MM-Space: Recreating Multiparty Conversation Space by Using Dynamic Displays

Kazuhiro Otsuka

Abstract

This article introduces our system, called MM-Space, which can recreate a multiparty conversation space at a remote place. This system features a novel visualization scheme that represents human head motions by controlling the poses of the screens displaying human facial images. This physical augmentation helps viewers understand more clearly the behaviors of remote conversation participants such as their gaze directions and head gestures. The viewer is expected to experience a more realistic feeling of the presence of the remote people.

1. Introduction

Face-to-face conversation is one of the most basic forms of communication in daily life, and group meetings are used for conveying and sharing information, understanding others' intentions and emotions, and making decisions. To support communication among remote places, videoconferencing systems have been developed and are widely used. However, they still feel unnatural and uncomfortable. To resolve the problems that have arisen in communications between remote places, which may be not only spatially but also temporally separated, NTT Communication Science Laboratories believes that it is important to deeply understand the mechanism of human-to-human communication and answer questions such as "How do we communicate with each other and what kinds of messages are exchanged by what types of behaviors?" On the basis of this concept, my colleagues and I have been conducting conversation scene analysis for multiparty face-to-face communications [1]. We are trying to extend it toward designing better future communication systems, and we have begun representation/visualization research on multimodal telecommunication and telepresence environments. As the first step, we targeted the problem of reconstructing/recreating the conversation space of a conversation that was originally held at a different location and different time and enabling viewers to visualize the conversation scene as if the people were talking in front them. This article overviews our novel representation scheme and a prototype system after reviewing some of the background.

2. Research progress in conversation scene analysis

In face-to-face conversations, people exchange not only verbal information, but also nonverbal information expressed by eye gaze, facial expressions, head motion, hand gestures, body posture, prosody, etc. Psychologists have suggested that such nonverbal information and behaviors play important roles in human communications [2]. Conversation scene analysis aims to understand human communication though these types of nonverbal information, which is captured by multimodal sensing devices such as cameras and microphones. The goal is to provide an automatic description of conversation scenes in terms of 6W1H, namely who, when, where, whom, what, why, and how. By combining some 6W1H information, we can define a number of problems from low-level (close to physical behavior) ones to high-level (contextual and social level) ones.

Let us consider some examples that NTT Communication Science Laboratories (CS Labs) has targeted.

"Who is speaking when?" is the most essential question: it is called speaker diarization [3]. The estimation of "Who looks at whom and when?" is also called the problem of the visual focus of attention [4], [5]. "Who is talking to whom and when?" is a question about the conversation structure [4]. "Who responds to whom and how?" is related to the problem of interaction structure estimation [6]. As a higher-level problem, "Who feels empathy/antipathy with whom?" is a question about inter-personal emotion [7]. "Who speaks what?" is known as the speech recognition problem [8]. For each of these problems, NTT CS Labs has devised automatic detection, recognition, or estimation methods.

In addition, NTT CS Labs developed the first real-time system for multimodal conversation analysis, which can automatically analyze multiparty face-to-face conversations in real time [9]. This system targets small-scale round-table meetings with up to eight people and uses a compact omnidirectional camera-microphone device, which captures audio-visual data around the table. From the audio-visual data, this system can estimate "who is talking" and "who is looking at whom" and display them on screen. The latest system has added speech recognition and displays "who speaks what" in semi-realtime [8].

3. MM-Space: Reconstructing conversation space by using physical representation of head motions

Beyond the analysis research toward future communication systems, we have recently devised a novel representation scheme and made a prototype system, called MM-Space [10], [11]. It aims to reconstruct multiparty face-to-face conversation scenes in the real world. The goal is to display and playback recorded conversations as if the remote people were talking in front of the viewers. The key idea is a novel representation scheme that physically represents human head motions by movements of the screens showing facial images of the conversation participants. This system consists of multiple projectors and transparent screens attached to actuators. Each screen displays the life-sized face of a different meeting participant, and the screens are spatially arranged to recreate the actual scene. The main feature of this system is *dynamic projection*: the screen pose is dynamically controlled in synchronization with the actual head motions of the participants to emulate their head motions, including head turning, which typically accompanies shifts in visual attention, and head nodding. We expect physical screen movement with image motion to increase viewers' understanding of people's behaviors. In addition, we expect background-free human images, which are projected onto the transparent screens, to be able to enhance the presence of the remote people. Experiments suggest that viewers can more clearly discern the visual focus of attention of meeting participants and more accurately identify who is being addressed.

The idea of this system comes from the importance of nonverbal behavior and nonverbal information in human conversations. This nonverbal information, which is exchanged in a face-to-face setting, cannot be fully delivered in a remote communication setting, e.g., videoconferencing. This insufficiency causes the unnaturalness of telecommunication environments. On the basis of this perspective, we introduced the key idea that physical representation of such nonverbal behaviors, especially head motions, can enhance the viewers' understanding of remote conversations. The nonverbal information expressed with the head motions includes the visual focus of attention, i.e., "who is looking at whom". Its function includes monitoring others, expressing one's attitude or intention, and regulating the conversation flow by taking and yielding turns [12]. A human tends to seize upon a target of interest at the center of his or her visual field: you orient your head toward the target, and various head poses appear according to the relative spatial position to the target. Our previous work indicated that interpersonal gaze directions can be correctly estimated with an accuracy of about 60–70% from head pose information and the presence or absence of utterances without actual eye-gaze directions. In addition, head gestures such as nodding, shaking, and tilting are important nonverbal behaviors. The speaker's head gesture is considered to be a sign of addressing or questioning, and the hearer's head gestures indicate listening, acknowledgement, agreement or disagreement, and level of understanding. Our system can replicate head gestures as physical motions of a screen, which gives viewers a stronger sensation of the presence of the meeting participants.

The physical representation of head motions is also related to human visual perception called *biological motion* [13]. Humans tend to anthropomorphize lifeless objects by assigning social context to movements, even if the objects are simple geometric shapes such as points and triangles. On the basis of the above findings, we hypothesize that realistic

Fig. 1. Overview of our system (MM-Space). (a) Reconstructed space and (b) actual meeting scene.

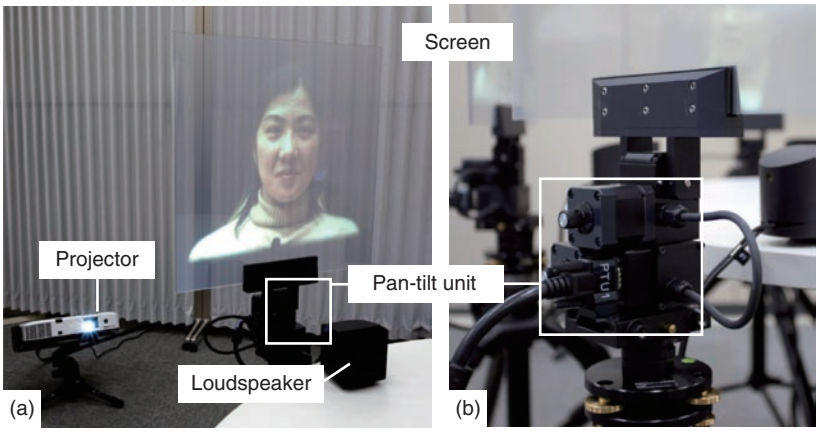

Fig. 2. Devices used in MM-Space.

kinematics offers strong cues for better understanding human communications, regardless of how it is implemented. Therefore, rather than pursuing realistic three-dimensional shape reproduction, our approach uses simple square screens and reproduces physical head motions to produce an augmented expression modality. We expect that combining this physical motion with image motion will boost the viewer's understanding. Moreover, it is known that human vision is highly sensitive to motion in the peripheral field. Therefore, viewers can perceive the entire conversation space including not only a person in front of them, but also the behaviors and presence of people located on their left or right.

3.1 System configuration

An overview of the MM-Space system is shown in **Fig. 1**. The reconstructed conversation space is shown in Fig. 1(a) and the actual conversation scene is shown in Fig. 1(b). In an actual conversation, multiple cameras and microphones capture the participants' faces and voices, respectively. In the reconstructed scene, multiple screens, projectors, actuators, and loudspeakers are placed to recreate the actual seating arrangement. Each participant's face is displayed on a flat transparent screen whose pose is dynamically changed in sync with his or her head motion. Each person's voice is play backed from the loudspeaker located in front of the screen displaying that person, so viewers can identify the speaker's position not only visually, but also aurally. The system described here was created to verify the effectiveness of head motion representation, so here we focus only on offline playback, but we plan to extend it to realtime telecommunication. In addition, our system provides a novel research platform for conversation analysis.

The screen, projector, and actuators are shown in **Fig. 2**. Each screen is highly transparent but includes

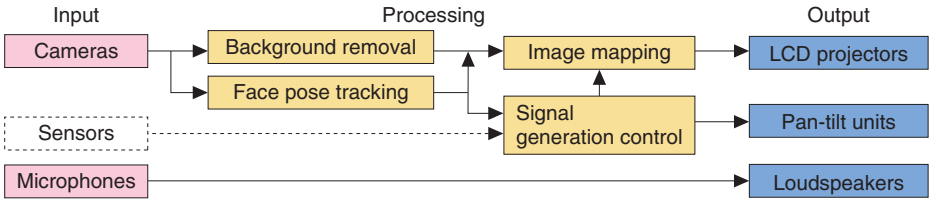

Fig. 3. Block diagram of MM-Space.

a diffusive material that catches the projector's output and makes it visible to the viewer. Each screen has its own liquid crystal display (LCD) projector behind it. Each screen is supported by an actuator, called the pan-tilt unit, that provides rotational motion in both the horizontal and vertical directions. We call this display device a *dynamic display*.

A block diagram of MM-Space is shown in **Fig. 3**. The processing parts provide visual face tracking, background removal, image mapping, and control signal generation. Visual face tracking measures the head position and pose of the meeting participants. Background removal creates images that emphasis the participants. The control signals drive the actuators holding the screens to reflect the participants' face poses, which are measured with visual face tracking and/or motion capture devices. Image mapping generates projected images that are skew-free.

3.2 Effectiveness

As the effect of the motion representation by dynamic screens, we hypothesized that a viewer can more clearly understand the gaze directions of meeting participants as well as who they are addressing. To verify this hypothesis, we compared two different conditions—with and without the motion representation—in terms of identification accuracy. Experimental results indicate that viewers could indeed more clearly recognize the gaze direction of meeting participants when the screens moved. In addition, the results statistically support the hypothesis that viewer can more accurately identify who is being addressed when the screens moved.

4. Conclusions and future perspective

This article introduced our system MM-Space for recreating a conversation space at different times and places. Its key feature is the physical representation of head motion as an additional expression modality. The synergy of physical screen motion and image motion on the screen is expected to boost our perception of social interactions involving the visual focus of attention. MM-Space is expected to be extended to realtime communication systems that can connect people located at different places. For that purpose, it will be necessary to explore in more detail the characteristics of the motion representation and evaluate how it can contribute to better expression and perception of addressing others and being addressed by others. In addition, other problems include the optimum camera configuration and the latency of telecommunications and physical systems. Finally, we believe that MM-Space will be a useful research platform for designing better communication systems and analyzing and understanding the mechanism of human communications.

References

[1] K. Otsuka, "Conversation Scene Analysis," IEEE Signal Processing Magazine, Vol. 28, No. 4, pp. 127–131, 2011.
[2] M. Argyle, "Bodily Communication—2nd ed. Routledge, London and New York, 1988.
[3] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," HSCMA2008, pp. 29–32, 2008.
[4] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "A Probabilistic Inference of Multiparty-Conversation Structure Based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances," Proc. of the 7th International Conference on Multimodal Interfaces (ICMI'05), pp. 191–198, Trento, Italy, 2005.
[5] S. Gorga and K. Otsuka, "Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection," Proc. of the ICMI-MLMI'10 International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, Article No. 54, Beijing, China, 2010.
[6] K. Otsuka, H. Sawada, and J. Yamato, "Automatic inference of cross-modal nonverbal interactions in multiparty conversations: "Who responds to whom, when, and how?" from gaze, head gestures, and utterances," Proc. of the 9th International Conference on Multimodal Interfaces (ICMI'07), pp. 255–262, Nagoya, Japan, 2007.
[7] S. Kumano, K. Otsuka, D. Mikami, and J. Yamato, "Analyzing empathetic interactions based on the probabilistic modeling of the co-occurrence patterns of facial expressions in group meetings," Proc. of the 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), pp. 43–50, Santa Barbara, CA, USA, 2011.
[8] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A.

Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-Latency Real-Time Meeting Recognition and Understanding Using Distant Microphones and Omni-Directional Camera," IEEE Trans. on Audio, Speech & Language Processing, Vol. 20, No. 2, pp. 499–513, 2012.
[9] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization," Proc. of the 10th International Conference on Multimodal Interfaces (ICMI'08), pp. 257–264, Chania, Crete, Greece, 2008.
[10] K. Otsuka, S. Kumano, D. Mikami, M. Matsuda, and J. Yamato, "Reconstructing multiparty conversation field by augmenting human head motions via dynamic displays," Proc. of the 2012 ACM Annual Conference (CHI EA'12), pp. 2243–2248, Austin, TX, USA, 2012.
[11] http://www.brl.ntt.co.jp/people/otsuka/ACM_MM2011.html
[12] A. Kendon, "Some functions of gaze-direction in social interaction," Acta Psychol., Vol. 26, pp. 22–63, 1967.
[13] G. Johansson, "Visual perception of biological motion and a model for its analysis," Attention, Perception, & Psychophysics, Vol. 14, No. 2, pp. 201–211, 1973.

Kazuhiro Otsuka
Senior Research Scientist, Distinguished Researcher, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and computer engineering from Yokohama National University, Kanagawa, in 1993 and 1995, respectively, and the Ph.D. degree in information science from Nagoya University, Aichi, in 2007. He joined NTT Human Interface Laboratories in 1995. He moved to NTT Communication Science Laboratories in 2001. He stayed at Idiap Research Institute, Switzerland, as a distinguished invited researcher in 2010. He has been a distinguished researcher at NTT since 2010. His current research interests include communication science, multimodal interactions, and computer vision. He received the Best Paper Award of the Information Processing Society of Japan (IPSJ) National Convention in 1998, the IAPR Int. Conf. on Image Analysis and Processing Best Paper Award in 1999, the ACM Int. Conf. on Multimodal Interfaces 2007 Outstanding Paper Award, the Meeting on Image Recognition and Understanding (MIRU) 2009 Excellent Paper Award, the Institute of Electronics, Information and Communication Engineers (IEICE) Best Paper Award 2010, the IEICE KIYASU-Zen'iti Award 2010, the MIRU2011 Interactive Session Award, and JSAI Incentive Award 2011. He is a member of IEEE, IPSJ, and IEICE.