Video Communication System Supporting Spatial Cues of Mobile Users

Keiji Hirata, Yasunori Harada, Toshihiro Takada, Naomi Yamashita, Shigemi Aoyagi,
Yoshinari Shirai, Katsuhiko Kaji, Junji Yamato, Kenji Nakazawa
NTT Communication Science Laboratories
hirata@brl.ntt.co.jp

Abstract

We propose a novel video mediation method that immerses remote users in a virtual shared space. In the implemented system using this method, video cameras and screens surround users, and on the screens placed behind them remote users and physical or virtual objects are all shown in life-size. Unlike conventional video conferencing systems, the method can support the user's mobility within a shared space and spatial cues exchanged by users. We introduce two properties of a shared space, sharedness and exclusiveness, and compare our method with conventional ones in light of these two properties. We present t-Room, the prototype video communication system that employs the proposed method. Furthermore, to study the cases where two t-Rooms of different layouts are connected, we introduce three parameters for properly describing the geometrical relationships of cameras, screens and users.

1. Introduction

Even though the quality of video and audio components of video conferencing systems (VCSs) has been remarkably improved, VCSs are still far from replacing face-to-face communications. One of the reasons for this is the lack of spatial cues (i.e. spatial relationships occurring between people and objects), which have been shown as critical to group activities [9]. For example, gaze direction is considered a very important spatial cue in the regulation of turn-taking during communication in larger groups [5].

Given that spatial cues have such an importance, a variety of systems are being developed to support spatial cues across distant sites (Hydra [13], MAJIC [11], Gaze-2 [16], MultiView [9,10]). However, most of these systems support spatial cues only when the users are positioned and remain still at a given location; in other words, these systems do not preserve spatial cues when users move within the space. While most VCSs do not allow users to move around the space, such ability to move within a shared environment (i.e. mobility) is regarded as one of the important affordances of face-to-face communications. The mobility gives users greater flexibility in adapting each other's perspectives; for example, a user moves closer to see what an opponent is looking at [6].

In this paper, we propose video mediation technologies that support spatial cues of users between distant sites, where users can freely move around inside a system (called mobile users). We then describe the design and implementation of the t-Room system, in which these technologies are integrated.

This paper is organized as follows. First, we examine two typical camera and screen layouts for VCSs. To clarify the distinction between the proposed method and conventional ones, we introduce the two properties of sharedness and exclusiveness for a shared space. Afterwards we present our ongoing project of building a prototype system that explores the possibilities of our method. In the Discussion section, we study practical cases where two t-Rooms of different camera and screen layouts are connected. Finally, we conclude this paper by describing our future work and perspectives.

2. Camera and screen layout for supporting mobility and spatial cues

We start by considering the geometrical relationships that arise among users in face-to-face interaction and how they dynamically change.

2.1 Geometrical relationships in face-to-face interaction

Consider this example: If the person A sees the person B from a diagonal perspective, then person B also sees person A from a diagonal perspective (Figure 1). The

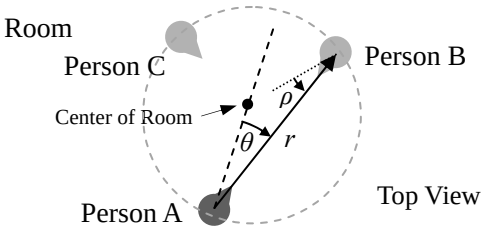


Figure 1: Parameters *r*, *θ*, and *ρ* for face-to-face interaction

figure depicts the top view of the geometry among users, thus a third dimension (height) dimension is neglected here. The apparent distance of A from B, r, is assumed to match the apparent distance of B from A. When A sees B at angle θ with regard to the center of the room, B necessarily sees A at the rotation angle ρ, which is the same as θ. If A moves, then the apparent direction and distance of A from B correspondingly change, and vice versa. Moreover, their positional relationship is also immediately apparent to a third person C who is looking at them from the side.

These phenomena, which people take for granted when they perform face-to-face interaction in the same room, are rarely conveyed by conventional VCSs. We argue that preserving the same values of parameters, r, θ, and ρ, as those in face-to-face interaction would help to support spatial cues of mobile users between distant rooms.

2.2. Front screen versus surrounding back screen

There are two methods for reproducing face-to-face interaction in VCSs: (1) separating the space of face-to-face interaction in Figure 1 into each subspace for a user (location) and projecting remote users to front screens (Figure 2), and (2) duplicating the space in Figure 1 and projecting remote users to surrounding back screen (Figure 3). The shadowed areas in the figures depict the areas where a video camera correctly captures users and objects in front of the camera. In both methods, to project users and objects where they should appear, the physical relationships characterized by the values of r, θ, and ρ among local and remote users and objects must be the same as in face-to-face interaction; this necessarily leads to projecting at life-size on a screen.

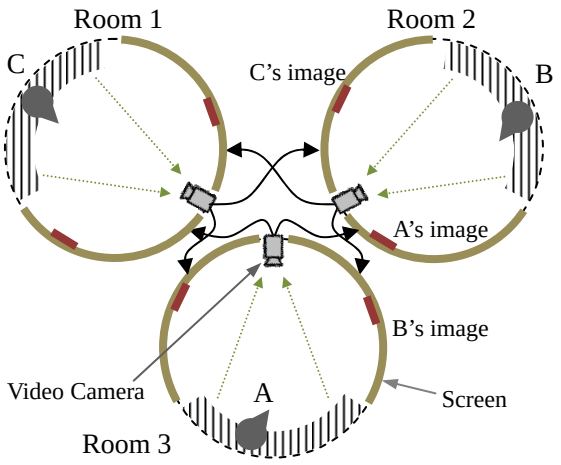


Figure 2: Splitting space and projecting to front screen

Conventional VCSs, and variations of them, employ the front screen method. This arrangement is characterized by having no screens behind the user in the shadowed area and the ability of the user to move around freely only within the area (Figure 2). For example, B's image is distributed to A and C through the right-hand front screen of A in Room 3 and the left-hand screen of C in Room 1. The shadowed area should not be overlapped onto any screen.

As an alternate method, to duplicate a space, we arrange cameras and screens so that they surround users and place screens behind users (Figure 3). As with the first method, the video camera capturing B is distributed to Rooms 1 and 3, but the preprocessing denoted by [−] and [+] in the figure is needed. The function of [−] is to extract only the light from real objects in front of the opposite screen and to cancel the light from the screen. That of [+] is overlapping or superimposing two images captured in Rooms 1 and 2 to correctly place images where they should be projected. In Figure 3, the entire wiring is omitted for simplicity.

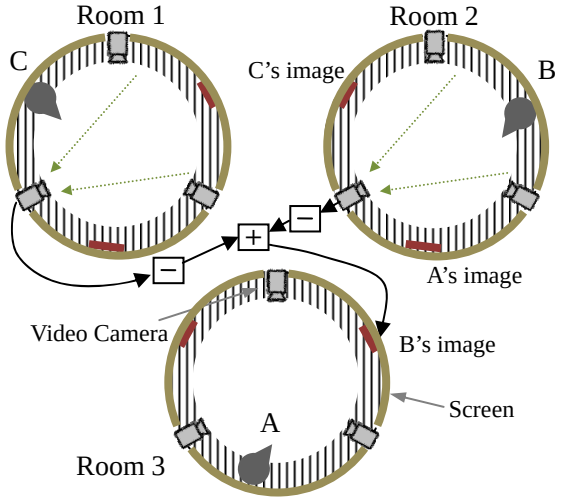


Figure 3: Duplicating space and projecting to surrounding back screen

2.3. Sharedness and exclusiveness

When we construct a shared space by a particular method, the method may be characterized by how this is done and what parts each distributed room shares and does not share. Consequently, we introduce two properties: sharedness and exclusiveness among local and remote rooms. We define sharedness to mean the total angle of a camera view, that is, which part of a scene or perspective in a room is projected to the screens of the other rooms. For example, the video camera of Room 3 in Figure 2 obtains a 120-degree view, but does not capture the remaining part of the room. Imagine the case where

an object is placed just in front of the right-hand screen of A. Although A can naturally see it, neither B nor C can because no camera captures the object in Room 3. Therefore, in Figure 2, 1/3 of the scene of each space is projected to the others, and we can only share the information of the captured part.

For exclusiveness, we first consider the area that a single user occupies and cannot be penetrated by another user. In the front screen method in Figure 2, for every room, there is an area excluded for other rooms, designated by the shadowed area. Although a user can freely move around within this area, the user cannot leave the area; in Figure 2, the area for each user spans 1/3 of the circumference. This ratio can be used as the index of space exclusiveness.

Next, let's consider the sharedness and exclusiveness of the surrounding back screen method in Figure 3. Since for the entire field of vision of every user, WYSIWIS ideally holds, the surrounding back screen method achieves full sharedness and the minimum exclusiveness. Accordingly, wherever a user is, spatial cues are correctly exchanged between moving users.

Sharedness and exclusiveness are highly related to direct pointing capability, which plays a crucial role in CSCW as well as in face-to-face interaction. Therefore, for direct pointing to work correctly, all of the users, including himself/herself, must be able to see the target object at its correct position, and he/she must be able to move to the place of a target object. That is, full sharedness is required, and exclusiveness must be avoided. Basically, users in the front screen method cannot perform direct pointing, while users in the surrounding back screen method can do this.

Some VCSs implement a shared plane by a method similar to the surrounding back screen method, which makes direct pointing possible [8,7,15]. However, because video cameras and screens in these systems do not surround users, full sharedness is not achieved. To resolve the presence disparity problem [14] closely related to sharedness and exclusiveness, some systems, such as ClearBoard [4] and VideoArms [14], successfully provided an equivalent rich awareness of a workspace by presenting only the parts of a body that appear within a workspace (e.g., arms and faces).

3. t-Room system

Based on the above discussion, we aim to demonstrate and explore the surrounding back screen method by developing a prototype system, called t-Room [3,17,2]. The t-Room system is designed as simply as possible to meet the demands of various styles of group activities.

3.1. Hardware design

Figures 4 and 5 show the hardware configuration of

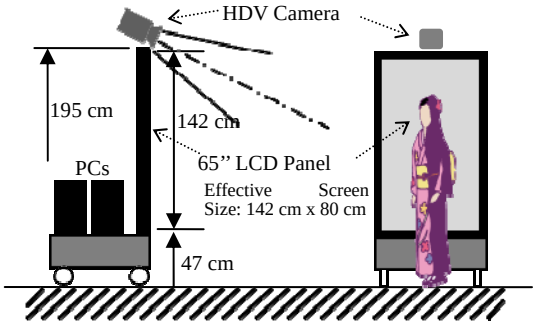


Figure 4: A "Monolith" building module: side view (left) and front view (right).

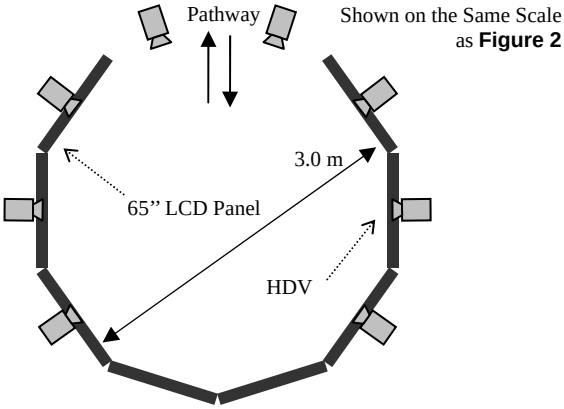


Figure 5: Top view of t-Room system of eight Monoliths arranged decagonally.

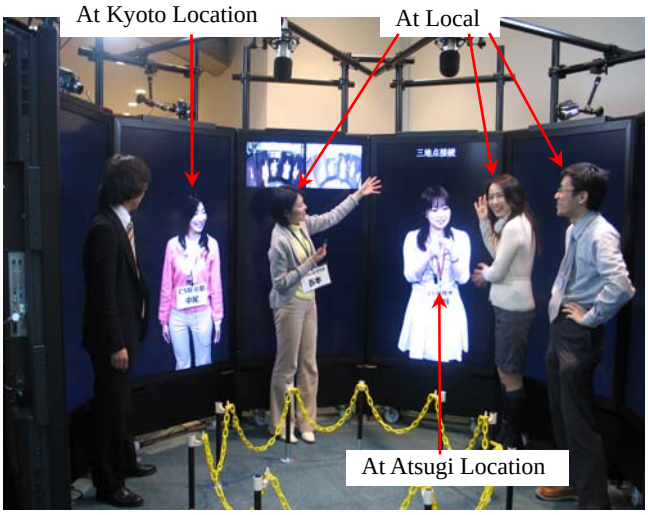


Figure 6: Demonstrating a t-Room made from eight Monoliths arranged decagonally, connecting three locations. Five central Monoliths shown in this photo.

the current t-Room system. A single t-Room consists of eight building modules (called Monoliths) arranged polygonally. With this setup, t-Room encloses a user space with surrounding LCD displays showing life-sized

images. The enclosed space is shared with other enclosed spaces by overlapping it onto them. As a result, users can freely come from and go into others' spaces, since there is no spatial barrier separating users such as the screen in a conventional videoconferencing system. Consequently, the overlapping enclosed spaces can provide full sharedness and minimum exclusiveness.

We installed three nearly identical t-Rooms in our labs located in Atsugi City and Kyoto Prefecture (Atsugi is in the Tokyo area, and Kyoto is approximately 400 km away from Tokyo). Currently, commercially available 100-Mbps optical fiber lines connect the one at Atsugi and the two at Kyoto.

Figure 6 shows a partial view of a working t-Room system in which four persons are standing and the other two are displayed on the screen alternatively. At "local" (one of the Kyoto t-Rooms), three spaces are overlapped: Atsugi, the other Kyoto, and local itself. These spaces are similarly overlapped at Atsugi and the other Kyoto. At local, the images of Atsugi and Kyoto are displayed by overlapping with a transparency ratio of 0.6 at present.

3.2. Standing position

The image displayed in the other t-Room depends highly on the standing position of a user. Standing as close as possible to the LCD panel of a Monolith is preferable, since a camera can correctly capture a user's view at the restricted area, corresponding to the shadowed area in Figure 3. In fact, when a user walks away from the LCD surface and moves toward the center of the t-Room, the displayed image of the user is magnified, and that person's life-size appearance is lost. Furthermore, the user is then captured not only by a front camera but also by the others (possibly both sides of the front camera), which leads to more than one image at different angles being displayed. To cope with the problem, the yellow chain is placed as shown in Figure 6.

3.3. Gaze

The geometrical relationships observed in Figure 7 show people's positions in gaze direction, face image orientation, and camera angles. In Room 1 of Figure 7, person B looks to his/her forward-right, and the camera captures B's face image from the front. However, considering the Mona Lisa effect, B's face image displayed in Room 2 appears too far to the side for person A. The same phenomenon takes place in terms of A's image as B looks at it. Therefore, projecting an image at a photorealistic rotation angle may not always be appropriate for exchanging spatial cues.

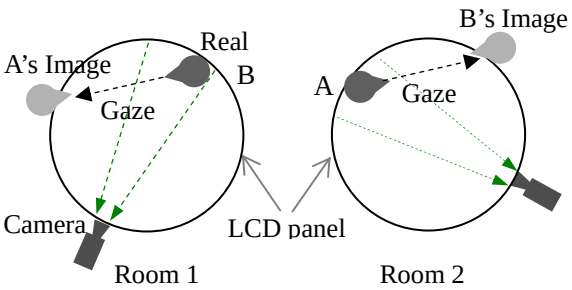


Figure 7. Gaze and face image orientation

Using our t-Room, through the process of interaction and collaboration, awareness of users' positions and body/face orientations are dynamically organized and shared among users within different t-Rooms. As a result, it was sometimes observed that such awareness compensates gaze error caused by the Mona Lisa effect and the photorealistic face image orientation to some extent.

4. Discussion

Even if full sharedness and minimum exclusiveness are achieved, in the case where two t-Rooms of different layouts are connected, we will still find areas where we cannot exchange correct spatial cues of mobile users. We demonstrate that by examining the values of r (distance between A and B), θ (direction of opposite person's image to the normal of the back screen), and ρ (rotation angle of opposite person's image to the normal of back screen), we can discriminate preferable situations from the others (Figure 8). The figure depicts the top view, so the height dimension is neglected.

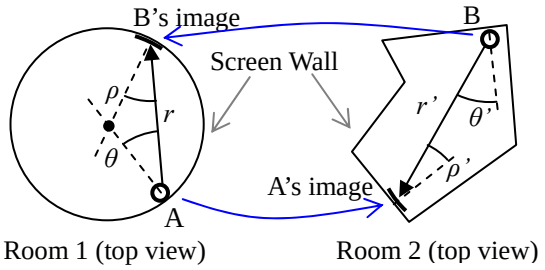


Figure 8: Comparison of physical parameters *r*, *θ*, and *ρ*

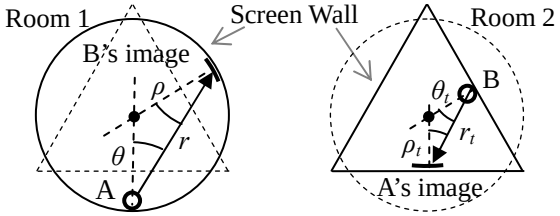
Here, we assume that one t-Room (Room 1) always has the circular layout of screens in which person A stays, and the other (Room 2) has one of the following four shapes in which person B stays: triangle, smaller circle, parallel segments, and circle with gap (the 'U' shape). For each case, except for the second, we assume that the two connected t-Rooms have the same overall circumference for projecting images at life-size, despite the different shapes. A camera captures a user's image from the direction of the normal of his/her back screen, that is, the direction perpendicular to the back panel surface[1].

[1] In fact, a camera cannot always capture a user from the direction of normal of his/her back screen, when he/she appears

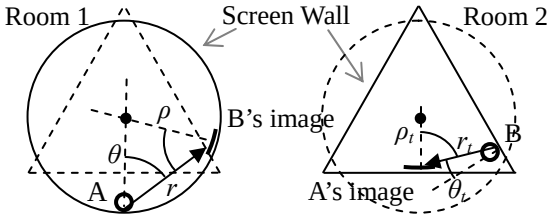
4.1. Circle connected to triangle

Figure 9 shows the values of r, θ, and ρ for Room 1 (circle) and those of r_t, θ_t, and ρ_t for Room 2 (triangle), and there are two patterns of user positions: (a) A's image and B in Room 2 are positioned at the centers of the triangle's sides, and (b) person B in Room 2 is positioned close to the triangle's corner.

For (a) in Figure 9, the values of r, θ, and ρ are well preserved, and only r is shortened to r_t in Room 2. In contrast, for (b), we find the difference that ρ is positive, while θ_t is negative[2], which may cause a misleading spatial cue, although full sharedness holds. This suggests that a user should not move close to the corner of an acute angle.



(a) Facing forward-right image



(b) Person B close to a corner of triangle

Figure 9: Triangle

4.2. Circle connected to smaller circle

In Figure 10, to project person A in Room 1 onto the screen of Room2 with full sharedness maintained, their sizes must be reduced, and in contrast, we need to magnify the images upon projection from Room 2 to Room 1. Apparently, we have $r > r_s$, $\theta = \rho_s$, and $\rho = \theta_s$. Since it is widely accepted that a life-sized image shown on a screen is desirable for a sense of presence [1] and appropriate cognitive arousal [12], the scaling may cause an undesirable effect upon exchanging spatial cues, although sharedness and exclusiveness are ideal; for example, it may be difficult to recognize a complicated gaze communication and corporal gestures through small-sized images.

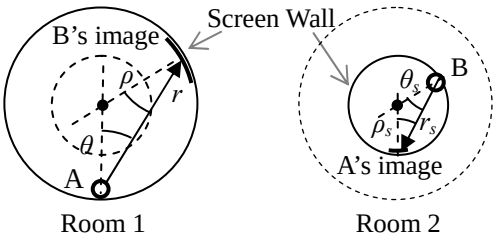
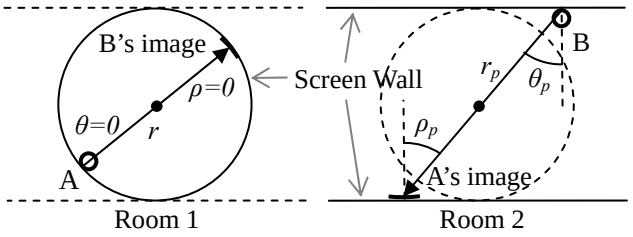


Figure 10: Smaller circle

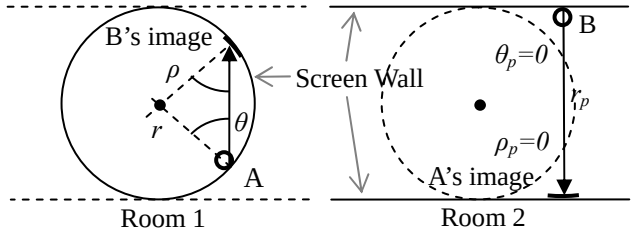
4.3. Circle donnected to parallel segments

In the parallel layout (Room 2 in Figure 11), within the area between the opposite end-points of the parallel segments, users and objects cannot be captured or projected to the other rooms. Within those areas, users in Room 2 can obtain the view of Room 1, but the reverse is not possible; this non-reciprocity of perspective produces a non-shared area, where sharedness does not hold. So, it is desirable to prevent users in Room 2 from staying there. Since the layout of Figure 11 contains such meaningless areas, r_p is, in total, longer than r.

Unlike the triangle case in Figure 9, the signs of θ and ρ_p are no longer reversed, as are those of ρ and θ_p. Around the center of the parallel segments, we can obtain the most accurate values of r_p, θ_p, and ρ_p. Hence, as long as users move around the center, spatial cues occurring in the layout of Figure 11 can be exchanged more accurately than the layout of Figure 9.



(a) Orthogonal



(a) Across to each other

Figure 11: Parallel segments

4.4. Circle connected to circle with gap

A gap in the circle creates a discontinuous screen, and

in the area apart from the center of the view angle. However, in this discussion, the phenomenon is ignored.

[2] For rotation angle, the clockwise direction is assumed to be negative, and the opposite positive.

two points close to each other in Room 1 may be transferred as two distant points in Room 2, which prevents accurate transfer of mobile spatial cues (Figure 12). Furthermore, the gap also produces a non-shared area.

As in Figure 11, around the center of the screen in Room 2, we can obtain the most accurate values of r_g, θ_g, and ρ_g, and the farther A and B in Room 2 are from the

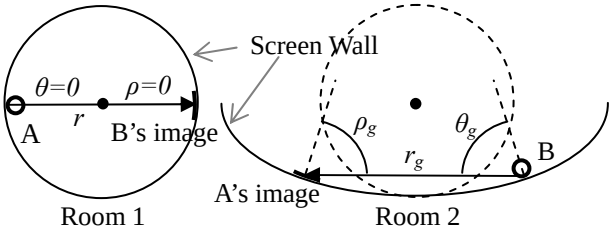


Figure 12: Circle with gap

center, the more mismatches of these values occur.

5. Conclusions

The contribution of this paper is to propose the surrounding back screen method and to introduce two properties and three parameters for classifying a shared space: sharedness and exclusiveness, and r, θ, and ρ. These properties and parameters enable us to compare video conferencing technologies and video mediation methods in a more quantitative manner.

Since the current working t-Rooms have almost identical circle layouts of cameras and screens, the problems pointed out in Section 4 do not occur. However, the more t-Rooms are connected in the future, the more various layouts of cameras and screens will emerge, as studied in Section 4. For such layouts, in spite of achieving full sharedness and minimum exclusiveness, a t-Room system may not correctly transfer mobile spatial cues due to the shape of a particular t-Room. Therefore, we think it is also important to investigate the effects of camera and screen layouts for collaborative work in a practical situation.

Although sound reproduction under this environment is not discussed in this paper, we think that reproducing the acoustic cues of face-to-face interaction is also of relevant importance for remote collaboration. Consequently, a system should transfer the acoustic cues provided by mobile users in a room to the correct positions in the other room, as done for visual information. The current t-Room system just satisfies the minimal requirements for experimental remote collaboration. Therefore, effective techniques of sound engineering (e.g., sound source localization and echo cancellation) should be used to deal with the sound reproduction problem. We believe that it is challenging to design and implement the practical acoustic environment of a t-Room having various camera and screen layouts.

References

[1] W. Buxton, "Telepresence: integrating shared task and person spaces", In *Proceedings of Graphics Interface '92*, pp.123-129.

[2] K. Hirata, Y. Harada, T. Ohno, T. Yamada, J. Yamato, and Y. Yanagisawa, "t-Room: Telecollaborative Room for Everyday Interaction", In *Proceedings of The 66th IPSJ Annual Convention*, 4B-3 (2004).

[3] K. Hirata, Y. Harada, T. Takada, S. Aoyagi, Y. Shirai, N. Yamashita, and J. Yamato, "The t-Room: Toward the Future Phone", *NTT Technical Review*, Vol. 4, No. 12, pp. 26-33 (2006).

[4] H. Ishii, M. Kobayashi, and K. Arita, "Iterative Design of Seamless Collaboration Media", *Communications of the ACM*, 37, 8 (August, 1994), pp. 84-97.

[5] A. Kendon, Conducting Interaction – Patterns of Behavior in Focused Encounters, Cambridge University Press (1990).

[6] R. E. Kraut, S. R. Fussell, Su. E. Brennan, and J. Siegel, "Understanding Effects of Proximity on Collaboration: Implications for Technologies to Support Remote Collaborative Work", In P. J. Hinds and S. Kiesler (Eds), Distributed Work, The MIT Press, pp.137-162 (2002).

[7] H. Kuzuoka, J. Yamashita, K. Yamazaki, and A. Yamazaki, "Agora: A Remote Collaboration System that Enables Mutual Monitoring", In *Proceedings of CHI'99 Extended Abstracts*, 190-191.

[8] O. Morikawa and T. Maesako, "HyperMirror: Toward Pleasant-to-use Video Mediated Communication System", In *Proceedings of CSCW '98*, 149-158.

[9] D. Nguyen and J. Canny, "Multiview: Spatially Faithful Group Video Conferencing", In *Proceedings of CHI2005*, pp. 799-808.

[10] D. Nguyen and J. Canny, "Multiview: Improving Trust in Group Video Conferencing Through Spatial Faithfulness", In *Proceedings of CHI2007*, pp. 1465-1474.

[11] K. Okada, F. Maeda, Y. Ichikawa, and Y. Matsushita, "Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design", In *Proceedings of CSCW'94*, pp. 385-393.

[12] B. Reeves and C. Nass, *The media equation: how people treat computers, television, and new media like real people and places.* Cambridge University Press, 1998.

[13] A. J. Sellen, "Speech Patterns in Video-Mediated Conversations", In *Proceedings of CHI '92*, 49-59.

[14] A. Tang and S. Greenberg, "Supporting Awareness in Mixed Presence Groupware", ACM CHI Workshop on Awareness systems: Known Results, Theory, Concepts and Future Challenges (2005).

[15] J. C. Tang and S. L. Minneman, "VideoDraw: A Video Interface for Collaborative Drawing", In *Proceedings of CHI '90*, 313-320.

[16] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction", In *Proceedings of CHI 2003*.

[17] N. Yamashita, K. Hirata, T. Takada, Y. Harada, Y. Shirai, and S. Aoyagi, "Effects of Room-sized Sharing on Remote Collaboration on Physical Tasks", *IPSJ Digital Courier*, Vol. 3, pp.788-799 (2007).