Combination of Local Interaction with Remote Interaction in ARM-COMS Communication

Teruaki Ito[1(✉)], Hiroki Kimachi[2], and Tomio Watanabe[3]

[1] Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, 111 Kuboki, Soja-shi, Okayama 719-1197, Japan
tito@ss.oka-pu.ac.jp
[2] Graduate School of Advanced Technology and Science, Tokushima University, 2-1 Minami-Josanjima, Tokushima 770-8506, Japan
c501732024@tokushima-u.ac.jp
[3] Faculty of Computer Science and System Engineering, Okayama Prefectural University, 111 Tsuboki, Souja, Okayama 719-1197, Japan
watanabe@cse.oka-pu.ac.jp

Abstract. ARM-COMS detects the orientation of a human subject face by the face-detection tool based on an image processing technique, and mimics the head motion of a remote partner during video conversation in an effective manner to enhance entrainment as reported before. However, ARM-COMS does not make any appropriate reactions if a communication partner speaks without move in video communication. Therefore, audio signal from the human subject is another option to use as a driving force of ARM-COMS to enhance the physical entrainment. In this study, a configuration of voice signal-based local interaction subsystem was implemented. Using this subsystem, handing of two types of individual input signals were studied: one is from the head-motion image of a remote partner, and the other one is from the combination of voice signals of a local user and its remote partner. This paper presents how the combination of remote interaction and local interaction was implemented in ARM-COMS communication, and discusses the feaibility of this approach.

Keywords: Embodied communication · Augmented tele-presence robotic arm · Face detection · Audio interaction · Combination of remote and local interaction

1 Introduction

A smartphone-based video communication tool is now one of the convenient popular tools freely available to many people. Supported by ICT (Information and Communication Technology) technologies, further enhancement of better quality in communication is being expected. In the meantime, this tool addresses the two types of critical issues, which are the lack of tele-presence feeling and the lack of relationship feeling in remote video communication as opposed to a typical face-to-face communication.

Several ideas of robot-based remote communication systems have been proposed as one of the solutions to the former issue; these robots include physical telepresence

S. Yamamoto and H. Mori (Eds.): HCII 2019, LNCS 11570, pp. 347–356, 2019.
https://doi.org/10.1007/978-3-030-22649-7_28

robots. Anthropomorphization is another new idea to show the telepresence of a remote person in communication system. Remote communication can be basically supported by the primitive functions of physical tele-presence robots, such as a face image display of the operator, as well as tele-operation function such as remote-drivability to move around, or tele-manipulation. However, there are still an open issue to be studied to narrow the gap between robot-based video communication and face-to-face one.

The second issue in the lack of relationship-type feeling in remote video communication is another big challenge. Recently, an idea of robotic arm-type systems draws researchers' attention. For example, Kubi, which is a non-mobile arm type robot, allows the remote user to "look around" during video communication by way of commanding Kubi where to aim the tablet with an intuitive remote control over the net. Furthermore, an idea of enhanced motion display has also been reported to show its feasibility over the conventional display. However, the usage of the human body movement of a remote person as a non-verbal message is still an open issue.

This research proposes an idea of motion-enhanced display that utilizes the display itself as the communication media, which mimics the motion of human head to enhance presence in remote communication. The idea has been implemented as an augmented tele-presence system called ARM-COMS (ARm-supported eMbodied COmmunication Monitor System). ARM-COMS is a solution to this second issue [3].

In order to mimic the head motion using the display, ARM-COMS detects the orientation of a face by face-detection tool based on an image processing technique. Even though ARM-COMS mimics the head motion of a remote partner, a reaction delay was recognized in communication experiments. Furthermore, ARM-COMS does not make appropriate reactions if a communication partner speaks without move in video communication, which has been often recognized during communication experiments. In order to solve these problems, this study proposes a voice signal during the video conversation as the driving force of local interaction and/or remote interaction. Therefore, the combination of local interaction activated by voice signal of a local user with the remote interaction activated by head motion of a remote user is the challenge of this study.

First, this paper overviews ARM-COMS, including its basic concept, basic functions, and experimental results conducted so far. Then, the paper focuses on the issue of communication without move, which has been recognized by the use of ARM-COMS. Configuration of voice signal-based local interaction prototype system will be presented to show its implementation. Handing of two types of individual input signals, one is from the head-motion of a remote partner, and the other one is from the combination of voice signals of a local user and a remote user, will also be shown to implement the combination of remote interaction and local interaction in ARM-COMS communication. Concluding remarks with some discussions will be given in the final part of this paper.

2 System Overview and Network Configuration of ARM-COMS (ARm-Supported eMbodied COmmunication Monitor System)

2.1 Basic System Overview of ARM-COMS

ARM-COMS (ARm-supported eMbodied COmmunication Monitor System) is composed of a tabletPC and a desktop robotic arm. The table PC in ARM-COMS is a typical ICT (Information and Communication Technology) device and the desktop robotic arm works as a manipulator of the tablet, of which position and movements are autonomously manipulated based on the behavior of a human user who communicates with remote person through ACM-COMS. This autonomous manipulation of ARM-COMS is controlled by the head movement, which can be recognized by a general USB camera.

Considering the two issues mentioned in the introduction section, this paper focuses on the nodding motion as a non-verbal message contents in remote communication using ARM-COMS. Figure 1 shows the system overview of ARM-COM for the experiment in this study. Face detection procedure of a prototype of ARM-COMS is based on the algorithm of FaceNet [6], which includes image processing library OpenCV 3.1.0, machine learning library dlib 18.18, and face detection tool OpenFace which were installed on a control PC with Ubuntu 14.04 as shown in Fig. 1. Using the input image data from USB camera, landmark detection is processed.

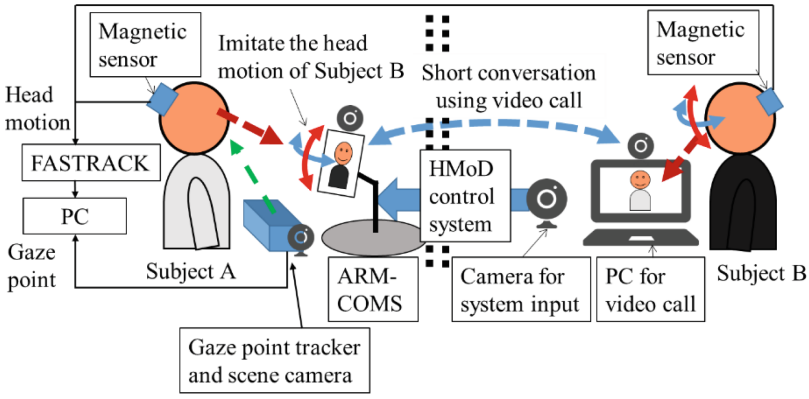

Fig. 1. Basic system configuration for ARM-COMS experiments

2.2 Network Configuration Overview of ARM-COMS

ARM-COMS is configured to implement network communication. The system is composed of various sensors to collect data, MQTT broker server, calculation server, database server, web server, client PC and application PC.

MQTT communication [5] is based on the combinaiton of publisher, MQTT server, and subscriber. Publisher defines each message as a topic and delivers it to the MQTT

broker, and then is transfered to the subscriber, which is illustraed in Fig. 2. The subscriber selects a message based on its topic and receives only the message which mathes the selected topic. Eash message is specified as three types of QoS (Quality of Service). QoS0 is not guaranteed to be delivered. QoS1 is to be sent at least one time, which is quick to be delivered if it works fine but its delivery would be without gurantee. QoS2 is guranteed to be delibered.

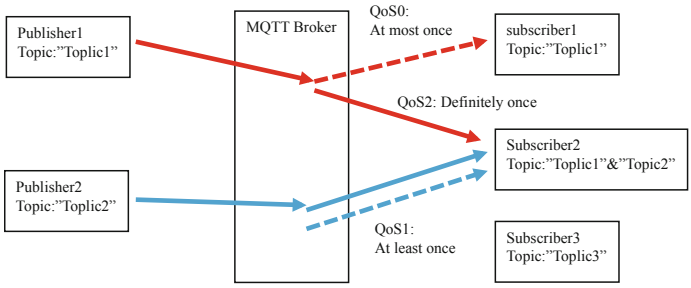

Fig. 2. MQTT communication process

Head motion of Subject A is used as a non-verbal communication to ARM-COMS which interact with Subject B. Video commmunication itself was performed by a typical software (for example, Skype) [1]. However, the head motion image data is processed by the face detection algorithm, which was used to trigger the motion of ARM-COMS installed at the site of subject B in Fig. 3.

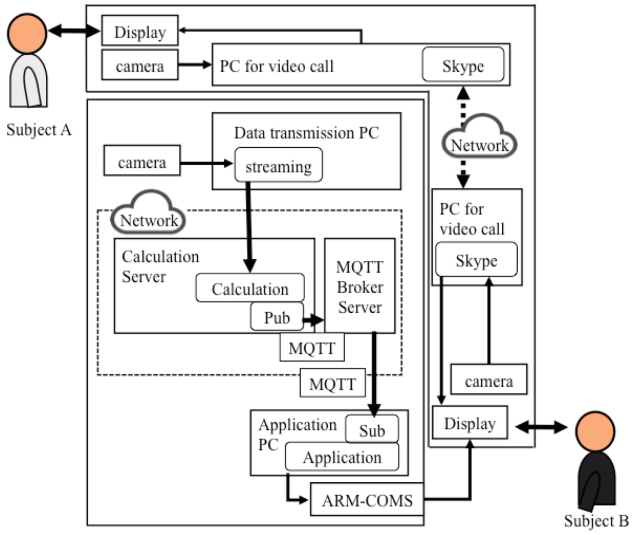

Fig. 3. Network-based configuration of ARM-COMS communication

3 Experimental Configurations of ARM-COMS System

3.1 Image-Based Interaction Subsystem

Using BGR image of human subject face collected from a USB video camera, the orientation of the subject face is calculated by OpenFace tool, which uses Contrrained Local Network Field (CLNF) composed of point distribution mode, patch expert, and fitting. In order to use this tool, the following data processing algorithm is conducted. First, a video image of Mjpeg-streamer from the USB video camera is streamed in and converted it from color image to black-and-white image by OpenCV tool. Haar Cascade method is used to detect the face area, from which feature point of 68 landmarks are extracted using dlib library tool as shown in Fig. 4.

Fig. 4. Image-based processing to determine the face orientation

Now, an image analysis for face detection is conducted by Haar Cascade face detector, which uses difference in brightness using a variety size of rectangles as shown in Fig. 5. Then 68 landmarks are defined using dlib library, and orientation of subject head is estimated by OpenFace tool. Using this orientation data, ARM-COMS can be controlled as head-up as shown in Fig. 5 and head-down shown in Fig. 6.

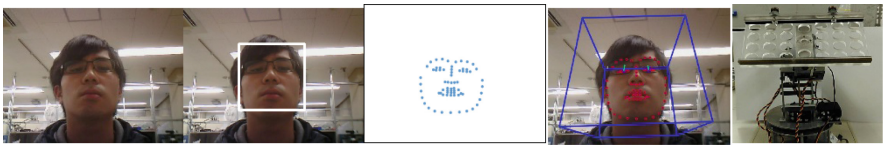

Fig. 5. Head-up detection for ARM-COMS control

Figure 7 shows the overview of ARM-COMS, which mimics the head motion of a human subject using the robotic arm of ARM-COMS in the image-based interaction subsystem.

When Fig. 8 shows the time delay comparison between standalone and network configurations, which was measured by the experimental setup illustrated in Fig. 1. This graph shows no significant difference between the two difference environment, which means that the experimental setup was appropriately configured.

Fig. 6. Head-down detection for ARM-COMS control

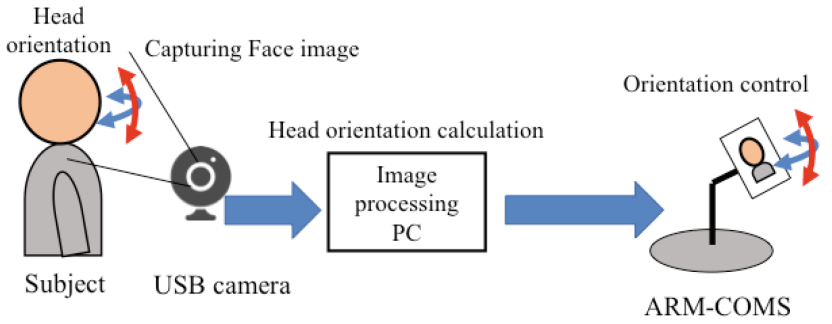

Fig. 7. ARM-COMS control to mimic head motion

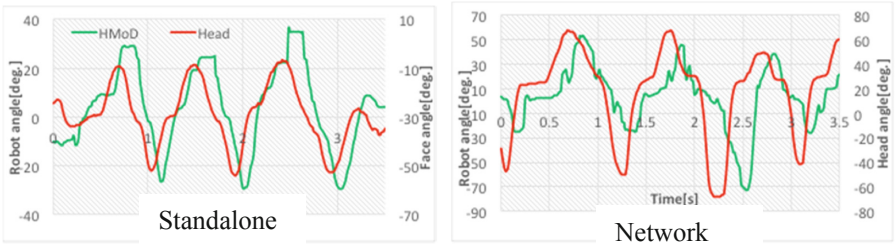

Fig. 8. Time delay comparison between standalone and network configurations for ARM-COMS

3.2 Combination of Audio and Video Interaction Subsystem

Combination of audio from local interaction and video from remote interaction was preliminaly implemented by the simple addition of the two signals as shown in Fig. 9. Pan angle and title angle generated from video signales are defined as *Mimic*p*(c)* and *Tiltc*t*(c)*. Then the output nodding angle of ARM-COMS will be given as scheme (**1**). Figure 9 also shows the combined signals to be used to control ARM-COMS.

$$\begin{cases} Pan(t) = Mimic_p + Nod(t) \\ \text{Tilt}(t) = Mimic_t(t) \end{cases} \tag{1}$$

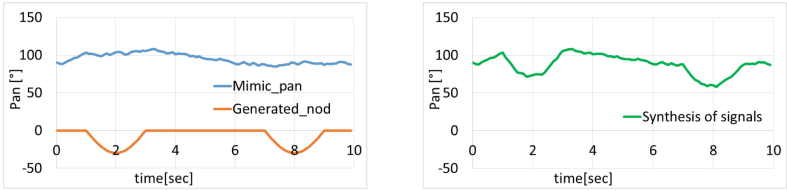

Fig. 9. Synthesis of nodding signals based on the combination of local and remote signals

3.3 Experimental Configuration for the Combination of Remote and Local Interaction

Based on the system configuration combined with image-based interaction and audio-based interaction, a new experimental environment was setup as shown in Fig. 10. Since image-based interaction was already implemented and tested so far, this configuration was based on audio signals only. A local user talks over the network with its remote partner using a video communication software, or Skype. ARM-COMS is setup in a local site only. When a local user talks, ARM-COMS detects the interval of the voice and makes nodding using the image-based interaction subsystem, which is supposed to enhance physical entrainment [7]. During that interaction, its remote user says "Yes" to interject the talk, which overrides the nodding of ARM-COMS.

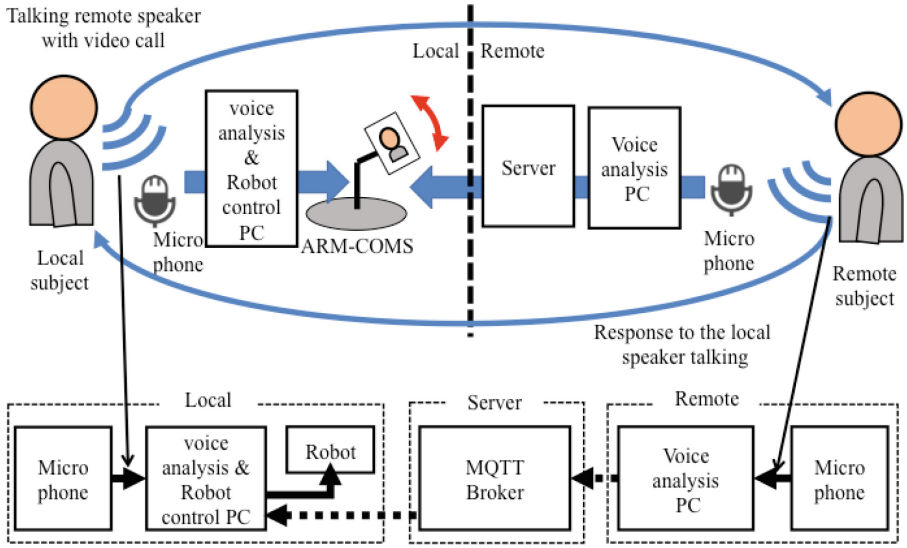

Fig. 10. Experimental setup for remote and local interaction

Figure 11 shows the experiment scene where audio signals and robot motion were recorded as shown in Fig. 12. In this example, nodding angle in local interaction is smaller than the nodding angle in remote interaction, which can be seen from the graph.

The scenario of this experiment was as follow: (a) Subject A (local) read 1 min. manuscript to Subject B through ACM-COMS, which makes local interaction based on the voice signals given by Subject A. Subject B listen to the talk of Subject A on a remote site, and says "Yes" to show that Subject B is listening to Subject A, which initiates the remote interaction in ARM-COMS with Subject A.

Fig. 11. Experiment scene under local and remote interactions

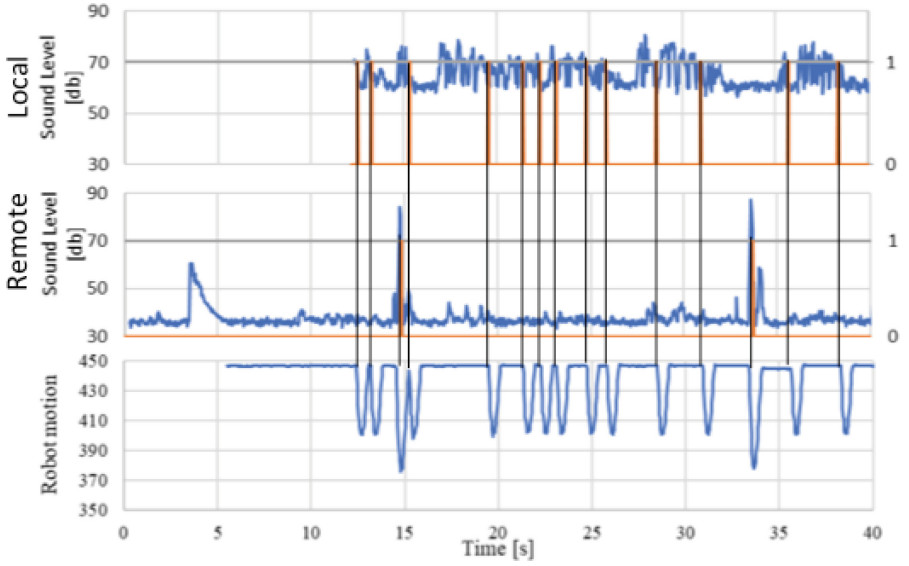

Fig. 12. Local interaction, remote interaction and robot interaction

3.4 Results and Discussion

ARM-COMS mimics the head motion of a remote subject during video conversation. Image-based interaction subsystem was implemented and evaluated [4] by the motion sensors [2] to calculate the time delay. The delay was ave. 120 [ms] for standalone environment and 210 [ms] for network environment. The delay in standalone environment was due to the process time of 100 [ms] in 10 [frames/sec] and physical motion delay in operation. The delay was ave. 209 [ms] in network environment, which is due to the streaming delay and MQTT communication delay.

Audio-based interaction subsystem was also implemented and tested to see if it works appropriately. Audio signals could be given by a local user or by a remote user. In either case, audio-based interaction worked fine to promote physical entrainment with ARM-COMS.

Combination of local and remote interaction was implemented using a two types of input signals, one of which comes from the local user and the other one of which comes from the remote user. The nodding angles could be the same or different. According to some preliminary experiments, it was recognized that the different angle was better to distinguish the local interaction from that of the remote partner. However, further studies and experiments are required to make accurate analysis.

4 Concluding Remarks

ARM-COMS detects the orientation of a subject face by the face-detection tool based on an image processing technique, and mimics the head motion of a remote partner in an effective manager as reported before. However, ARM-COMS does not make any appropriate reactions if a communication partner speaks without move in video communication. Therefore, audio signal is another option to use as a driving force of ARM-COMS. Configuration of voice signal-based local interaction subsystem was presented to show how it was implemented. Using this subsystem, handling of two types of individual input signals, one is from the head-motion image of a remote partner, and the other one is from the combination of voice signals of a local user and the remote partner, was presented to show how the combination of remote interaction and local interaction was implemented in ARM-COMS communication. This papers mainly focuses on the system implementation and only a small number of user experiments. For future work, this research will evaluate the effectiveness of the idea to find appropriate system parameters.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Numbers JP16K00 274. The author would like to acknowledge all members of Collaborative Engineering Labs at Tokushima University, and Center for Technical Support of Tokushima University, for their cooperation to conduct the experiments.

References

1. Bertrand, C., Bourdeau, L.: Research interviews by Skype: a new data collection method. In: Esteves, J. (ed.) Research Methods, pp. 70–79. IE Business School, Spain (2010)
2. FASTRK. http://polhemus.com/motion-tracking/all-trackers/fastrak
3. Ito, T., Watanabe, T.: Motion control algorithm of ARM-COMS for entrainment enhancement. In: Yamamoto, S. (ed.) Human Interface and the Management of Information: Information, Design and Interaction. LNCS, vol. 9734, pp. 339–346. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40349-6_32
4. Krafka, K., et al.: Eye Tracking for everyone. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
5. Light, R.: Mosquitto: server and client implementation of the MQTT protocol. J. Open Source Softw. **2**(13), 265 (2017). https://doi.org/10.21105/joss
6. Schoff F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: IEEE Conference on CVPR 2015, pp. 815–823 (2015)
7. Watanabe, T.: Human-entrained embodied interaction and communication technology. In: Fukuda, S. (ed.) Emotional Engineering, pp. 161–177. Springer, London (2011). https://doi.org/10.1007/978-1-84996-423-4_9