Evaluation of Window Interface in Remote Cooperative Work Involving Pointing Gestures

Ryo Ishii, Kazuhiro Otsuka
NTT Communication Science Laboratories,
NTT Corporation
Kanagawa, Japan 243–0198
Email: {ishii.ryo, otsuka.kazuhiro}
@lab.ntt.co.jp

Shiro Ozawa, Harumi Kawamura
and Akira Kojima
NTT Media Intelligence Laboratories,
NTT Corporation
Kanagawa, Japan 239–0847
Email: {ozawa.shiro, kawamura.harumi,
kojima.akira}@lab.ntt.co.jp

Yukiko I. Nakano
Faculty of Science and Technology,
Seikei University
Tokyo, Japan 180–8633
Email: y.nakano@st.seikei.ac.jp

***Abstract*—Gazes and pointing gestures are important in performing collaborative work involving instructions with shared objects. However, in general video conferencing systems, the geometrical consistency of size and positional relationships of remote spaces are not displayed correctly on the display screen. This inhibits the transmissions of gazes and pointing gestures vis-a-vis shared objects. It is thus important to demonstrate how gazes and gestures can be smoothly transmitted by video and develop an advanced system that can do it. We previously proposed a "MoPaCo" window interface system that can reproduce a communication partner's space within a display as if the display were a glass window to achieve geometrical consistency between remote spaces. Experiment results demonstrated it enables users to feel the distance between themselves and their conversational partners on video is about the same as in a face-to-face situation and the partner is actually present. We also consider MoPaCo can generate video images that smoothly transmit gazes and pointing gestures; this paper describes experimental tests of the system's effectiveness in doing so. Results suggest MoPaCo allows users to accurately identify target objects as they could under face-to-face conditions through an actual glass window. Results of experiments on conversation quality show MoPaCo facilitates smooth conversation and communication among users and strengthens their memories of the conversations, suggesting the users actively engage in conversation and the system makes a strong impression on them.**

Keywords-Remote cooperative work; full gaze awareness; pointing gesture; window interface.

I. INTRODUCTION

Our objective is to achieve an advanced media space providing a seamless connection between two remote spaces. This will enable users to work closely together while sharing their respective spaces, discuss things, such as furniture layouts, and smoothly perform collaborative work involving the following of operating instructions. As an example of this, we simulate a situation where users in two seamlessly connected remote places discuss a certain burden and an appropriate place to put it before transferring it from their space to a remote space. In such a situation, the media space is expected to enable smooth transmission of nonverbal behavior such as gazes and pointing gestures (hereafter "gestures"). Nonverbal behavior is known to play an important role in ensuring smooth performance of collaborative work and instruction work [1], [2], [3]. However, since video images are displayed as-is in general videoconferencing systems, the geometrical consistency of size and positional relationships of the remote spaces are not displayed correctly. Thus, gaze and gesture directions cannot be correctly transmitted [4]. A major topic in human-computer interaction research has been the need for an advanced system and method giving users video images that look like face-to-face situations and allow smooth transmission of their nonverbal behavior. However, to date, no such method or system has been developed for an environment where two remote spaces are connected and actual objects are shared in them.

We previously proposed a window interface system called "MoPaCo" that reproduces a communication partner's space within a display as if the display were a glass window to achieve geometrical consistency between two remote spaces [5]. Since MoPaCo imparts motion parallax that adjusts to a user's viewpoint position, users can feel as if the remote spaces are connected smoothly as if separated only by a glass window. Experiment results demonstrated the users feel the distance between themselves and their conversational partners on video is about the same as in a face-to-face situation where the partner is actually present [6]. Since MoPaCo achieves geometrical consistency between two remote spaces, it is considered to have excellent potential for enabling smooth transmission of gazes and gestures. However, its effectiveness in doing so has never been tested. If this could be demonstrated, it would demonstrate that achieving video images connecting two remote spaces seamlessly as if they were separated merely by a glass window would be effective in transmitting gazes and gestures. This knowledge would make a significant contribution as a guide for designing new remote collaborative systems.

This paper describes experiments conducted to determine whether MoPaCo accurately transmits gazes at and gestures made to shared objects. It also describes evaluation experiments performed involving remote collaborative work to determine whether correct gaze transmission positively affected communication smoothness. The results indicate the system allows gazes and gestures to be transmitted in a similar manner

 ISBN: 978-1-61208-325-4

as in face-to-face conditions. They suggest MoPaCo users could refer to target objects smoothly, as if speaking face-to-face through a glass window, and conversation partners could predict the next target to be explained. Subjective assessments indicate MoPaCo encourages natural conversation and communication, facilitates conversation smoothness, and strengthens users' memories of conversations. This demonstrates the system contributes to improved conversation quality.

In the rest of this paper, Section 2 reviews related work and highlights of the paper. Section 3 presents details of the MoPaCo system. Sections 4 and 5 describe the evaluation of the system's gaze and gesture transmission and the evaluation experiments conducted involving remote collaborative work. Section 6 discusses the evaluation results in detail and Section 7 concludes the paper with a summary.

II. Related work

A. Importance of gazes and pointing gestures

Nonverbal behavior is known to play an important role in social psychology for performing collaborative work and instruction work smoothly. When conversation participants share the same physical environment and their tasks require complex reference to and joint manipulation of physical objects, the participants frequently observe a shared object most of the time instead of paying direct attention to their partner [1], [2], [7]. In such situations, establishing joint attention by paying attention to the shared object signals the listener's engagement in the conversation, and functions as evidence for comprehension in conversation grounding [8]. For example, if the listener asks for directions while observing a map, the listener's behavior in directing his or her gaze at the map to indicate sharing of the map information gives effective nonverbal feedback serving as evidence of comprehension. Suzuki et al. analyzed the relationship between gaze behavior and task completion, demonstrating nonverbal information such as gazes and gestures governs the success of a task [3].

In indication work, when referring to a target object in an indication, projectability, i.e., the predictability of which object a partner is observing and what he or she will explain or do from the direction of a partner's body or gaze, is shown to be important in making reference to objects easy [9]. In connection with this finding, Goodwin analyzed nonverbal behavior under face-to-face conditions, in which a speaker indicates a target object to a listener [10]. First, the listener appropriately adjusts the direction of his or her body so as to share a mutual gaze at the object with the speaker. This indicates the listener is actively listening to the speaker. Conversely, when the speaker gives the indication, he or she changes position so both the object and listener are visible. The listener then comprehends the speaker's target of interest and directs attention at the next object to be indicated. In this way, when the speaker refers to a target object, the listener can smoothly identify it.

B. Gazes and gestures in communication systems

Video expression enabling transmission of gazes and gestures has been a major challenge in human-computer interaction research. Here, a person's awareness of the conversational partner's gaze is defined as gaze awareness. Gale and Monk divided gaze awareness into three levels, as follows [11].

- Mutual gaze awareness: A person can understand he/she is being observed by a conversational partner. This is generally known as "eye contact".
- Partial gaze awareness: A person can understand the eye direction (up, down, left, right) of the conversational partner.
- Full gaze awareness: A person can understand what object the conversational partner is observing.

This classification also applies to gestures. Many studies have focused on achieving gaze awareness in video conferencing systems. First, methods for achieving mutual gaze awareness in remote face-to-face communication have been considered. Methods have been proposed using a half mirror [12], a liquid crystal shutter [13] and a stereoscopic camera or time-of-flight camera [14] to generate frontal facial images. In addition, a widely used method has been developed in which the deviation of the face and camera positions is five degrees or less and thus eye contact is achieved [15].

Furthermore, systems have been proposed for extending multi-party conversations, i.e., HYDRA [16], Browser Magic [17], GAZE Groupware system [18], and GAZE-2 [19]. These systems enable users to understand the direction a person faces from the person's head direction; thus both partial and mutual gaze awareness are achieved. In addition, the Browser Magic system [20] enables users to understand whom the conversation partners are observing; thus, full gaze awareness is achieved assuming users at three remote sites. Furthermore, a method has been proposed to achieve full gaze awareness in many-to-many human conversations, i.e., MultiView [21], which presents parallax images in accordance with each user's viewpoint using a camera and projector for each user. However, these systems focused on who the participants observe and did not address the issue of correctly transmitting gaze behavior to objects in shared spaces. Although Clearboard [22] enables gazing at a shared display surface, it is limited to the display surface and does not achieve full gaze awareness for objects in a shared space. Therefore, insufficient study has been done on video expression techniques connecting two remote spaces in a media space smoothly and achieving full gaze awareness allowing users to understand what objects their conversational partner is observing.

In another attempt to achieve effective transmission of gazes at and gestures made towards shared objects in a remote space, the idea of having a vicarious robot stand in for the user has been proposed [23]. This robot acts as a substitute for a remote user and reflects gestures and head direction (pseudo eye direction) in real time. In a test at a surrogate robot exhibition, it was able to smoothly establish mutual gazing by directing its attention to audience members observing it and referring to objects pointed out to it. This research showed the importance of transmitting gazes and gestures in achieving smooth remote communication, but focused only on transmitting three nonverbal information factors (gazes, gestures, and body positions), using a vicarious robot device as a human substitute. However, there is a need to convey multiple, complex nonverbal information factors in addition to gazes, gestures, and body positions, e.g., facial expressions and nodding. From this viewpoint, it must be considered important to transmit all nonverbal information emanating from the person in the video to transmit nonverbal behavior in the same manner as in face-to-face situations.

In contrast to these methods, our aim is achieving geometrical consistency for the size and positional relationships of two remote spaces on a video display. We suggest the MoPaCo

 ISBN: 978-1-61208-325-4

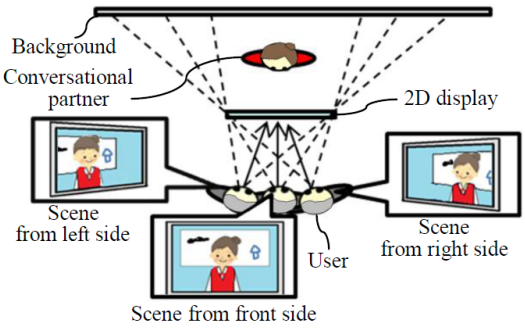

Figure 1: Concept images of video representations caused by motion parallax.

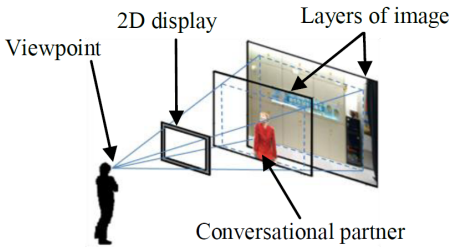

Figure 2: The person layer and background layer are projected.

system as a means of presenting images as clearly as if the spaces were merely separated by a glass window [5]. Since MoPaCo reproduces the size of the spaces and their positional relationship, it transmits body positions, gestures, and gazes naturally and correctly. We have previously performed experiments with the system demonstrating it allows users to feel the interpersonal distance between themselves and their conversational partners in a remote space so they can feel the reality of face-to-face communication and the partner's presence [6]. Since MoPaCo achieves geometrical consistency between two remote spaces, it is considered to have excellent potential for smoothly transmitting gazes and gestures.

III. Window Interface: MoPaCo

A. System Summary

We previously proposed a real-time video communication system called Motion Parallax Communication (MoPaCo) that reproduces a communication partner's space within a display as if the display were a glass window to achieve geometrical consistency between two remote spaces [5], [6]. Figure 1 shows MoPaCo-produced motion parallax video images of a conversational partner that correspond to the viewpoint positions of different users. The display for a user some distance from the partner in the video can give the user and partner the feeling they are linked as if seeing each other through a glass window. We consider this motion parallax video representation will eliminate spatial separation, improve the conversational partner's presence, and enable the transmission of nonverbal information associated with depth by imparting depth information to video images. Presenting a motion parallax video of a partner on a 2D display corresponding to the viewpoint positions of different users requires the following process:

(I) Measuring each user's viewpoint position.

(II) Constructing a 3D space having information on the dimensions and positional relationships of the people and the background, based on information obtained from a camera or other means.

(III) Rendering the 3D space constructed in step (II) on a 2D display, to correspond to each user's viewpoint position obtained in step (I).

MoPaCo implements steps (I) and (II) with a single monocular camera. This section describes the detailed process for steps (I) to (III).

B. Measuring User's Viewpoint Position

We proposed using a single monocular camera to detect each user's viewpoint. Before calculating the 3D position from parts information (coordinate position) of each face in the 2D image, the system performs preprocessing by measuring the eye separation distance of each user. It then acquires the distance of each user from the camera, using the depth from focus function used for achieving focus in ordinary cameras. The lens distortion of the image was eliminate by Zhang's lens distribution correction method [26] before this process. During this process, template matching is performed on the image captured from the camera to measure the positions of both eyes (2D coordinates within the image) and the orientation of the head. The system calculates the eye separation distance of each user from the user-to-camera distance, the information measured from the image, and the camera's angle of view and resolution. With this information, real-time capture starts and the system obtains the positions of both eyes (2D coordinates within the image) and the orientation of the head from the captured image, and calculates the viewpoint position z of that user from the camera from there at that time. The x- and y-coordinates are calculated from the 2D coordinates within the image and the image's pixel pitch.

C. Construction of 3D Space

We proposed constructing 3D information for an image captured from a single camera by performing background difference processing using background information acquired beforehand (Images of several seconds were captured for background information), maintaining the 2D plane and dividing it into personal and background areas, and creating a multi-layer structure with those areas arranged as layers in accordance with their depth-wise positions (see Figure 2). Using 2D images ensures a high-resolution display; furthermore, subjecting only the background difference to image processing lowers processing costs and enables real-time processing. The system generates a "person layer" showing a full size image of a person and a "background layer" showing a full size image of the background. These layers have a distance relationship from the camera. The distance information measures for the background layer are calculated beforehand using the depth-from-focus method of the camera's auto focus function, when the background difference image is acquired. For the person layer, the user viewpoint position is used. These distances become the information about the distance from the camera to the person layer and the background layer, respectively. The system then uses (1) to calculate the full size (width w_i × height h_i) of each layer i from the thus-acquired distance information d_i and the camera's angle of view (width θ_x, height θ_y). This procedure configures a 3D space having full size and position information.

$$w_i = 2 * d_i * tan(\theta_x/2), h_i = 2 * d_i * tan(\theta_y/2) \quad (1)$$

 ISBN: 978-1-61208-325-4

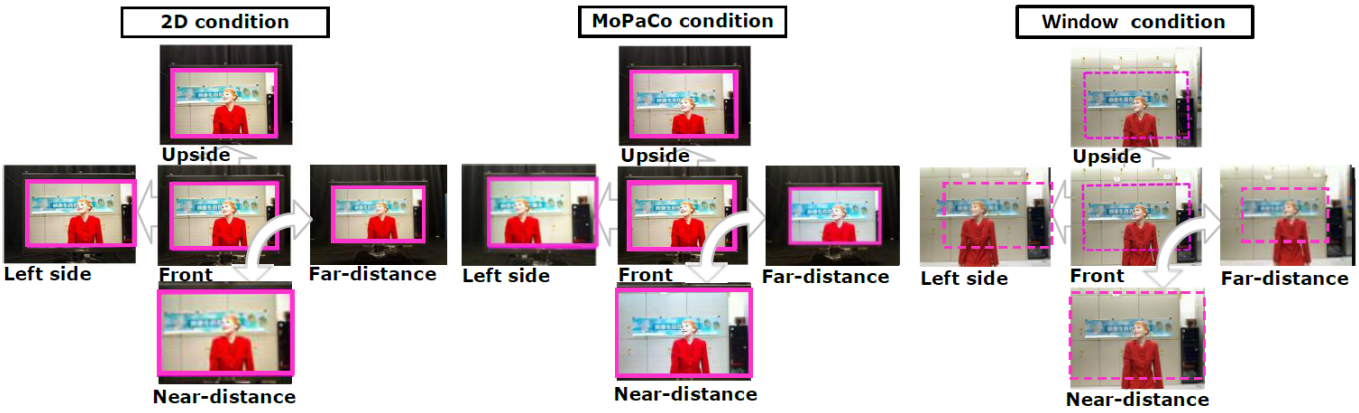

Figure 3: Scenes for 2D, MoPaCo, and Window conditions.

D. Rendering 3D Space on User Viewpoint Basis

As shown in Figure 2, the person layer and background layer generated by the 3D spatial information module are projected in perspective to match the user's viewpoint position, using the 2D display as a projection surface. Thus, motion parallax video is implemented.

E. Implementation

Using the above-described methods, we implemented the MoPaCo system, which enables real-time bidirectional viewing of motion parallax video. The development environment was a camera with HD size resolution (1920 × 1080), a computer with Intel Core i7 Extreme 980X as the CPU and 12 GB of memory, and a NVIDA GeForce GTX480 graphics board. Table I shows the implementation results; "lag from viewpoint movement" is the time from the user's viewpoint position moving to the time motion parallax appears in the video; "lag of camera image" is the time until the captured video appears.

Figure 3 shows scenes used in experiments conducted to enable users to evaluate the MoPaCo system. They show differences in the visibility of a conversational partner from the same position under 2D, window, and MoPaCo conditions. Under window conditions, users can observe the conversational partner through an actual glass window. Five view positions were used, i.e., the front, left, and upper sides of the display, and the near distance to and far distance from the display. In comparing the scenes under the window and 2D conditions, since there was no parallax in the video under the 2D conditions even if the user's head moved, the human dimensions and positional relationships did not match. Under the MoPaCo condition, in contrast, the dimensions and positional relationship between the person and background were reproduced in the video.

IV. Evaluation of Accuracy of Gaze and Pointing Gesture Transmission to Objects

A. Experimental Method

We conducted experiments to determine whether the MoPaCo system correctly transmitted gazes and gestures, and to compare and verify transmission accuracy when it was directed through the window and the actual 2D video as general experimental conditions, in addition to MoPaCo conditions.

- 2D condition: observing the conversational partner in an image taken with a camera directly on a 2D display. This condition is for the use of a classic 2D video conferencing system. In this case, the user's viewpoint position is where the image is displayed at a position when the user is sitting straight in the chair.
- MoPaCo conditions: observing the conversational partner with MoPaCo.
- Window conditions: observing the conversational partner through a glass window.

TABLE I: Performance of MoPaCo system.

	Frame rate	Response
Lag of camera image	30 fps	260 ms
Lag from viewpoint movement	30 fps	300 ms

B. Experiment Results

In the experimental setup, the subject was seated on a chair 80, 150, or 230 cm from a partition with a glass window installed between the subject and his/her conversational partner (Figure 4) and was able to observe the partner's space through the glass window. Since the glass window size (46 cm high × 80 cm wide) was less than the display size, the subjects could not see the display edges. A camera was installed immediately above the glass window so as to match the participant's gaze [15].

Rectangular 50 × 50 panels for use as indication targets were placed in a 3 × 18 panel arrangement on a wall 200 cm behind the participant (panel rows were labeled from A to C vertically from the top; columns were labeled 1 to 18 from the left). Four panels were chosen as indication targets:

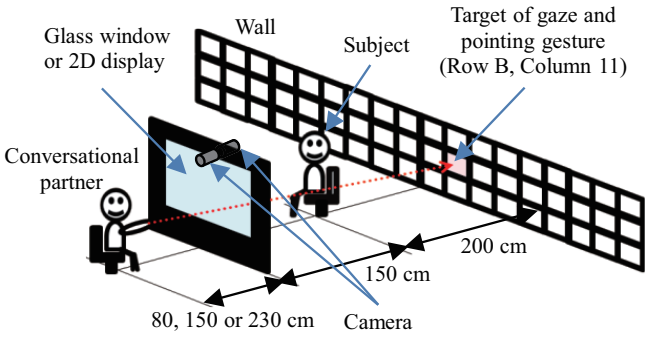

Figure 4: Experimental setting for measuring gaze and pointing gesture transmission accuracy.

 ISBN: 978-1-61208-325-4

TABLE II: Pointing gesture accuracy rate and Turkey-Kramer multiple comparisons.

Conditions		Accuracy rate (%)			Multiple comparison		
Targets	Partner's position	2D	MoPaCo	Window	2D vs MoPaCo	2D vs Window	MoPaCo vs Window
Row B, Column 4	80 cm	3.0	21.2	24.2	$*$	$*$	n.s.
Row B, Column 4	150 cm	0	27.3	24.2	$**$	$**$	n.s.
Row B, Column 4	230 cm	0	13.3	15.2	n.s.	n.s.	n.s.
Row B, Column 8	80 cm	3.0	30.3	33.3	$*$	$*$	n.s.
Row B, Column 8	150 cm	3.0	30	27.3	$**$	$*$	n.s.
Row B, Column 8	230 cm	0	24.2	27.3	$*$	$*$	n.s.
Row B, Column 11	80 cm	3.0	33.3	33.3	$*$	$**$	n.s.
Row B, Column 11	150 cm	0	27.3	30.3	$*$	†	n.s.
Row B, Column 11	230 cm	0	21.2	24.2	†	$*$	n.s.
Row B, Column 15	80 cm	0	12.1	21.2	†	†	n.s.
Row B, Column 15	150 cm	0	9.1	9.1	n.s.	n.s.	n.s.
Row B, Column 15	230 cm	0	9.1	9.1	n.s.	†	n.s.

†: $p<.10$, $*$: $p<.05$, $**$: $p<.01$

B4, B8, B11, and B15. During the experiment, the participant was shown someone performing a gesture either in a video or through the window and verbally answered which object was being indicated. Three trials were performed for each condition. To minimize order effects, experiment conditions were randomly chosen from combinations of three observation conditions, three indicator positions, and the four indication targets.

Table II shows the experiment results obtained for the 11 participants (9 males and 2 females in their 20s). The table shows the average accuracy rate of participant answers regarding the indication target under each of the three experiment conditions. We performed a repeating two-way factorial analysis of variance for each of the four indication targets to determine whether the conversation partner's position or observation conditions affected the accuracy rate. This showed the conversation partner's position did not have a significant effect but the observation conditions did (B4: $F(2,90) = 13.92, p<.01$, B8: $F(2,90) = 10.23, p<.01$, B11: $F(2,90) = 15.56, p<.01$, B15: $F(2,90) = 8.14, p<.01$). Since the observation conditions had a contributing effect, multiple comparisons were performed for each of the three observation conditions using the Tukey-Kramer method. Table II shows the test results; the accuracy rates for the 2D condition were 0% in most cases but increased dramatically under the MoPaCo and window conditions, showing significant differences and trends. No significant differences were seen between the results for the MoPaCo and window conditions. This shows similar precision is obtained regardless of distance when transmitting indication actions under the MoPaCo and window conditions, i.e., MoPaCo successfully reproduces an actual window's size and location relationships.

V. Experiment in Remote Collaboration Involving Pointing gestures

A. Experimental Procedure

We investigated the communication smoothness MoPaCo provides using nonverbal behavior such as gazes and gestures in remote collaborative work. Specifically, we evaluated video expression achieving geometric integrity involving actual size and position relations with motion parallax adjusting to the user's viewpoint to confirm MoPaCo achieves smooth communication by smoothly transmitting gazes and gestures. As the evaluation method, we propose a hypothesis that the video expression can allow a user to recognize a shared object that the conversational partner indicates with gazes and gestures. We also propose a hypothesis that smooth transmission of nonverbal behavior such as gazes and gestures will facilitate smooth remote communication and improve users' impressions of conversations and conversation quality factors such as the user's conversational engagement. Accordingly, we evaluated the system for the smoothness and the impressions of conversations it provides.

In carrying out the evaluation, two subjects were placed in a conversational setting and tasked with choosing the furniture layout in each other's rooms. As the specific method of evaluating the smoothness with which they could identify the objects their partner indicated, we measured the time required to identify objects and the number of utterances one of the subjects had to make about the object's position before the other could positively identify the object. We expected the required time would become shorter and the number of utterances would become smaller if gazes and gestures were used to help the user identify the object. Conversation quality was assessed through 6-level subjective evaluations made using the Rickert method, with questionnaires asking questions about conversation smoothness and impressions. Subjective evaluation items are shown in Table III. In addition to subjective evaluations, we measured the participants' memory of the conversation and the furniture used as an indicator of whether they actively participated in the conversation. We consider that active participation and strong impressions of a conversation create stronger memories. Specifically, 80 pieces of furniture were shown in the questionnaire form and the subjects answered whether a given piece of furniture was in the partner's room. Then, we measured the accuracy rate of the subjects' responses.

To perform a comparative investigation between ordinary conversations, 2D video conversations, and conversations through an actual glass window, tests for this activity were conducted under these conditions.

- 2D condition: a conversation through images taken with a camera displayed as-is on a 2D display (the display and

TABLE III: Contents of subjective evaluation.

- Conversation smoothness: Did the conversation progress smoothly?
- Communication: Was communication achieved?
- Window feeling: Did you feel as though you were speaking through a window?
- Enjoyment: Did you enjoy the conversation?
- Affinity: Did you feel an affinity toward your conversation partner?

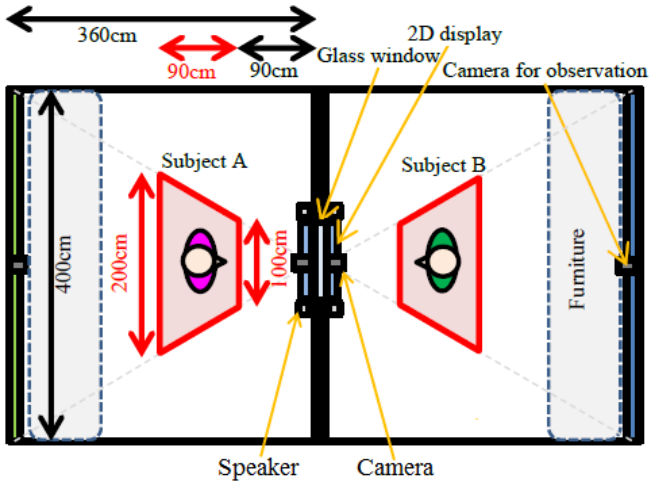

Figure 5: Top view of experimental equipment.

Figure 6: Example of arbitrarily-placed furniture.

camera angle were adjusted to include all objects to allow the participant to see the entire region of objects the indicator would refer to). This condition is for the use of a classic 2D video conferencing system.

- MoPaCo conditions: a conversation through a window image using MoPaCo.
- Window conditions: a glass window was placed between two adjacent rooms and participants conversed through it.

Figure 5 shows the experiment environment. Two participants entered adjacent rooms (360 cm × 400 cm) assigned individually to them and stood in a space in which they could move (a trapezoid 90 cm tall, 100 cm at the top, and 200 cm at the base) 90 cm away from the wall separating the two rooms. Participants were permitted to move freely within the movement space. They were not allowed to touch and move the furniture. A glass window 49.8 cm tall × 88.4 cm wide was installed 120 cm above the floor on the wall separating the rooms. Under the window conditions, conversations took place through this window. Under the 2D and MoPaCo conditions, a 40-inch 2D display (1920 × 1080 resolution) identical in size to the window was installed in front of the window. Participants communicated while watching the video. Image delay was 300 ms under both 2D and MoPaCo conditions. A camera was installed immediately above the window so as to match participant gaze [15]. Voice was collected through a microphone located in front of the display, and was output to speakers located directly beside the partner's window. Sound was delayed by 200 ms using a delay generator to ensure lip-sync under both 2D and MoPaCo conditions. Each room was arranged with 14 items (poster, table, TV, etc.) chosen randomly from a set of 84 items. Figure 6 shows an example object layout in the room.

The experiment began with the participant standing in the center of the movable space. At a signal to begin, images and voice of the participant's partner were output, and the pair conversed for ten minutes. Participants were instructed to discuss how to preferably rearrange items placed haphazardly in the two rooms. Afterwards, participants were tasked with choosing one item from their partner's room and considering where they would place it in their own room.

To minimize order effects, the three experiment conditions were used in experiments randomly. Each pair used a different set of items under each condition. After executing the experiment under each condition, the participants filled in a questionnaire concerning subjective assessments and assessments for measuring participant memory of items in the room. Sixteen participants (10 males and six females in their 20s-40s) were formed into eight pairs of friends or family members.

B. Collected Conversation Corpus

Participants' utterances, gaze behavior, and gestures were collected for analysis through the following methods.

- Utterances: voices were recorded and transcribed.
- Gaze targets: wearable Tobii Glasses [24] were used to measure the participants' gaze behavior. This allows measurements to be taken using only a pair of transparent glasses, putting little burden on the user and avoiding blocking the view of a participant's gaze direction by covering the eyes. Tobii Glasses output the gaze location in the participant's view image as a 2D coordinate plane at 30 fps. We used the annotation tool Anvil [25] to annotate gaze target objects from video images. Each room contained 14 labeled gaze target objects and one participant.
- Pointing gestures: participant actions were collected on video images and then annotated using Anvil. Gestures were defined in three steps: "preliminary action", from when the participant began moving his or her arm to perform the gesture, "during indication", when the participant pointed at the indication target object, and "returning action", when the participant finished indicating and returned his or her arm to the starting position. Gestures were annotated in these three steps.

After synchronizing these three types of data, video and annotation data were integrated into a single file of Anvil data, and conversation corpus data was created. Figure 7 shows an example; the total data comprised 24 conversations (three conditions, eight participant pairs) of five minutes each for a total of 120 minutes of corpus data.

C. Results for Target Identification Smoothness

1) Time required for identifying objects: When a participant indicated an item using demonstrative pronouns ("here","there", etc.) the name of the item, or gestures, the time the partner needed to identify the item was measured. The time started when the item was indicated and ended when the partner started gazing at it. Since 2D and MoPaCo conditions included a 300 ms image and voice delay, the starting time was set to when the indicator's voice was output from the speakers.

 ISBN: 978-1-61208-325-4

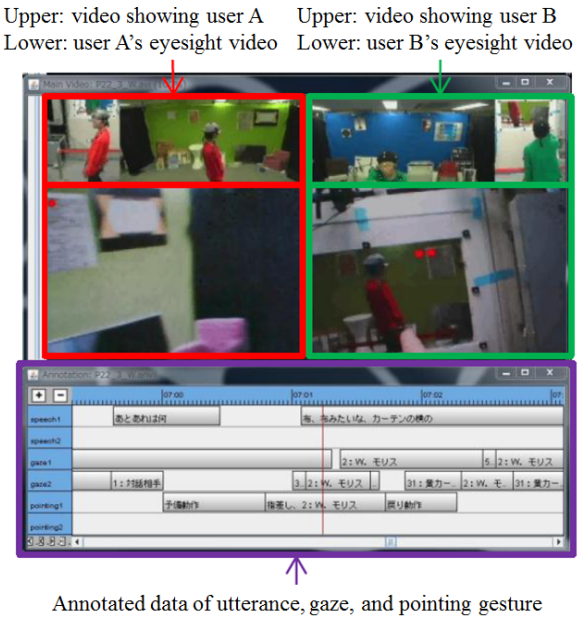

Figure 7: Corpus data in Anvil annotation tool.

TABLE IV: Result of analysis of time required for identification of indicated object.

Condition	Average time (ms)	One-way analysis of variance	Multiple comparison		
			vs 2D	vs MoPaCo	vs Window
2D	2800	*	–	**	**
MoPaCo	1100	(F(2, 447)	–	–	n.s.
Window	1500	=13.6)	–	–	–

†: p<.10, *: p<.05, **: p<.01

Table IV shows average required reference times for all conversations. The 2D conditions required the longest average time; approximately 2.8 seconds was required for 2D conditions, 1.1 seconds for MoPaCo conditions, and 1.5 seconds for window conditions. To determine whether experimental conditions made a difference in the time required for identifying target objects, we performed a one-way factorial analysis of variance. The results show a significant difference between experimental conditions ($F(2, 612) = 59.63, p<.01$). Next, we performed multiple comparisons using the Tukey-Kramer method to identify differences between pairs of conditions. These tests showed significant differences only between 2D and MoPaCo conditions ($p<.01$) and 2D and window conditions ($p<.01$). The results demonstrate 2D conditions make the identification time longer than for the MoPaCo and window conditions, and confirm MoPaCo conditions allow smooth target identification similarly to window conditions. This suggests our hypothesis was correct.

2) Number of indicator's utterances about object's position: We counted the number of utterances participants made about an object's position. Example sentences used to indicate the position included, "It's on the edge of the right-hand side of XXX (the name of another object)", "Not over there", and "It's on the opposite side". Table V shows the results obtained for the average number of utterances about an object to be identified. Under the 2D conditions the number (0.27) was highest; it was 0.06 under the MoPaCo conditions and 0.09 under the window conditions. We performed one-way factorial analysis of variance to determine whether the experimental conditions affected the differences found in the number of utterances made in indicating an object's position. The results showed there was a significant difference due to the conditions (F (2,447) = 8.38, p <.01).

TABLE V: Result of analysis of number of instructor's utterances about object's position.

Condition	Average number (per second)	One-way analysis of variance	Multiple comparison		
			vs 2D	vs MoPaCo	vs Window
2D	0.27	*	–	**	*
MoPaCo	0.06	(F(2, 447)	–	–	n.s.
Window	0.09	=8.38)	–	–	–

†: p<.10, *: p<.05, **: p<.01

TABLE VI: Result of memory of furniture in partner's room.

Condition	Accuracy rate (%)	ANOVA	Multiple comparison		
			vs 2D	vs MoPaCo	vs Window
2D	86.9	*	–	*	†
MoPaCo	94.8	(F(2,45)	–	–	n.s.
Window	93.5	=4.56)	–	–	–

†: p<.10, *: p<.05, **: p<.01

Next, multiple comparisons using the Tukey-Kramer method were performed to confirm the differences between pairs of individual criteria. Results showed significant differences between the 2D and MoPaCo conditions (p <.01) and between the 2D and window conditions (p <.05), but none between the MoPaCo and window conditions. They show subjects make more utterances to indicate an object's position under the 2D conditions than under the window and MoPaCo conditions. They also show the MoPaCo conditions enable users to identify objects with the same small number of utterances as for the window conditions. This suggests our hypothesis was correct.

D. Results for Conversation Quality

1) Memory of furniture in partner's room: We calculated the accuracy rates obtained in a memory test the subjects took regarding the furniture in their partner's room. Table VI shows the average accuracy rate for all 16 subjects' answers. We performed one-way factorial analysis of variance to verify whether the experimental conditions affected the differences found in the rate. The results showed the conditions produced significant differences (F (2,45) = 4.56, p <.05).

Next, multiple comparisons using the Tukey-Kramer method were performed to confirm the differences between pairs of individual criteria. Results showed significant differences between the 2D and MoPaCo conditions (p <.01) and between the 2D and window conditions (.05 <p <.10), but none between the MoPaCo and window conditions. They show the accuracy rate of memory about the furniture in the partner's room is lower under the 2D conditions than under the window and MoPaCo conditions. They also show the MoPaCo conditions enable users to remember conversations as well as they can under the window conditions. This suggests our hypothesis was correct.

2) Subjective evaluation results: Table VII shows the average values for participants' subjective evaluations. We performed one-way factorial analysis of variance for each of five items to determine whether experimental conditions affected the values. Since an effect of experimental conditions on the evaluation values was shown, multiple comparisons using the

 ISBN: 978-1-61208-325-4

TABLE VII: Subjective Evaluation Results.

Items of subjective evaluation	Average of subjective score			ANOVA	Multiple comparison		
	2D	MoPaCo	Window		2D vs MoPaCo	2D vs Window	MoPaCo vs Window
Conversation smoothness	3.0	4.0	4.3	** (F(2, 45)=11.27)	**	**	n.s.
Communication	3.6	4.5	4.6	* (F(2, 45)=3.64)	†	†	n.s.
Window feeling	3.0	4.0	4.4	* (F(2, 45)=3.42)	†	†	n.s.
Enjoyment	3.5	4.5	4.3	* (F(2, 45)=4.15)	*	†	n.s.
Affinity	3.0	4.1	4.0	** (F(2, 45)=7.25)	**	**	n.s.

†: p<.10, *: p<.05, **: p<.01

Tukey method were performed for each condition. Table VII shows the test results; significant differences and trends were found for each item between 2D conditions and MoPaCo and window conditions, but none between MoPaCo and window conditions. This suggests "conversation smoothness", "communication", "window feeling", "enjoyment" and "affinity" were all higher under MoPaCo and window conditions than under 2D conditions, but no significant differences were found between them under MoPaCo and window conditions.

Next, we demonstrate whether the results obtained for smooth transmission of identification and improved memory about communication content have a major effect on improving communication smoothness. We evaluated the correlation between the subjective score results for items relevant to conversation smoothness and (a) the time required to identify an indicated object, (b) the number of utterances indicating the object's position, and (c) the accuracy rate of memory about furniture. The correlation coefficient between the subjective values for conversation smoothness items and the required time was a negative correlation, -0.45. The coefficient between the values and the average number of utterances indicating the position was a low negative correlation, -0.22. This shows the differences in smoothness in identifying the object possibly affected the users' introspection regarding the conversation smoothness. Finally, the coefficient between the subjective values for conversation smoothness and the accuracy rate of memory about furniture was a positive correlation, 0.31. This shows the differences in memories of the furniture possibly affected the users' introspection regarding the conversation smoothness.

VI. Discussion

Evaluations of communication precision of indication actions showed that indication actions performed through MoPaCo were similarly precise to those performed through an actual glass window, regardless of the distance between the indicator and the display. We therefore consider MoPaCo successfully reproduced similar sizes and positional relationships seen in an actual glass window. While the difference was insignificant, the average accuracy rate was 23.0% under the window conditions and 21.4% under the MoPaCo conditions, i.e., the former was slightly higher. We consider this is because MoPaCo displays people as a flat layer, and thus even when users change their viewpoint their partner's arm direction does not actually change. For example, if one participant stretches his or her arm toward another, the latter should be able to see the former's arm stretching to the left when he or she moves to the right. In MoPaCo, the arm will still be shown stretching straight ahead. Post-experiment interviews with participants showed some of them detected a change in the direction of their partner's arm as they moved through parallax, even though the direction did not actually change. We consider this is an illusion caused by parallax in the background. This suggests the possibility that since arm movements are slight when the user does not move much from the front of the display, even if the person is shown as a flat layer, this does not greatly affect the precision of indication actions. This leads us to consider that using MoPaCo to perform collaborative work while sharing the spaces and items in two locations allows work to progress smoothly through the natural use of indicative actions.

We conclude the required times for referring to an object indicated in a partner's space were the same for the MoPaCo and window conditions. Subjective evaluations showed similar assessment results for conversations and communication smoothness, suggesting MoPaCo usage results in smooth conversation and transmission of indications. In other words, this suggests indication actions were smoothly referenced by presenting through-window images, considering size and positional relationships in the media space as if the two spaces were actually joined by a glass window. From the experiment participants' activities, we consider two reasons contributed to this.

Since MoPaCo presented spaces while preserving the geometrical consistency of width and positional relationships, gestures made at objects within the space and gaze targets were correctly transmitted. Under the 2D conditions, listeners would mistakenly look in the opposite direction of that being indicated, and indicators were often forced to name the object or otherwise provide concrete supplementary information. Figure 8 shows an example of this; the indicator (participant A) pointed to a clapperboard behind and to the right of the listener (participant B) using a gesture and gaze while saying, "Over there's a thing making a clapping noise, what's that called, a clapperboard?" (7m7s947 from conversation start). Participant B gazes at participant A and quickly identifies the target object, but the indicated direction is not transmitted directly, and participant B gazes in the opposite direction. Participant B says to participant A, "Which one?" (7m11s410). Participant A explains the location of the clapperboard in detail, saying "That thing that goes clap in TV and movies" (7m13s245). After that, participant B finally directs his gaze at the clapperboard, and says "Oh" while making a gesture (7m13s840). In this manner, since the direction of an indicator's gaze and gestures cannot be accurately transmitted under the 2D conditions, finding an indication object often requires confirmation. Conversely, under the MoPaCo and window conditions, this sort of confirmation is not required. In other words, we consider using MoPaCo to reproduce the size and positional relationships of a space enables gaze and gesture directions to be accurately transmitted, allowing indication work to progress smoothly as if through an actual glass window.

We consider that through the window metaphor, since

 ISBN: 978-1-61208-325-4

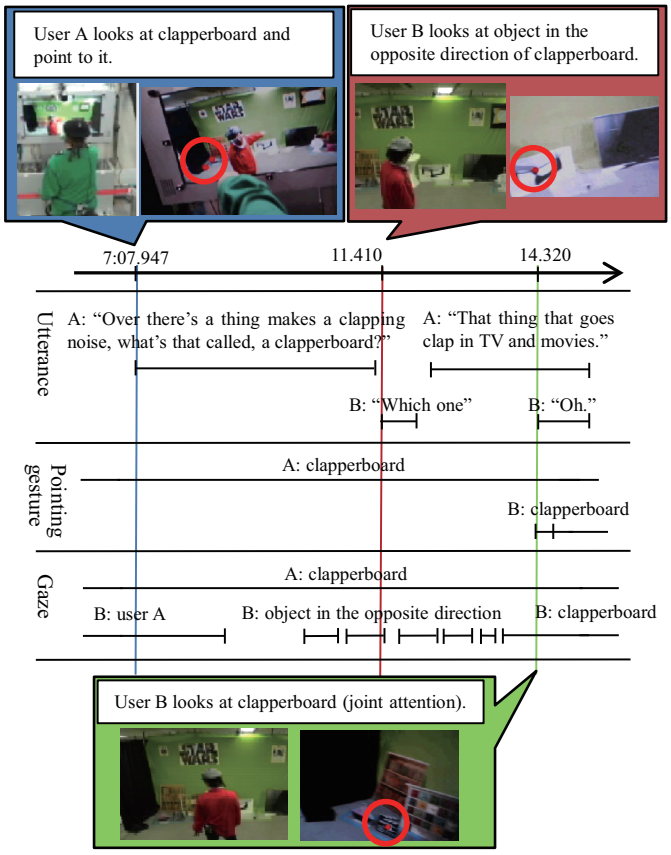

Figure 8: Example scene of instructor's action and recognition in 2D situation.

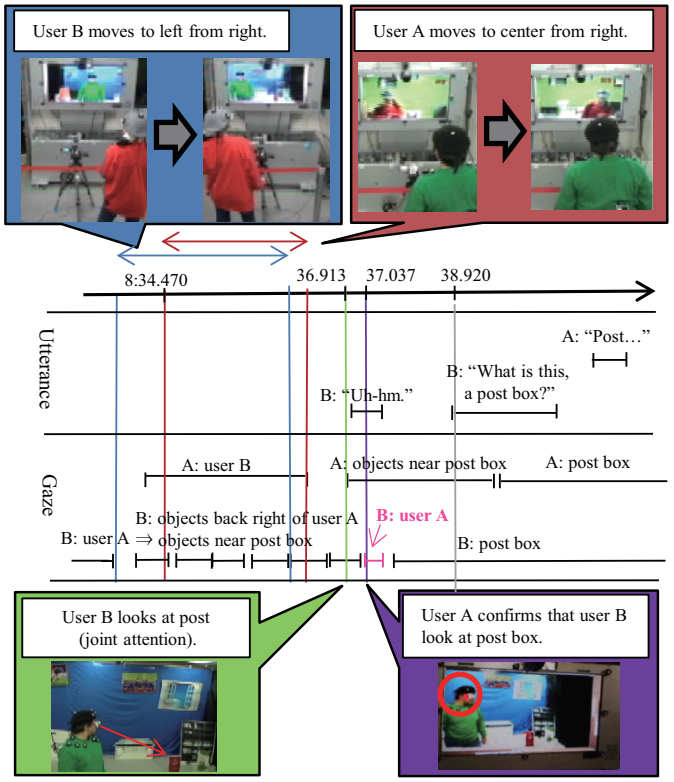

Figure 9: Example scene of instructor's action and recognition in MoPaCo situation.

gazing at objects was accompanied by a physical movement made by the user, changes in the user's position and direction were clear, and the partner could easily predict the object of the user's interest. Under the 2D conditions, since the entire room was displayed, users would move only their gaze without changing the position or direction of their head when gazing at an object. This made it difficult to grasp the direction of their gaze in the video, and participants were not often seen matching the gaze direction of their partner, moving their bodies in the same direction, or sharing mutual gazes. In contrast, under the MoPaCo and window conditions, when a participant gazed at an object, this was accompanied by a change in physical position or direction in most cases. The partner would then often change his or her position or direction to match the gaze. An example of this behavior under the MoPaCo conditions is shown in Figure 9. Participant B observes items in participant A's space from right to left. User A gazes at user B, and when he notices this movement, he moves from right to left to match user B's movement so he can always be seen from user B's position (8m34s470). When user B stops moving, user A also stops moving, turns his body toward the direction in which user B is looking, and shares a mutual gaze (8m36s913). In this case, we consider that user B is predicting the next instruction or explanation. Generally, user B directs his or her gaze at user A, and confirms user A is looking in the same direction in which he himself or she herself is looking (confirming he or her is sharing a mutual gaze) (8m37s037). He or she then indicates the post box they are both observing and says "What is this, a post box?" (8m38s920). At this point, user B's use of the demonstrative pronoun "this" indicates user B shares an interest target with user A and is predicting the next instruction or explanation. This sort of predictive behavior was seen under both MoPaCo and window conditions. Under both conditions, it was possible for users to firmly express targets (directions) of interest by observing objects using the glass window as a metaphor, which can be considered as a cause for this behavior. In other words, we consider users would show greater movements of the direction of their head and body by peeking through the glass window at something, allowing the partner to predict their target of interest (in the example shown in Figure 9, user A moves in response to user B's movements, and performs a mutual gaze). Since MoPaCo presented a window, when observing a space with a degree of size such as the one used in the experiment in this study, not all items could be observed at once, and participants were forced to move. However, it was shown nonverbal communication transmission was smoother than it is when simply displaying a 2D image in which the entire room could be seen. Thus, if the objective is to allow collaborative work using indicative actions to be performed smoothly, it is important to allow the natural transmission of nonverbal information such as gazes and gestures to be performed even if the entire room cannot be seen at all times. From this viewpoint, the MoPaCo window interface can be considered effective. Improvements to enjoyment and affinity seen in subjective evaluations are thought to be secondary to the improvement in smooth nonverbal communication. Moreover, increases in memory show MoPaCo gives more impressive images and possibly has the effect of making users engage more actively in conversation.

 ISBN: 978-1-61208-325-4

VII. CONCLUSION

This paper described evaluations of our proposed MoPaCo window interface system, which allows the size and positional relationships of two remote spaces to be reproduced using one stationary camera. The results obtained in implementing the system and performing evaluation experiments on it show it allows gazes and pointing gestures to be transmitted in a similar way to transmitting them through an actual glass window. We also performed experiments to determine whether indicative actions, which are important in performing remote indicative work, could be smoothly referenced with the system. Experiment results suggest MoPaCo users can accurately identify target objects as if under face-to-face conditions through an actual glass window. Results of experiments on conversation quality show the system facilitates smooth conversation and communication and strengthens memories of the conversations, suggesting users actively engage in conversation and the system makes a strong impression on them.

REFERENCES

[1] A. H. Anderson, E. G. Bard, C. Sotillo, G. Doherty-Sneddon, and A. Newlands, "The effects of face-to-face communication on the intelligibility of speech," Perception and Psychophysics, 59, 1997, pp. 580–592.

[2] M. Argyle and J. Graham, "The Central Europe Experiment - looking at persons and looking at things," Journal of Environmental Psychology and Nonverbal Behavior, 1, 1977, pp.6–16.

[3] N. Suzuki et al., "Nonverbal behaviors in cooperative work: a case study of successful and unsuccessful team," Proceedings of the 29th Annual Conference of the Cognitive Science Society, 2007, pp. 195–196.

[4] C. Heath and P. Luff, "Disembodied Conduct: Interactional Asymmetries in Video Mediated Communication," Proceedings of ACM Conference on Human Factors in Computing Systems, 1991, 99–103.

[5] R. Ishii, S. Ozawa, H. kawamura, and A. Kojima, "MoPaCo: high telepresence video communication system using motion parallax with monocular camera," IEEE International Workshop on Human-Computer Interaction: Real-time vision aspects of natural user interfaces (ICCV Workshops), 2011, pp. 463–464.

[6] R. Ishii, S. Ozawa, T. Mukouchi, and N. Matsuura, "MoPaCo: pseudo 3D video communication system," Proceedings of the 1st international conference on Human interface and the management of information: interacting with information - Volume Part II, 2011, pp. 131–140.

[7] S. Whittaker, "Theories and methods in mediated communication," The Handbook of Discourse Processes, New Jersey: Erlbaum, 2003, pp.253–293.

[8] Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell, "Towards a model of face-to-face grounding," Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ALC'03), 2003, pp. 553–561.

[9] P. Auer, "Projection in interaction and projection in grammar," Text, 25, 1, 2002, pp. 7–36.

[10] C. Goodwin, "Conversational organization: Interaction between speakers and hears," Academic Press, New York, 1981.

[11] C. Gale and A. F. Monk, "Where am I looking? The accuracy of video-mediated gaze awareness," Perception and Psychophysics, 62, 2000, pp. 586–595.

[12] T. V. Crater, "The picturephone system: service standards," Bell System Technical Journal, 50, 1971, pp. 235–269.

[13] E. D. Mynatt,, J. Rowan, S. Craighill, and A. Jacobs, "Digital family portraits: supporting peace of mind for extended family members," Proceedings of Conference on Human-Factors in Computing Systems, 2001, pp. 333–340.

[14] Jiejie Zhu, Ruigang Yang, and Xueqing Xiang, "Eye contact in video conference via fusion of time-of-flight depth sensor and stereo," Journal of 3D Research, 2,3, 2011.

[15] S. M. Anstis, J. W. Mayhew, and T. Morley, "The perception of where a face or television portrait is looking," American Jounal of Psychology, 82, 4, 1969, pp. 474–489.

[16] A. Sellen, B. Buxton, and J. Arnott, "Using spatial cues to improve videoconferencing," Video proceedings of Conference on Human-Factors in Computing Systems (CHI), 1992, pp. 651–652.

[17] S. Tanaka, K. Okada, S. Kurihara, and Y. Matsushita, "Desktopconferencing System Using Multiple Still-Pictures: Desktop-MAJIC," Proceedings of the ninth conference on European Conference on Computer Supported Cooperative Work (CSCW), 1996, pp.5–6.

[18] R. Vertegaal and Y. Ding, "Explaining effects of eye gaze on mediated group conversations: amount or synchronization?", Proceedings of the 2002 ACM conference on Computer supported cooperative work, 2002, pp. 41–48.

[19] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2003, pp. 521–528.

[20] K. Okada, F. Maeda, Y. Ichikawa, and Y. Matsushita, "Multiparty videoconferencing at virtual social distance: MAJIC design," Proceedings of the ninth conference on European Conference on Computer Supported Cooperative Work (CSCW), 1994, pp. 383–393.

[21] D. T. Nguyen and J. Canny, "MultiView: improving trust in group video conferencing through spatial faithfulness," Proceedings of Conference on Human-Factors in Computing Systems (CHI), 2007, pp. 1465–1474.

[22] H. Ishii and M. Kobayashi, "ClearBoard: a seamless medium for shared drawing and conversation with eye contact," Proceedings of Conference on Human-Factors in Computing Systems (CHI), 1992, pp. 525–532.

[23] H. Kuzuoka, "Spatial workspace collaboration: a sharedView video support system for remote collaboration capability," Proceedings of ACM Conference on Human Factors in Computing Systems, 1992, pp. 533–540.

[24] Tobii glass, [retrieved: http://www.tobiiglasses.com/scientificresearch/, January, 2014]

[25] M. Kipp, "Anvil - a generic annotation tool for multimodal dialogue," Proceedings of the 7th European Conference on Speech Communication and Technology, 2005, pp. 1367–1370.

[26] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), 2000, pp. 1330-1334.

 ISBN: 978-1-61208-325-4