Evaluation of Superimposed Self-character Based on the Detection of Talkers' Face Angles in Video Communication

Yutaka Ishii and Tomio Watanabe

Faculty of Computer Science and System Engineering,
Okayama Prefectural University, Japan
{ishii,watanabe}@cse.oka-pu.ac.jp

Abstract. We build upon an embodied video chat system, called E-VChat, in which an avatar is superimposed on the other talker's video images to improve the mutual interaction in remote communications. A previous version of this system used a headset-type motion capture device. In this paper, we propose an advanced E-VChat system that uses image processing to sense the talker's head motion without wearing sensors. Moreover, we confirm the effectiveness of the superimposed avatar for face-to-face communication in an experiment.

Keywords: Multimodal interaction, Human Interface.

1 Introduction

The effectiveness of modern video-based teleconference systems is confirmed by their widespread commercial use [1], [2], [3], [4]. Video images help individuals to directly observe nonverbal information, such as nodding, gestures, and facial expressions. As such, video is considered to be a very useful communication media. However, remote talkers have difficulty interacting because they do not share the same communication space; hence, they rely only on a video image. Some telecommunication systems have been developed using computer generated (CG) characters in cyber-space, or avatars, which allow remote talkers to communicate through a common virtual space [5], [6], [7]. Recently, new devices and methods have been proposed for movement of the avatars. Takahashi et al. suggested a head motion detection method that uses an active appearance model (AAM) that is sensitive to eye blinks [8]. A virtual communication system for human interaction, called "VirtualActor," uses a human avatar that represents the upper body motion of the talker [9]. This system experimentally demonstrated the effectiveness of communication through avatar embodiment.

A more effective remote communication system can be developed by allowing talkers to observe each other's nonverbal information, such as facial expressions. To take advantage of both avatars and video images, we developed an embodied video communication system in which VirtualActor is superimposed on each partner's video image in a virtual face-to-face scene [10]. Moreover, we developed a headset motion-capture device that uses an acceleration sensor and gyro sensor to track the talker's head movements directly. The device allows an avatar to mimic the talker's

S. Yamamoto (Ed.): HIMI/HCII 2013, Part I, LNCS 8016, pp. 431–438, 2013.

motion and automatically move in response to the on-off pattern of the talker's voice. The combined system was implemented in a prototype communication system called "Enhanced-VideoChat (E-VChat)" [11].

In this paper, we propose an advanced E-VChat system that uses image processing to track the talker's head motion and facial angles, thereby avoiding the need for a wearing sensor and expanding the practicality of the system. The effectiveness of the new system is confirmed experimentally.

2 Concept of E-VChat System

Figure 1 shows the E-VChat system concept. Talkers communicate using the video image of the other. In the E-VChat system, a voice-driven substitute character is superimposed on a video image of the partner. The character's motions are automatically generated based on the talker's voice and head motions as measured by a motion capture device.

Figure 2 depicts how a talker's gaze line is dependent on camera position. In general, the web camera is placed on the periphery of the monitor, and the talker casts his/her eyes outside of the monitor. By including an embodied avatar on the talker's screen, each talker can observe a virtual face-to-face interaction between the self-avatar and an image of the other talker.

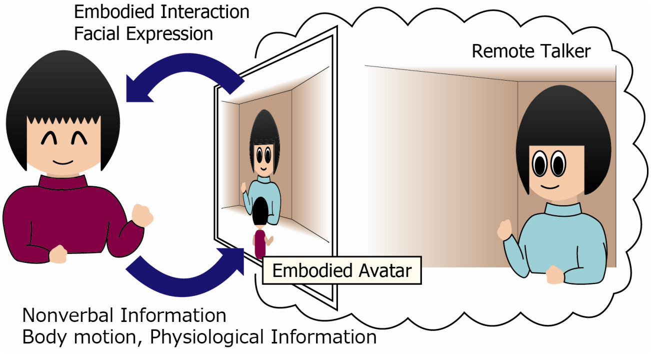

Fig. 1. Concept of the E-VChat System

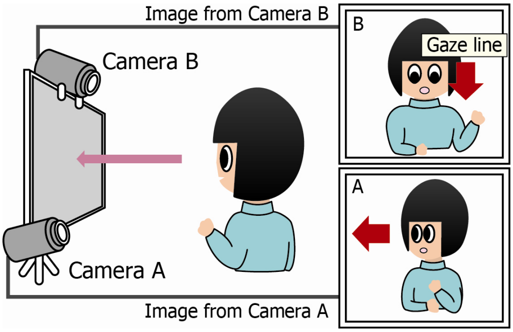

Fig. 2. Talker's gaze lines based on camera positions

This system is unique in that the talker's avatar is only visible to the talker. On the screen, the talker observes the avatar overlaid on the other talker's image, but the other talker is unaware of the avatar. We expect to extend this effect for all communication between remote talkers.

3 System Configuration

Talkers using the original E-VChat systems were required to wear a headset device or magnetic sensors to track the talker's motion. Our simple device using an acceleration sensor and gyro sensor also had problems for the practical or universal use, and the magnetic sensors for detecting talker's motions had some problems, such as a sense of restraint owing to the sensor cables and the sensors' lack of port-ability. To address this issue, the advanced E-VChat system uses image processing to track the talker's head motion. A description of the image processing algorithm is provided here.

3.1 Character Motions Generated Automatically on the Basis of Speech Input

An avatar motion generation method that is based on the talker's voice has already been developed [12]. Nonverbal actions that express a talker's intention are important in serious situations such as negotiations, counseling, and agreements. The avatar's head motion, which plays an important role in communication, synchronizes with the talker's motion to facilitate communication.

The Moving-Average (MA) model, which times nodding on the basis of a talker's voice data, is used to auto-generate avatar motions for the "listener" [12]. The MA model estimates nodding timing from a speech on-off pattern, using a hierarchy model consisting of a macro stage and a micro stage. When *Mu*(i) exceeds a threshold value, the nodding value, $M(i)$, is estimated as the weighted sum of the binary speech signal, $V(i)$. Avatar body movements are introduced when speech input timing exceeds a body threshold. The body threshold is set lower than that of the nodding prediction of the MA model, which is expressed as the weighted sum of the binary speech signal to nodding.

$$M_u(i) = \sum_{j=1}^{J} a(j)R(i-j) + u(i) \tag{1}$$

$$R(i) = \frac{T(i)}{T(i) + S(i)} \tag{2}$$

$a(j)$: linear prediction coefficient
$T(i)$: talkspurt duration in the i-th duration unit
$S(i)$: silence duration in the i-th duration unit
$u(i)$: noise

$$M(i) = \sum_{k=1}^{K} b(j)V(i-j) + w(i) \tag{3}$$

$b(j)$: linear prediction coefficient
$V(i)$: voice
$w(i)$: noise

The MA model of the "speaker" allows the avatar's head and body motions to be linked to the on-off pattern of speech.

3.2 Character Motions Measured by a Headset Motion-Capture Device

More effective communication would be supported by not only the auto-generated motions based on the voice input as described in the previous section but also talkers' own measured motions for their intentions by a motion-capture device (Kinect for Windows L6M-00005). The talker's head motions are detected based on three-axis angles and three-dimensional positions. The positions are measured using a depth sensor. The face angle detection range is shown in Figure 3. The avatar's motions are represented based on measured head motions.

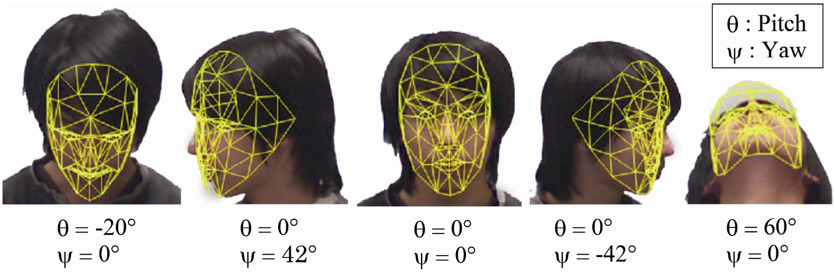

Fig. 3. Detection range of the face angle

4 Communication Experiment

4.1 Experimental Setup

A communication experiment was conducted in three modes: mode A used only head-motion, mode B used auto-generated motion based on speech input, and mode C used both head motion and auto-generated motion. A population of 12 pairs of subjects was evaluated. The talkers in each pair were familiar with one another and were observed using the system in unrestrained conversations. Subjects could select from seven different types of avatar, such as human, robot, or animal. An example of a communication scene using the E-VChat is shown in Figure 4. The experimental setup is shown in Figure 5.

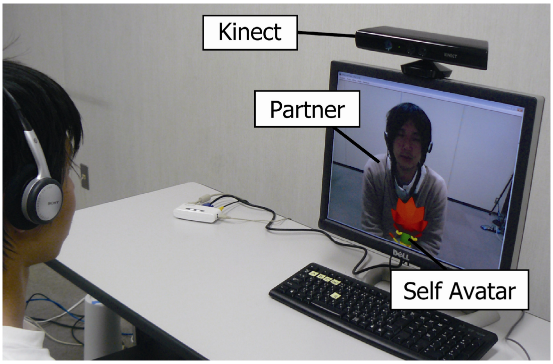

Fig. 4. Communication scene using the E-VChat

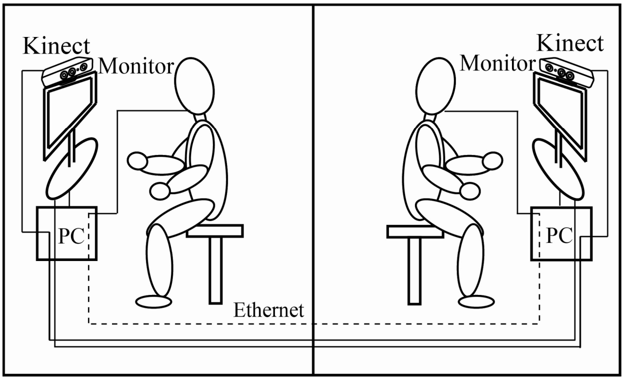

Fig. 5. Experimental setup

4.2 Result of Sensory Evaluation

The results of the paired comparisons of the three modes, in terms of talker preference, are shown in Table 1. Figure 6 shows the Table 1 data calculated using the Bradley-Terry model given in Equation (4). Mode C, which used both head motion and auto-generated motion, received the most positive talker feedback.

Table 1. Result of pair comparison in the communication experiment

	A	B	C	Total
A		11	3	14
B	13		1	14
C	21	23		44

$$P_{ij} = \frac{\pi_i}{(\pi_i + \pi_j)} \tag{4}$$

$$\sum_i \pi_i = const.(= 100)$$

(π_i : intensity of *i*, P_{ij} : probability of judgment that *i* is better than *j*.)

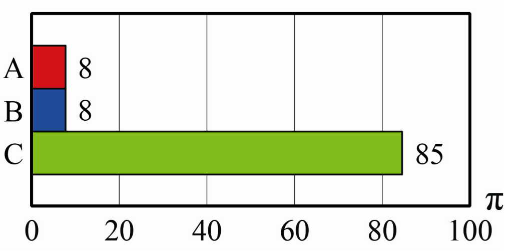

Fig. 6. Preference based on the Bradley-Terry model

Six additional factors were evaluated in each of the three modes using a seven-point scale ranging from -3 (lowest) to 3 (highest) with 0 denoting a moderate score. The six factors were "Enjoyment: Did you enjoy the conversation using the system?," "Sense of unity: Did you have a sense of unity with your partner?," "Ease of talking: Did you feel it was easy to talk using the system?," "Relief: Were you able to communicate with relief?," "Like: Do you like this system?," and "Preference: Would you like to use this system?"

For readability, the means and standard deviations of the questionnaire results are shown in Figure 7. The significant differences between each of the three modes were obtained by administering Friedman's test, in which a significance level of 1% was obtained for all factors. Significant differences were also obtained by administering the Wilcoxon's rank sum test for multiple comparisons. A significance level of 1% was obtained for all factors between Modes A and B and for the "Sense of unity,"

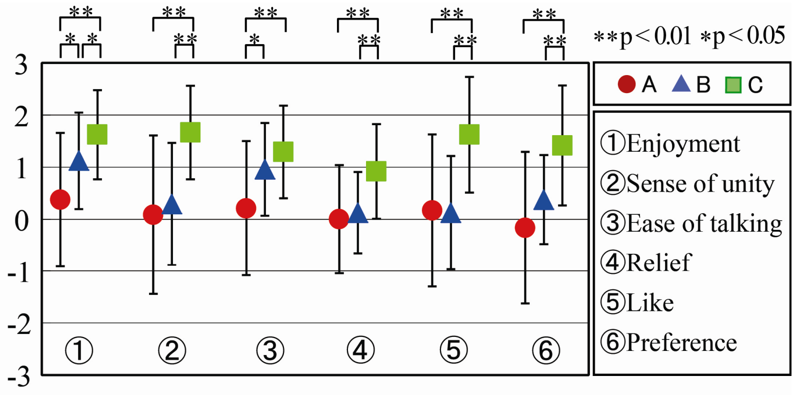

Fig. 7. Seven points bipolar rating

"Relief," "Like," and "Preference" factors when comparing Modes B and C. A significance level of 5% was obtained between Modes B and C for the "Enjoyment" factor. Mode C was positively evaluated as the paired comparison.

5 Conclusion

In this paper, we described an advanced E-VChat system that allows talkers to smoothly communicate using nonverbal information via a self-avatar that is displayed alongside video images. The avatar motions are based on image processing. The effectiveness of self-avatar using talker's head motion and auto-generated motion based on the speech input was confirmed in a communication experiment that evaluated 12 pairs of subjects in three separate communication modes: one mode in which only head-motion was used, another mode in which motion was auto-generated based on speech input, and a final mode that used both head-motion and auto-generated motion based on speech input.

References

1. Sellen, A.J.: Speech Patterns In Video-Mediated Conversations. In: Proc. of CHI 1992, pp. 49–59. ACM (1992)
2. Buxton, W.A.S.: Living in Augmented Reality: Ubiquitous Media and Reactive Environments. In: Finn, E.K., et al. (eds.) Video-Mediated Communication. Computers, Cognition, and Work, pp. 363–384 (1997)
3. Ishii, R., Ozawa, S., Mukouchi, T., Matsuura, N.: MoPaCo: Pseudo 3D Video Communication System. In: Salvendy, G., Smith, M.J. (eds.) HCII 2011, Part II. LNCS, vol. 6772, pp. 131–140. Springer, Heidelberg (2011)
4. Kim, K., Bolton, J., Girouard, A., Cooperstock, J., Vertegaal, R.: TeleHuman: Effects of 3D Perspective on Gaze and Pose Estimation with a Life-size Cylindrical Telepresence Pod. In: Proc. of CHI 2012, pp. 2531–2540 (2012)
5. Cassel, J., et al.: An Architecture for Embodied Conversational Characters. In: Proc. of WECC 1998, pp. 21–29 (1998)
6. Yahoo! Inc., Yahoo! Avatar, http://avatars.yahoo.com/
7. Linden Lab, Second Life, http://secondlife.com/
8. Takahashi, K., Mitsukura, Y.: Eye Blink Detection Using Monocular System and its Applications. In: Proc. of 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2012), pp. 743–747 (2012)
9. Ishii, Y., Watanabe, T.: An Embodied Avatar Mediated Communication System with VirtualActor for Human Interaction Analysis. In: Proc. of the 16th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2007), pp. 37–42 (2007)
10. Ishii, Y., Watanabe, T.: An Embodied Video Communication System in which Own VirtualActor is Superimposed for Virtual Face-to-face Scene. In: Proc. of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004), pp. 461–466 (2004)

11. Ishii, Y., Watanabe, T.: A Video Communication System in Which a Speech-driven Embodied Entrainment Character Working with Head Motion is Superimposed for a Virtual Face-to-face Scene. In: Proc. of 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2012), pp. 191–196 (2012)
12. Watanabe, T., Okubo, M., Nakashige, M., Danbara, R.: InterActor: Speech-Driven Embodied Interactive Actor. International Journal of Human-Computer Interaction 17(1), 43–60 (2004)