Face-to-Face Tabletop Remote Collaboration in Mixed Reality

Shinya Minatani*
University of Tsukuba

Itaru Kitahara †
University of Tsukuba

Yoshinari Kameda ‡
University of Tsukuba

Yuichi Ohta §
University of Tsukuba

ABSTRACT

This paper proposes a novel remote face-to-face Mixed Reality (MR) system that enables two people in distant places to share MR space. Challenging issues to realize such an MR system include capturing, sending, and rendering each user's appearance in real time. We developed a method to represent user's upper body and hands on the table as a single deformed-billboard. An MR Othello game is implemented as a test bed of the remote face-to-face MR system. Users can play the tabletop game as if their opponent were sitting across from the table, despite being physically separated. By detecting and sending the status of each real game board to the other site, both users feel that they are sharing tabletop objects.

CR Categories: H.4.3 [Information Systems Applications]: Communications Applications - Computer conferencing, teleconferencing, and videoconferencing; H.5.1 [Information Interfaces and Presentation (e.g., HCI)]: Multimedia Information Systems - Artificial, augmented, and virtual realities; I.3.2 [Computer Graphics]: Graphics Systems - Remote systems.

Keywords: Shared Mixed-Reality, Billboard, Tele-Immersion, Real-Time Modeling and Rendering.

1 INTRODUCTION

In Shared Mixed-Reality (Shared-MR) space, multiple users sharing MR space can improve communication with each other by watching both the real and virtual objects [1][2]. To see each other's appearance, on the other hand, all users have to gather at the same place. Thus, it is impossible to share MR space between distant places. In MR space, however, a user wears a video see-through Head-Mounted Display (HMD) to observe the MR space. Therefore, even if an opponent sharing identical MR space does not exist in identical real space, it is possible to virtually make them feel as if they were located together by correctly capturing, sending, and rendering the appearances of remote opponents in real time, as illustrated in Figure 1. We call this concept "Remote Shared-Mixed Reality."

Tele-immersion have been studied for a long time [3][4]. Recently, as a remote communication technique using MR/AR, many types of tele-collaboration system that used a screen or wall as a spatially immersive display were presented [5][6]. Even though these systems can be used to interact spatially, the users cannot interact with objects in space extended over both users' sides (ex., on a tabletop) because displayable space is limited on the screen. Since nonverbal information (ex., gestures) and spatial interaction is important to communicate naturally, the interacting space must be placed between users in space, not limited on a screen [7]. To easily extend the interacting space on all areas of the tabletop and represent the remote user's appearance naturally, we introduce a method using HMDs and billboard-based modeling/rendering, which uses a deformed-billboard as a virtual spatially immersive display placed in front of the user and the tabletop.

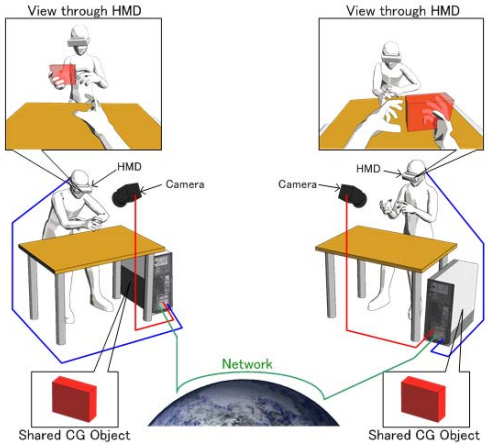


Figure 1. Concept of Remote Shared-MR.

2 REMOTE SHARED-MR SYSTEM

Three important consistencies must be satisfied to merge real and virtual space without causing discomfort to users: geometric, photometric, and temporal (processing delay). We concentrate on temporal consistency because it is the most important for realizing smooth communication. The resolution of appearances of users and/or objects is crucial because it affects the quality of nonverbal communication, such as changes in facial expression. So we designed a remote shared-MR system that focuses on capturing high-resolution video and sending/rendering the appearances of remote users and real objects in real time.

To generate the appearances of remote users and objects, 3D model-based rendering [8][9] or image-based rendering [10][11] are often used. However, these methods require much computational cost for building 3D models and rendering appearances from arbitrary viewpoints. The size of the 3D model is usually huge. As a result, 3D model-based rendering is difficult in real time. When the resolution of captured video is increased, the ability of computers must be increased, too.

Our system employs a billboard-based rendering instead of 3D model-based rendering. Since it does not estimate the 3D shape of objects but the 3D position and orientation of a billboard, just a single camera is required to capture the appearance of target scene, and the modeling and rendering processes do not consume computing cost. When the system captures appearances by a high-definition camera, the quality of the generated video is improved by mapping fine texture onto the billboard. Furthermore, it is possible to send it on a conventional broadband network in video rate, since there is only one capturing camera, and the data amount to locate a billboard in 3D space is not affected by the increase of video resolution.

Since the shape of an ordinary billboard (a vertical flat plane) is usually different from the target shape, there might be visual

*e-mail: minatani@image.esys.tsukuba.ac.jp
† e-mail: kitahara@iit.tsukuba.ac.jp
‡ e-mail: kameda@iit.tsukuba.ac.jp
§ e-mail: ohta@acm.org

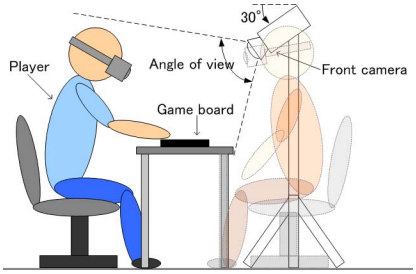


Figure 2. Layout of the proposed system.

distortion in the rendered appearance when the position of an observer's viewpoint moves away from the position of the capturing camera. Thus, it is not practical to apply a billboard-based method to interactions held in a wide area of the MR space. We solve this problem by two measures. First, we assume that interaction is done by two users sitting in front of a table and facing each other, and the interaction area is near a tabletop. Second, we deform the billboard shape to minimize the difference of shapes between target object and billboard as much as possible. We call this novel-shaped billboard a deformed-billboard.

As a test bed of remote shared-MR system that satisfies above conditions, we developed a remote collaboration system based on Othello, a board game that resembles Reversi, which is played by two people facing each other. Our shared-MR system enables two remote users to naturally play Othello by capturing, sending, and rendering the opponent's appearance and the game board.

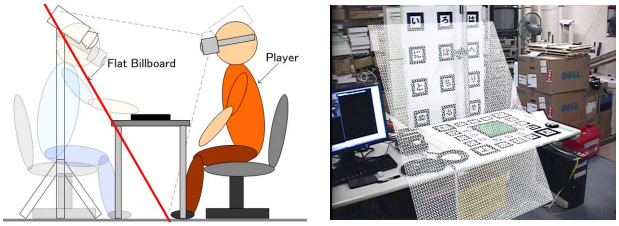
3 BILLBOARD-BASED RENDERING

Figure 2 shows the layout of our proposed remote shared-MR system. The two users sit 'virtually' facing each other. We placed a camera with a super-wide-angle lens at each site to share visual information of the distant (opponent) site. Each camera is fixed in front of the opponent and tilted 30° downward to capture the entire opponent's appearance and the tabletop. Here, the camera position should almost be identical to the viewpoint of remote user. When the remote user gazes around (rotating line of sight), the system generates the view by virtually rotating the captured image with 2D projective transformation. When the user moves (shifts) his/her viewpoint, the system generates this view by using the billboard as a 3D shape.

Figure 3 shows the layout of a "flat billboard" that faces the front to a virtual viewpoint from which the target space is observed. Thus, the flat billboard is tilted 30° backwards. The picture on the right side hand in Figure 3 shows an overlapping view of the flat billboard on the target scene. When we map texture on the billboard without considering distortion of the super-wide-angle lens, the appearance is distorted. This distortion can be removed by geometric transformation using the camera's intrinsic parameters[12]; however, this process is time-consuming.

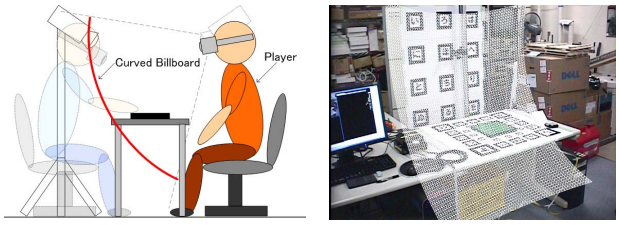
We developed a method to undistort the appearance using a "curved billboard" shown in Figure 4. It is spherically curved to fit into the imaging geometry of the lens, and then lens distortion can be removed by mapping the captured image on the billboard. When the PC has a graphic accelerator, this process is done by hardware. The curved billboard has another advantage: The curved shape reduces the difference of shapes between the target object and the billboard.

By using a curved billboard, the appearance of the remote user's upper body is represented well, since the 3D pixels displayed on the billboard are located near the surface of the



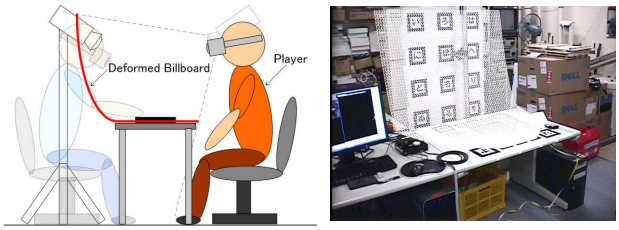
(a) Shape and position. (b) Visualized in real scene.

Figure 3. Flat billboard.



(a) Shape and position. (b) Visualized in real scene.

Figure 4. Curved billboard.



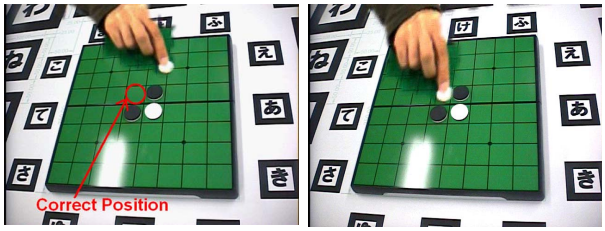
(a) Shape and position. (b) Visualized in real scene.

Figure 5. Deformed billboard.

user's body. However, the appearances of arms and hands of the remote user on the tabletop are not represented well, since the 3D pixels on the billboard under the table are located apparently different from the 3D positions of the user's body above the table. When the user does not move his/her head at all, the discrepancy of appearance is not so large; on the other hand, when the user moves his/her head, as shown in Figure 6(a), the appearance of hand on the tabletop is strongly out of position. The remote user tries to put a piece on the marked position in the figure.

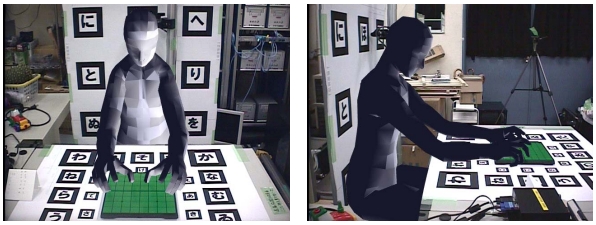
We developed a novel billboard called a "deformed-billboard" to solve this problem. As shown in Figure 5, the upper part of the billboard, where the appearance of the remote user's upper body is mapped, is curved to fit into the imaging geometry of the lens. The lower part of the billboard, where the appearances of the user's arms and hands on the tabletop are mapped, is fit to the tabletop surface. These two areas are joined smoothly to make a single billboard, and their mapping grids are deformed to adjust the scale and to compensate the distortion of mapped appearance. As a result, it is possible to naturally display the appearance of the remote user's hand at a correct position, even if the user moves his/her head, as shown in Figure 6(b). In both pictures of Figures 6(a) and (b), the remote user is trying to put a white piece on an identical square.

To evaluate the validity of the deformed billboard, we measured the distance between the surfaces of a billboard and a 3D human model shown in Figures 7(a) and (b). In Figures 7(c), (d), and (e), the distance is indicated as a grayscale map, where darker means



(a) With curved billboard. (b) With deformed billboard.

Figure 6. Comparison of curved/deformed billboards observed from inclined head.



(a) 3D Human Model (Front). (b) 3D Human Model (Right).

(c) Flat billboard. (d) Curved billboard. (e) Deformed billboard.

Figure 7. Distance map between billboard and 3D human.

larger difference. As a result, since the distance of the deformed billboard is smaller than the others, the deformed billboard can reduce the distortion of the rendered appearance near the tabletop.

4 Implementation

Figure 8 presents an overview of our prototype Remote Shared-MR system. Each player is sitting in front of a table on which a game board is placed. A camera, front camera, is set at the opposite side of the table. A super-wide-angle lens is attached to the camera and its optical axis is tilted 30° downward to capture both the user's body and the objects on the tabletop. Status of the real objects on the table, *i.e.*, pieces on the game board, is extracted from the captured image. Each player wears a video see-through HMD with built-in binocular cameras. To estimate the pose of HMD, a multi-marker method of the ARToolkit [13] is used. The cameras and the HMD are connected to a PC at each player's site, and the two PCs are connected via a network. Players' appearances and the statuses of tabletop objects are transmitted in real time to each other.

The processing flows of our system are shown in Figure 9. There are two stages in this implementation; image capturing and MR space rendering. They are executed at each player's site in video rate.

4.1 Capturing player's appearance

Figure 10(a) shows an image captured by the front camera and transmitted to the PC by IEEE1394b interface. A mask image

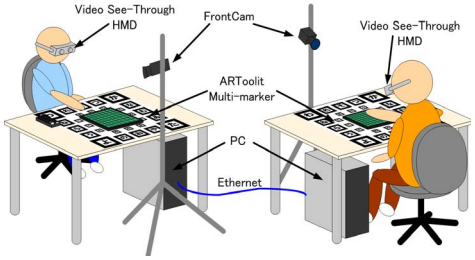


Figure 8. Overview of remote shared-MR test bed.

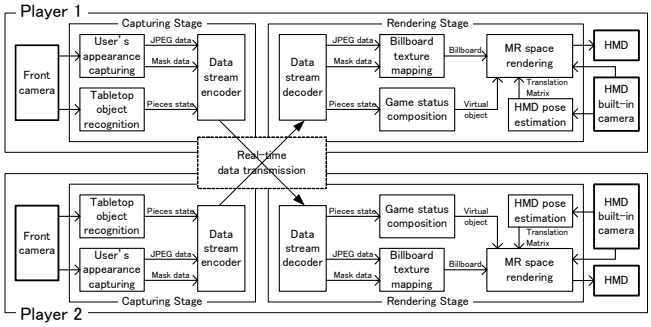


Figure 9. Processing stages implemented in the test bed.

representing the player's silhouette is generated by background subtraction. We assume that the background image is captured in advance. Then the player's region is extracted as texture from the captured image and compressed by JPEG. For each frame, the JPEG data of the texture, the run-length coded mask image, and other additional data (e.g., data length and image size) are packeted up into a data stream.

4.2 Recognition of tabletop objects

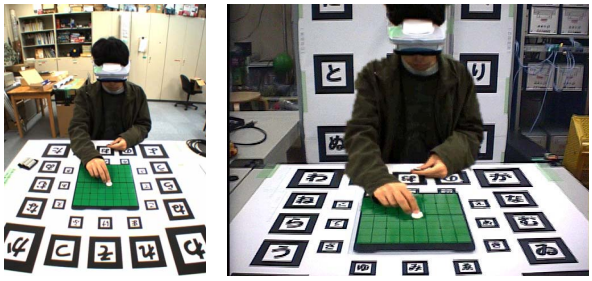
In our test-bed Othello game system, the positions and the colors of all pieces on the game board are automatically recognized from the images captured by the front camera. Othello pieces are either white or black, and the surface of the game board is green. Then we implemented a simple recognition method based on the L*a*b* color space. The center pixel of every square on the game board, which is calibrated beforehand, is always scanned to detect a piece using the following algorithm:

1) If both of a* and b* are less than a threshold, there is a piece; otherwise there is no piece.
2) If L* is less than a threshold, the piece is black; otherwise it is white.

To prevent misrecognitions, if the position of a scanning pixel is in the mask region, *i.e.,* occluded by player's hands, a former state is kept. When a player puts a new piece on a square of the game board or he/she reverses a piece on a square, the status of the square is recognized as "changed." Then the new status is sent to the opponent as additional data in the packet.

4.3 Real-time data transmission

We implemented a transmission scheme combined with TCP for synchronization and UDP for data streaming. A data stream is split into lengths shorter than the maximum data size of UDP and transferred continually. To synchronize frames and to control delays, synchronization signals are sent to each other through TCP at every frame.



(a) Front camera view (b) Remote user's view through HMD

Figure 10. Snapshot of our remote shared-MR system.

4.4 Billboard texture mapping

After receiving a data stream from the opponent, a smoothing filter is applied to the mask image to shade the edge of the mask region and it is jointed as the alpha value of the texture image. The texture image with alpha value is mapped on the deformed billboard using Alpha Blending to naturally merge the appearance of remote player with the real scene.

4.5 HMD pose estimation

We use a multi-marker method of the ARToolkit to estimate the pose of HMD. Many markers are located on the tabletop and on a flat board behind the front camera. The markers are studded so that the opponent's appearance and the objects on the tabletop can always be rendered, regardless in what direction the player's HMD turns. When a built-in HMD camera captures an image containing some markers, the translation matrix from the datum point of the Shared-MR space to the HMD can be estimated.

4.6 MR space rendering

The rendering of MR space is necessary for both of right and left eyes of the HMD, and it must be video-rate. First, the image captured by each built-in binocular camera of the HMD is rendered as the background real world of the MR space. Secondly, a deformed billboard, on which the remote player's texture transmitted via network is mapped, is placed at a correct position in the virtual space. Thirdly, the pieces captured from the remote player's game board are rendered as CG objects at their correct positions in the virtual space. Finally, to follow the position and orientation the player's head, the transformation matrix obtained in Section 4.5 is multiplied to the virtual space. By executing these processes in video rate with small latency, the players feel naturally in the remote shared-MR space as if they are facing each other in the real world.

5 Experimental Results

In our experimental environment, using Dell XPS710 with Intel Core2Extreme 2.67 [GHz] as the player's PC, PGR Flea2 (IEEE1394b camera, 768x1024 [pixels], 30 [fps]) as the front camera, and video see-through HMD with 640x480 [pixels] LCD panels and NTSC built-in binocular cameras, the frame rate was between 15 [fps] and 30 [fps]. The frame rate depends on the mask size and the error rate of the network. The bit rate was less than 130 [KB/frame], and delay was less than 1 [sec] at the local host or by directly connected 1000BASE-T Ethernet.

As shown in Figure 10(b), the remote player's appearance looks so natural that it is possible to play Othello without facing each other in the real world. Only with a single deformed billboard, the appearances of the upper body and the hands are represented well, and the objects on the tabletop are represented at correct positions in the real world.

6 Conclusion

We introduced a remote face-to-face tabletop collaboration system by using Mixed Reality technology. As a test bed of our remote shared-MR, a remote MR Othello game system was implemented. By using a novel deformed-billboard with real-time image data streaming, the remote opponent's appearance was represented naturally and his hands placed near the tabletop were appeared at correct positions in the real world. The pieces on the game board at one site were detected and displayed at the other site as CG objects to support natural playing of the game.

By implementing a method to detect and display known objects on the tabletop, as well as Othello, various applications can be realized in our Tabletop Remote Shared-MR scheme.

In the current system, however, some problems remain unsolved; the player's hands in the real world can be occluded by the texture and/or the virtual objects mapped on the billboard. A possible solution will be to segment the player's hands in the image captured by the built-in HMD camera and to render them above the billboard.

Reference

[1] Y. Ohta and H. Tamura. Mixed Reality - Merging Real and Virtual Worlds -. Splinger-Verlag, 1999.

[2] K. Kiyokawa, H. Iwasa, H. Takemura and N. Yokoya. Collaborative Immersive Workspace through a Shared Augmented Environment. In Proc. of SPIE '98, Vol.3517, pp.2-13, 1998.

[3] H. Takemura, Y. Kitamura, J. Ohya and F. Kishino. Distributed Processing Architecture for Virtual Space Teleconferencing. In Proc. of ICAT'93, pp.27-32, 1993.

[4] S. Sugawara, G. Suzuki, Y. Nagashima, M. Matsuura, H. Tanigawa, M. Moriuchi. InterSpace: Networked Virtual World for Visual Communication. In IEICE transactions on information and systems, E77-D(12), pp.1344-1349, 1994.

[5] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In Proc. of SIGGRAPH'98, pp.179–188, July 1998.

[6] J. Mulligan, K. Daniilidis. View-independent scene acquisition for tele-presence. In Proc. of ISAR2000, pp.105-108, 2000.

[7] K. Kiyokawa, M. Billinghurst, S. E. Hayes, A. Gupta, Y. Sannohe, and H. Kato. Communication Behaviors of Co-located Users in Collaborative AR Interfaces. In Proc. of ISMAR2002, 2002.

[8] H. Towles, Wei-Chao Chen, R. Yang, Sang-Uok Kum, H. Fuchs, N. Kelshikar, J. Mulligan, K. Daniilidis, L. Holden, B. Zeleznik, A. Sadagic, and J. Lanier. 3D Tele-Collaboration Over Internet2. In Proc. of ITP2002, 2002.

[9] K. Daniilidis, J. Mulligan, R. McKendall, G. Kamberova, D. Schmid, and R. Bajcsy. Real-Time 3D Tele-immersion. In the Confluence of Vision and Graphics, A. Leonardis et al. (Eds.), Kluwer Academic Publishers, pp. 253-266, 2000.

[10] E. Cooke, P. Kauff, and O. Schreer. Image-Based Rendering for Tele-Conference Systems. In Proc. of WSCG2002, 2002.

[11] P.J. Narayanan, P. Rander, and T. Kanade. Constructing Virtual Worlds Using Dense Stereo. In Proc. of ICCV'98, pp. 3 - 10, 1998.

[12] B. Watson, L. F. Hodges. Using Texture maps to Correct for Optical Distortion in Head-Mounted Displays. In Proc. of the Virtual Reality Annual Symposium '95, IEEE Computer Society Press, pp.172-178, 1995.

[13] H. Kato and M. Billinghurst. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In Proc. of IWAR99, ACM, pp.85–94. 1999.