17]

Speech recognition and synthesis are indispensable for human-computer communication. In particular, developing the function of emotional speech synthesis[18] offers a way to create a paradigm whereby a computer system such as a robot can work seamlessly with humans.

6. Conclusions

To develop a complex computer system such as a robot that can communicate smoothly with humans, it is necessary to equip the system with the ability to both understand human emotion and express emotional signals to humans. From both points of view, facial expression is a promising research field. In developing a method for recognizing facial expressions, we have used infrared-ray images as well as visible-ray images. For expressing emotional signals to humans, we have developed a personified agent. Developing the function of emotional speech synthesis is the next target of our studies.

Acknowledgements

We would like to thank Dr. J. Narumoto, Professor of the Kyoto Prefectural University of Medicine, Dr. M. Tabuse, Professor of the Kyoto Prefectural University, and Dr. T. Asada, Associate Professor of the Kyoto Prefectural University, for their valuable cooperation during the course of this research.

References

1. Y. Yoshitomi, N. Miyawaki, S. Tomita and S. Kimura, Facial expression recognition using thermal image processing and neural network, *in Proc. of 6th IEEE Int. Workshop on Robot and Human Communication*, (Japan, Sendai, 1997), pp. 380-385.
2. 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. of of 9th IEEE Int. Workshop on Robot and Human Interactive Communication*, (Japan, Takamatsu, 2000), pp.178-183.
3. Y. Oka, Y. Yoshitomi, T. Asada, and M. Tabuse, Emotion recognition of a speaker using facial expression intensity of thermal image and utterance time, *J. Robotics, Networking and Artif. Life* **3**(3) (2016) 148-151.
4. Y. Yoshitomi, T. Asada, R. Kato, and M.T abuse, Method of facial expression analysis using video phone and thermal image, *J. Robotics, Networking and Artif. Life* **1**(1) (2014) 7-11.
5. T. Asada, Y. Yoshitomi, R. Kato, M. Tabuse, and J. Narumoto, Quantitative evaluation of facial expressions and movements of persons while using video phone, *J. Robotics, Networking and Artif. Life* **2**(2) (2015) 111-114.
6. T. Asada, Y. Yoshitomi, A. Tsuji, R. Kato, M. Tabuse, N. Kuwahara, and J.Narumoto, Facial expression analysis while using video phone, *J. Robotics, Networking and Artif. Life* **2**(4) (2016) 258-262.
7. R. Shimada, T. Asada, Y. Yoshitomi, and M. Tabuse, Real-time system for horizontal asymmetry analysis on facial expression and its visualization, *J. Robotics, Networking and Artif. Life* **6**(1) (2019) 7-11.
8. Y. Yoshitomi, T. Asada, K. Mori, R. Shimada, Y. Yano, and M. Tabuse, Facial expression analysis and its visualization while writing messages, *J. Robotics, Networking and Artif. Life* **5**(1) (2018) 37-40.
9. T. Asada, Y. Yano, Y. Yoshitomi, and M. Tabuse, A system for posting on SNS portrait selected using facial expression analysis while writing message, *J. Robotics, Networking and Artif. Life* **6**(3) (2019) 199-202.
10. T. Asada, R. Adachi, S. Takada, Y. Yoshitomi, and M. Tabuse, Facial expression synthesis using vowel recognition for synthesized speech, *in Proc. 2020 Int. Conf. on Artificial Life and Robotics*, ed. M. Sugisaka (Japan, Beppu, 2020), pp.398‐402.
11. Julius, http://Julius.osdn.jp/, Accessed 24 November 2020.
12. MMDAgent, http://www.mmdagent.jp/ Accessed 24 November 2020.
13. MikuMikuDance, https://sites.google.com/view/vpvp/, Accessed 29 November 2020.
14. A. Matsui, M. Sakurai, T. Asada, and M. Tabuse, Music recommendation system driven by interaction between user and personified agent using speech recognition, synthesized voice and facial expression, *in Proc. 2021 Int. Conf. on Artificial Life and Robotics*, ed. M. Sugisaka (Japan, 2021), in press.
15. S. Yoshizaki, Y. Yoshitomi, C. Koro, and T. Asada, Music recommendation hybrid system for improving recognition ability using collaborative filtering and impression words, *J. Artif. Life and Robotics* **18**(1-2) (2013) 109–116.
16. 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.
17. 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.
18. R. Makino, Y. Yoshitomi, T. Asada, and M. Tabuse, Speech synthesis of emotions in a sentence using vowel features, *J. Robotics, Networking and Artif. Life* **7**(2) (2020) 107-110.