ImmerseBoard: Immersive Telepresence Experience using a Digital Whiteboard

Keita Higuch[1], **Yinpeng Chen**[2], **Philip A Chou**[2], **Zhengyou Zhang**[2], **and Zicheng Liu**[2]

[1]The University of Tokyo
[2]Microsoft Research

ABSTRACT

ImmerseBoard is a system for remote collaboration through a digital whiteboard that gives participants a 3D immersive experience, *enabled only by an RGBD camera (Microsoft Kinect) mounted on the side of a large touch display*. Using 3D processing of the depth images, life-sized rendering, and novel visualizations, ImmerseBoard emulates writing side-by-side on a physical whiteboard, or alternatively on a mirror. User studies involving three tasks show that compared to standard video conferencing with a digital whiteboard, ImmerseBoard provides participants with a quantitatively better ability to estimate their remote partners' eye gaze direction, gesture direction, intention, and level of agreement. Moreover, these quantitative capabilities translate qualitatively into a heightened sense of being together and a more enjoyable experience. ImmerseBoard's form factor is suitable for practical and easy installation in homes and offices.

Author Keywords

Immersive Experience; Collaboration; Telepresence

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Computer-supported cooperative work

INTRODUCTION

A physical whiteboard can enhance collaboration between people in the same location by allowing them to share their ideas in written form. The existence of the written representations in turn allows the participants to express their relationships to the ideas in physical terms, through pointing, gaze direction, and other forms of gesture. These are important ways, besides the written information itself, that a physical whiteboard enhances collaboration beyond the usual important elements of collaboration between co-located people, such as eye contact, body posture, and proxemics.

When collaborators are remote, a digital whiteboard makes it possible for remote collaborators to share their ideas graphically. Digital whiteboard sharing is a facility found in many modern video conferencing systems. However, it is mostly used to convey information through writing. The ability for the participants to relate with each other and with the writing through pointing, gaze, and other forms of gesture is often lost. Preserving such context, as if the participants were co-located, has been a goal of research on remote collaboration for some time [26][7]. The well-known Clearboard [7] deeply affected the remote collaboration field. It shows the video of the remote participant on a shared workspace as if the participants talk through and draw on a transparent glass window. However, Clearboard has several limitations: (a) it requires a rear projector/camera and special screen (either liquid crystal screen switched between transparent and scattering states, or 45 degree tilted projection screen with a polarizing film and half-silvered mirror), whose bulk and cost make large deployment difficult, (b) gaze is correct only when both participants' heads are simultaneously located at the virtual camera positions, (c) collaborating through a glass window is not as familar for users as collaborating in front of a whiteboard, and requires an unexplained image flip, and (d) the writing and remote participant's video are overlapped, which may distract participants. The metaphor that participants talk in front of a whiteboard was discussed in the Clearboard paper and was considered too difficult to implement without using head-mounted displays and special gloves.

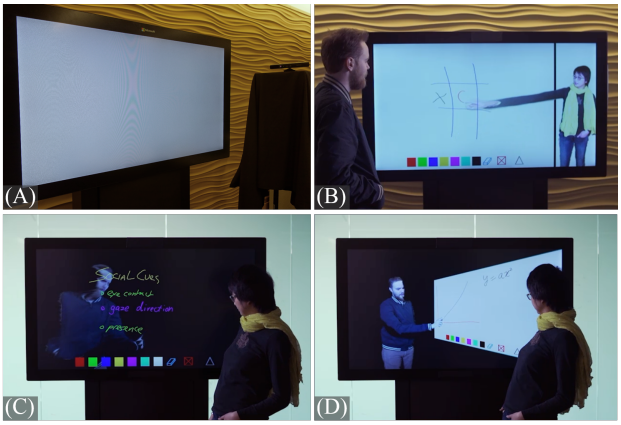

Figure 1. ImmerseBoard setup and conditions. (A) Large touch display and a Kinect camera, (B) Hybrid, (C) Mirror, (D) Tilt board.

In this paper, we propose ImmerseBoard, which provides an immersive telepresence experience [15] around remote whiteboard collaboration with a simple setup, using only a large touch display and an RGBD camera. ImmerseBoard preserves the remote participants' physical relation to the whiteboard and to each other, while overcoming ClearBoard's lim-

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

http://dx.doi.org/10.1145/2702123.2702160

itations. We design and implement a prototype system, which supports three novel immersive conditions, called Hybrid, Mirror, and Tilt board conditions, shown respectively in Figures 1(B)–(D). The Hybrid condition is an augmentation of 2D video conferencing with a whiteboard, extending the remote person's hand out of the video window to reach the location where he or she is writing. The Mirror condition emulates side-by-side collaboration while writing on physical mirror. Though visually similar to ClearBoard, the Mirror explains the flip and extends easily to multiple parties. The Tilt board condition emulates side-by-side collaboration while writing on a physical whiteboard. This has not been possible before without a head-mounted display. Key contributions include the following:

- We use *only an RGBD camera (Microsoft Kinect), mounted on the side of a large touch display*, to enable 3D immersive collaboration in a desirable form factor, practical for home or office use.
- We introduce three new visualization metaphors, including the completely new 2.5D Hybrid and 3D Tilt visualizations, which provide to remote collaborators a sense of the spatial relationships to each other and to their shared writing, thereby preserving varying degrees of gaze direction, gesture direction, intention, and proximity, for immersive collaboration.
- We implement all three visualizations (Hybrid, Tilt, Mirror) in a single system, and allow users to choose their preferred visualization depending on task. To our knowledge, this is the first implementation of any of these visualizations in a simple and practical setup.

We also run a user study to validate the system. In the user study, we design three games that reflect important aspects of real-world collaboration. A total of 32 participants in 16 pairs play these games on ImmerseBoard. The results show that compared to standard video conferencing with a digital whiteboard, ImmerseBoard provides participants with a quantitatively better ability to estimate their remote partners' eye gaze direction, gesture direction, intention, and level of agreement. Moreover, the participants have a heightened sense of being together and a more enjoyable experience.

RELATED WORK

ImmerseBoard draws from several fields, including computer supported cooperative work (CSCW) and telepresence. In this section, we explain how ImmerseBoard is related to prior work in these fields.

Large Screen Collaboration

Large displays, and large touch screens in particular, have been used to support collaboration between many people in the same place. Streitz et al. proposed a collaboration workspace, including the DynaWall, which can be jointly operated by two people [23]. Khan et al. proposed a method for showing attention on a large display using a spotlight [11]. Birnholtz et al. evaluated the effectiveness of large screens in negotiation [2]. Other collaboration researchers aimed to enhance co-located collaboration using digital whiteboard systems that use the pen's buttons [5], and handheld-computers [21] . In contrast, we focus on remote collaboration, through a digital whiteboard. Specifically, we focus on generating a sense of presence between the remote participants as well as a seamless collaboration environment.

Remote Collaboration Systems

CSCW researchers aim to realize remote collaboration with an experience similar to local collaboration. Tang and Minneman proposed VideoWhiteboard, which is a remote collaboration system that shows the remote participant's shadow [26]. Apperley et al. also developed a collaboration system that shows shadow information on a large display [1]; however, shadowed facial information does not preserve eye contact. Ishii et al. introduced Clearboard, which shows the video of the remote participant on a shared workspace, as if the participants are looking at each other through a glass wall (on which they can write), approximating eye contact [7]. Clearboard flipped the remote video to fix the inverse writing problem. Ishii's research deeply affected the remote collaboration field [24]. Roussel designed THE WELL, which introduced the looking down display model [22]. These works all used tabletop computers to show shadow [25] or photographic hands [6] [4] for user's attention. In contrast, our paper extracts a 3D representation of the remote participant in order to reconstruct a more informative representation.

Immersive Human Reconstruction

Raskar et al. introduced immersive telepresence for remote collaboration in an office environment [20]. Various other works on immersive telepresence also involved reconstruction of human images in 2D/3D environments, including 3D human images from stereo or depth cameras [17] [13]. Zhang et al. made realistic human 3D images in real time using a hybrid camera system, consisting of a depth camera, IR cameras, color cameras, and IR laser projectors [29]. Morikawa et al. proposed Hyper Mirror, which mixed images from two places using background subtraction [16]. Several researchers displayed reconstructed humans using tetrahedral displays [8], omni-projection [12], and face-shaped displays [14]. In contrast, ImmerseBoard reconstructs the remote participant as a life-sized human body on a whiteboard in real-time using an RGBD camera for immersive collaboration.

Immersive Telepresence with a Whiteboard

Some prior work in immersive telepresence employs whiteboard collaboration [27]. Kunz et al., in Collaboard, extracted the remote participant from video and used background subtraction for showing attention [18]. Uchihashi et al. proposed a system for mixing remote locations using stereo cameras that can show the touch position of the remote person [28]. Junuzovic et al. created a shared work space on any surface using a camera-projection system [9]. However their IllumiShare system loses eye contact and face-to-face communication because the camera position is behind the user. For eye contact through video, the camera and the display should be at or close to the same position. In our work, we use 3D capture in order to solve the problem of displaying the remote person from the correct point of view, which solves the eye contact problem if the visual quality is sufficiently high.

Zillner et al. proposed 3D-Board, which can capture and transmit a user's whole body for remote whiteboard collaboration using multi-kinects [30]. Their solution is similar to the Mirror condition in this paper. However, we provide two additional novel visualizations - Hybrid and Tilt, which are valuable alternatives to Mirror. Our study finds that it is important to provide different conditions for participants to choose from, as their preferences are diverse and task dependent. In addition, there are several differences between our Mirror condition and 3D-Board: (a) 3D-Board is asymmetric. The operator can see the instructor's image, but the operator's image is not sent to the instructor. In contrast, ImmerseBoard provides a symmetric collaboration experience where participants can see each other. (b) 3D-board needs a Kinect away from the board to track the operator's head for motion parallax, while ImmerseBoard uses the Kinect on the side of the display to perform head tracking. (c) 3D-Board uses two Kinects to reconstruct the remote user with a better image quality than the Mirror condition in the ImmerseBoard, which we have left to future work.

Remaining Challenges

There are two major issues in the existing works. The first issue is the tradeoff between the form factor of the system and the level of immersion that it can provide. It is challenging to provide an immersive telepresence experience for whiteboard collaboration in a form factor simple enough for practical installation in homes and offices. The second issue is the difficulty of implementing the whiteboard metaphor (collaborating side by side on a whiteboard), and furthermore providing the glass wall metaphor for users to choose from within the same system. Therefore, we design and implement ImmerseBoard to address these two challenges.

TWO GUIDING METAPHORS

ImmerseBoard aims to connect remote collaborators as if they were co-located. Two metaphors of physical collaboration guide the design of ImmerseBoard. The first is the metaphor of side-by-side writing on a physical whiteboard, as shown in Figure 2(A). Each participant views the whiteboard and the other participant from his or her personal view, seeing the whiteboard in perspective and seeing the other participant from the side, in front of the whiteboard. The second is the metaphor of side-by-side writing on a physical mirror, as shown in Figure 2(B). Each participant sees the image of the other participant reflected in the mirror. The participants write on the mirror. In each metaphor, the participants are able to convey eye contact, eye gaze direction, pointing, hand gestures, body proximity, and other aspects of body language in relation to the other participant as well as to the writing. In addition, there is shared space in front of the writing surface for physical interaction and manipulation. The whiteboard metaphor is discussed in the Clearboard paper, but is considered hard to implement without using head mounted display. The mirror metaphor is similar to the glass wall metaphor in ClearBoard. In this paper, we implement both metaphors with a much simpler setup, i.e., just setting a Kinect on the side of large touch display (see next section), allowing users to choose their prefered metaphor.

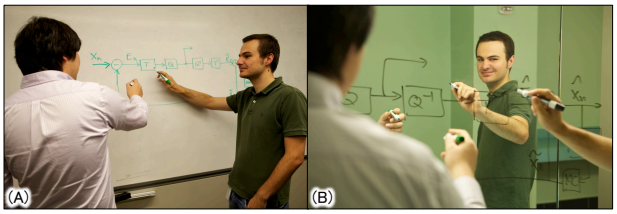

Figure 2. Metaphors: Side-by-side writing (A) on a whiteboard, (B) on a mirror.

SYSTEM

To emulate the above metaphors, we built ImmerseBoard around a large touch screen and a color plus depth (RGBD) camera, as shown in Figure 1(A). In our prototypes, the touch screen is a 55-inch Microsoft Perceptive Pixel (PPI) display, and the camera is a Microsoft Kinect camera. The PPI board is a multi-touch screen that can be used with either pens or fingers. We built two prototypes, called *Left* and *Right*, respectively configured with the PPI board to the left and right of the Kinect camera. (See Figure 3.) In this setup, users can move freely in the capture range of the Kinect camera (0.4-4.5 meters, 70° FOV), which allows them to roam up to 2 meters away from the board at its center. The remote user is rendered on the display close to the Kinect, so that the local user naturally stays within the capture range of and faces the Kinect in order to look at the image of the remote user and to write on the board. In the event that the local user faces the board directly, there may be some minor but not critical occlusions. One hand may be occluded by the torso when the hand is not active, but will be observable when the hand is writing or pointing.

The ImmerseBoards transmit to each other stroke data (position and color), color video data, depth video data, and skeleton data. The color and depth data allows us to extract an image and 3D point cloud of the participant without the background, while the skeleton data allows us to track the positions of the limbs of the participant. The depth data and skeleton data are expressed in the coordinate system of the capturing camera. In order to understand the pose of the participant in relation to the board, we transform the data from the camera's coordinate system into the board's coordinate system. This requires prior calibration of the pose of the camera with respect to the board.

We implemented a simple calibration system, which allows a user to tap four points in the corners of the PPI. When the user

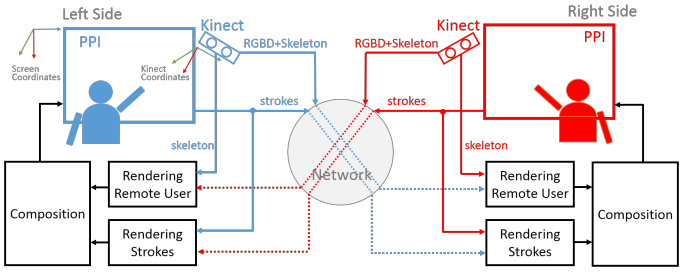

Figure 3. Left and right ImmerseBoard prototypes.

taps a point, the system records his 3D hand position from the skeleton information. From these four 3D positions, the system calculates a transformation matrix relating the coordinate systems of the camera and the board.

Once the data are transformed into the board's coordinate system, it can be processed and rendered with different visualizations (to be described in the next section). We use C++ and OpenGL for 2D/3D video processing and rendering, and use TCP for data communication.

IMMERSEBOARD CONDITIONS

ImmerseBoard supports several visualizations, or *conditions*. The first condition emulates the metaphor of participants writing shoulder-to-shoulder on a physical whiteboard. The second condition emulates the metaphor of the participants writing shoulder-to-shoulder on a mirror. Both of these conditions use 3D capture and rendering of the remote participants. The third condition is a hybrid between a standard 2D video conference and a 3D writing experience. A fourth condition is simply a standard 2D video conference with standard digital whiteboard. We now explain (in reverse order) the conditions and their implementations.

Video Condition

We begin with a standard video condition, in which the left or right side of the display is reserved for standard 2D video, leaving the bulk of the screen as a shared writing surface. The video is captured by the color camera in the Kinect, and displayed on the same side of the PPI as the camera, so that the eye gaze discrepancy is about 15 degrees. The display is large enough to show the upper body of the remote participant, life-sized. The video is processed so that the background is removed and the participant is framed properly regardless of where he is standing.

Hybrid Condition

The Hybrid condition is a hybrid of the above Video condition and a 3D experience. In the Hybrid condition, the remote participant's hand is able to reach out of the video window to gesture, point, or touch the board when writing, as shown in Figure 1(B). From the remote participant's hand position, the local participant is often able to understand the remote participant's intention as well as his attention.

ImmerseBoard implements the Hybrid condition using 3D depth and skeleton information from Kinect to guide 2D color video processing, as shown in Figure 4. The Kinect determines foreground (person) and background pixels. Each foreground pixel has a 3D coordinate. ImmerseBoard uses these 3D coordinates to segment pixels into body parts according to the pixels' 3D proximity to bones in the skeleton. The foreground pixels are framed within the video window of the display such that the upper body pixels are displayed. (This is the same as in the Video condition.) When the reaching arm is close to the PPI board, the system redraw arm and hand pixels by (a) moving the hand pixels to the appropriate location (orthogonal projection of the hand on the PPI board), and (b) stretching the image of the arm to seamlessly connect the upper body to the hand using texture mapping and deformation.

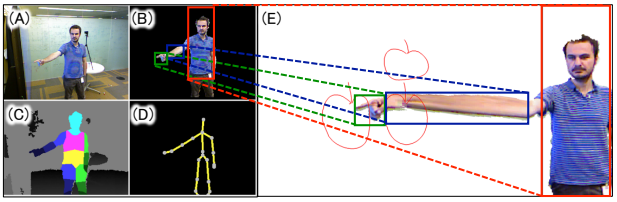

Figure 4. Video processing in Hybrid condition: (A) Source RGB image, (B) Extracted human image, (C) Segmentation, (D) Skeleton, (E) Result.

The hand and upper body images are not stretched. Aside from the stretched arm, the foreground image is identical to that coming from the color camera. Thus image quality and eye gaze discrepancy are the same as in the Video condition.

Mirror Condition

The Mirror condition, shown in Figure 1(C), is an emulation of the mirror metaphor. The remote participant's full upper body is seen life-sized, conveying body posture, body proximity, gesture direction, pointing direction, and eye gaze direction, in relation both to the board and to the local participant. Both participants are able to write on the entire surface, and see each other in any part of the surface, as if it were a large mirror.

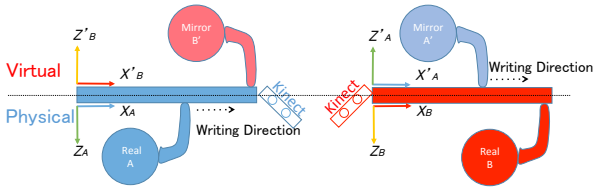

Figure 5. Mirror Condition: The system flips the z-axis in both sides

ImmerseBoard implements the Mirror condition by transforming the 3D colored point cloud from the Kinect coordinate system to the PPI coordinate system, and then flipping the z-axis (z to $-z$). The remote participant's point cloud is rendered using a 3D polygonal mesh. The viewpoint from which the remote participant is rendered onto the display can either be fixed at a default position, or for maximum accuracy, can track the head of the observer.

When head tracking is used at both sides, the relative geometry between the participants is precise, and eye contact is possible if the video quality is sufficiently high. Moreover, head tracking allows either participant to move to look around either the figures on the board or around the remote participant, as shown in Figure 6. However, the side of the remote participant not seen by his Kinect camera cannot be rendered, leading to a significant loss of perceived visual quality. Adding a second Kinect camera on the other side of the PPI board would solve the problem.

Tilt Board Condition

The Tilt board condition, shown in Figure 1(D), is an emulation of the metaphor of side-by-side writing on a physical whiteboard. As in the Mirror condition, the remote participant's full upper body is seen life-sized, conveying body posture, body proximity, gesture direction, pointing direction,

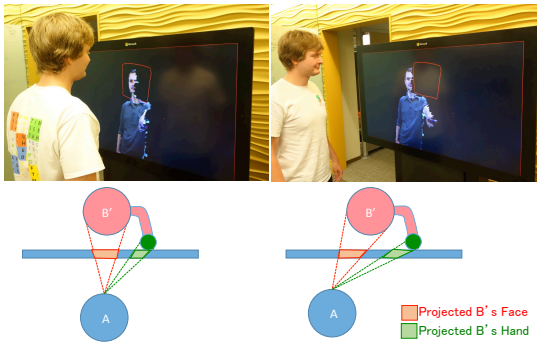

Figure 6. Mirror Condition with Head Tracking: The system can change perspective based on the user's head position.

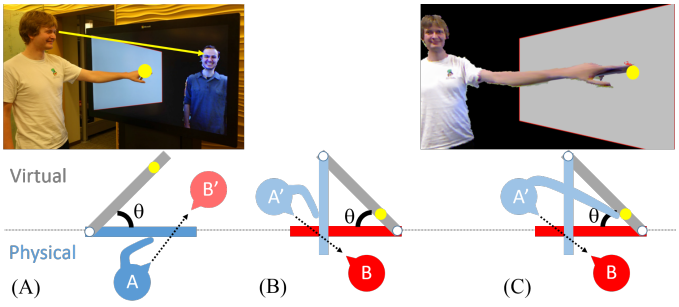

Figure 7. Tilt Board Condition: (A) The user touches the projection of the tilted board and looks at the remote person's face on the physical display. (B) The remote user's touch position would be incorrect if the system directly reconstructs the physical environment using the virtual board as the reference. (C) The system extends the remote participant's arm to correct the touch point.

and eye gaze direction, in relation both to the board and to the local participant. However, to fit the remote participant's image on the display, the image of the rectangular drawing surface is tilted back by 45 degrees (which is adjustable) and rendered in perspective. That is, the drawing surface is now virtual. Participants are able to write on the virtual drawing surface, by writing onto its projection on the physical PPI surface. At the same time, they can see each other as if they were side by side.

If writing onto the projection of a tilted virtual surface becomes awkward, optionally the tilted virtual surface can be rectified so that it coincides with the physical surface. When the board is rectified, the remote participant is no longer visible. Thus, typically, a user will use the tilted board to watch the remote participant present, and will use the rectified board to write detailed sketches. The tilting and rectification are visualizations for the benefit of the local user only, and can be done independently on either side.

To reduce the gaze divergence between the participants, the remote participant's image should be placed as close as possible to the Kinect camera. Thus, the direction of the tilt is different for the left and right boards, as shown in Figures 7 (A) and (C), respectively. For the left board, the Kinect camera is located on the right, and the virtual board is tilted to the left (Figure 7A). For the right board, the Kinect camera is located on the left, and the virtual board is tilted to the right (Figure 7C). As byproduct, this increases the overlap of the remote participant seen by the local participant and captured by the remote Kinect camera, resulting higher image quality, compared to the Mirror condition.

However, when the remote participant writes on a tilted board, he is actually writing on the image of the tilted virtual surface projected onto the physical surface of the PPI. Therefore, if the system directly reconstructs the physical environment (i.e., rotating the remote participant such that the virtual boards from both sides align) and changes only the viewpoint, the remote participant has correct eye gaze direction but points at the wrong place as shown in Figure 7B. Figure 7C shows that the correct touch point can be realized by extending the remote participant's arm to reach the correct position in the virtual environment.

To extend the remote participant's arm, the system calculates an appropriate hand position in the virtual environment. For example, if the participant is touching the physical board, this corresponds to a position on the virtual board (Figure 8 (A)). The hand is moved to this position in the virtual environment. However, if only the hand is moved to this position, it would be disconnected from the body (Figure 8 (B)). Thus, the system uses a coefficient α to interpolate the positions for points on the hand ($\alpha = 1.0$), arm ($0.0 < \alpha < 1.0$) and shoulder ($\alpha = 0.0$). The system also uses a coefficient β, based on the hand skeleton position in PPI coordinate system, to perform the interpolation only near the board. The system has two thresholds: $min(= 5cm)$ and $max(= 20cm)$. If the participant's hand is closer than min, β is 1.0. If it is further than max, β is 0.0. Otherwise, β is determined linearly ($0 < \beta < 1.0$). The system transforms each point on the hand, arm, or shoulder to a point $P_t = P_h(1-\alpha\beta)+P_p(\alpha\beta)$, where P_h is the original point and P_p is the projected point.

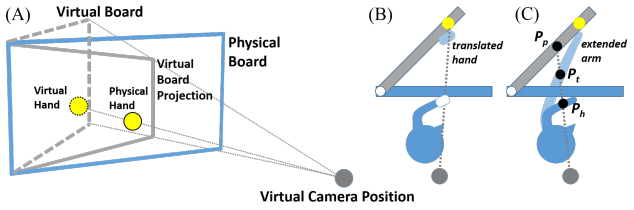

Figure 8. Tilt Board Geometry. (A) Projection of physical hand position on virtual board, (B) Hand translation towards virtual board, (C) Arm extension that preserves the proper hand-board relationship and arm-torso connection.

The major limitation of the Tilt board is the shape imprecision due to the perspective. It also causes fewer pixels to use on the side close to the remote user's image. Our system provides a remedy by allowing a user to rectify the board if needed.

USER STUDY

We performed a user study to compare the three conditions (video, hybrid, mirror, and tilt board), using both objective and subjective measures (the latter based on user feedback) to analyze key elements of the immersive experience such as gesture direction, intention, and eye gaze direction.

Participants and Studies

We recruited 32 subjects (5 female and 27 male) between the ages of 19 and 66 (mean 36). All subjects were right-handed information workers, with normal vision, hearing and movement ability. They all had video conferencing experience as part of their work.

The participants were partitioned into two disjoint studies (1 and 2) in order to answer three questions: (a) Is reference, e.g., pointing, useful for remote collaboration? (b) Does 3D rendering provide a better experience than 2D? (c) Which 3D condition (Mirror or Tilt) is more effective? Study 1 was formed to answer the first two questions and Study 2 for the third. This partition reduces the number of subjects needed to counterbalance all conditions tested in the studies.

Study 1 had 12 subjects, working in 6 sessions. Each session had a pair of subjects to act as remote collaborators (or partners). Study 1 compared 2D visualizations (Video, Hybrid) and a 3D visualization (Mirror). The Video and Hybrid conditions have good image quality as they are generated from the RGB source, but do not have exact eye contact nor do they preserve the positional relationship between the user and the board. The Mirror condition preserves eye direction and relationship to the board, but its image quality is relatively poor, due to rendering the remote participant from a viewpoint much different from that of the camera. We distributed the six sessions evenly over the six possible sequences of three conditions.

Study 2 had 20 subjects, working in 10 sessions. The study compared Mirror and Tilt conditions as well as a variation of the Mirror with headtracking and a variation of the Tilt with optional board rotation. Half the sessions evaluated first Mirror, then Tilt. The other half evaluated first Tilt, then Mirror.

Setting

The study session took place in a room with two ImmerseBoards (one left prototype and one right prototype). The two ImmerseBoards were separated by a curtain; thus the two subjects could see each other only through the ImmerseBoard. Subjects could write on the board using either fingers or stylus. They could talk to each other directly. We did not capture and transmit audio, since we focused on visual experience.

Procedure

At the beginning of each session, the two subjects filled out background questionnaires on their prior experience with video conferencing. Then they performed one subjective task (teaching) and two objective tasks (gaze estimation, symbol matching) to evaluate each condition. The three tasks will be discussed in detail in the next section. Each condition was introduced at the beginning in terms of the appropriate metaphor in Figure 2 (whiteboard or mirror). We did not observe any difficulties understanding the metaphors. For Study 2 , we also evaluated a variation of each condition (i.e. the Mirror with head-tracking and the Tilt board with optional rotation). Participants experienced the variation immediately after its base condition and were asked to perform only the teaching task in the interest of time. After each condition or variation, the subjects filled out a brief questionnaire on the condition or its variation. At the end of each session, the subjects filled out an overall questionnaire to compare the conditions. We also debriefed each subject with an interview.

Figure 9. Three Games for User Study. (A) Teaching game in Video condition. (B) Gaze estimation game in Mirror condition. The left participant guesses where the right participant is looking. (C) Symbol matching game in Hybrid condition. The right participant looks at where the left participant is about to touch.

Task Design

We designed the three tasks to be realistic, to be fun, and to reveal the strengths and weaknesses of the different conditions on aspects important to real-world collaboration. The first task is a subjective but practical task: teaching. Teaching is a important and broadly representative use case for remote collaboration systems [3]. The remaining tasks are games with measurable objectives. The subjects are instructed to play each game to maximize (or minimize) its objective. We used the game outcomes to evaluate aspects of each condition.

Subjective Task — Teaching

As shown in Figure 9(A), one subject plays the role of Teacher, and the other Student. The subjects are free to decide among themselves who will be Teacher, and what the Teacher will teach. We suggested teaching the rules of a card game, board game, or sports game, but many other topics came up during the user study. The subjects had about 3-5 minutes to teach and learn, just enough time to get some experience and understanding of the condition. In addition, the teaching task evaluates how well the condition is able to convey social cues about the level of agreement and understanding through questionnaire and interview.

Objective Game 1 — Gaze Estimation

The first game (Figure 9B) evaluates the accuracy with which a participant can estimate the eye gaze direction of the remote participant in each condition. It is well known that eye contact

and eye gaze direction are important elements of communication [10] [14]. This is an asymmetric game with a leader and a follower, so the game is played twice in each condition, to give each subject a chance to play both roles. The players are shown an eight by eight grid of cells on their shared surface. However, on the leader's side, one of the cells is colored red, at random. The red cell is not visible on the follower's side. The leader is prompted to look at the red cell in a natural way. The follower observes his partner via the visual condition in effect, and tries to guess which cell his partner is looking at. The follower clicks on the estimated cell, and then is shown the correct answer. Before the follower clicks a button to move on to the next trial, the system blanks the follower's visual condition and shows the leader a new prompt. This gives the leader time to move to the new prompt without the follower seeing the leader's direction of movement. There are 16 trials per side. We record the estimation accuracy.

Objective Game 2 — Symbol Matching

The second game (Figure 9C) evaluates the ability of a participant to follow his partner's gestures as cues of attention and intention. Again, this is an asymmetric game with a leader and a follower, so the game is played twice in each condition, to give each subject a chance to play both roles. The players are shown on their shared surface ten pairs of symbols, randomly permuted on a four by five grid. The symbols come in five shapes (circle, square, diamond, up and down triangles) and two colors (red and blue). The leader taps a colored shape to make it disappear. The follower is instructed to tap the corresponding colored shape, as quickly as possible, to make it disappear. If he taps the wrong shape, nothing happens. When all shapes are gone, the game ends. We recorded the follower's response time.

STUDY RESULTS

In this section, we will demonstrate and discuss the user study results. We will show the quantitative evaluation for the two games (gaze estimation, symbol matching) as well as the results from questionnaire and interview.

Result of Gaze Estimation Game

Figure 10 shows the average shift, or bias, from the cell that the follower clicked to the cell that the leader was looking at over four conditions in two studies. The bias is calculated per block that includes 4 cells over all subjects, since each block had only one cell selected in each game.

Figure 11 shows the mean and standard deviation of the horizontal and vertical errors for all four conditions. The horizontal (or vertical) error is defined as the absolute distance from the cell which the follower clicked to the cell that the leader was looking at, along the horizontal (or vertical) direction, in units of the number of cells. A Repeated Measures ANOVA revealed a significant main effect in Study 1 on both horizontal ($F_{2,22} = 9.12, p = .001$) and vertical error ($F_{2,22} = 14.75, p < .001$). Post-hoc pairwise comparison (with bonferonni corrections) revealed that (a) Video had significantly more horizontal error than Hybrid ($p = .046$) and Mirror ($p = .003$), and (b) Mirror had significantly more vertical error than Video ($p = .001$) and Hybrid ($p = .011$).

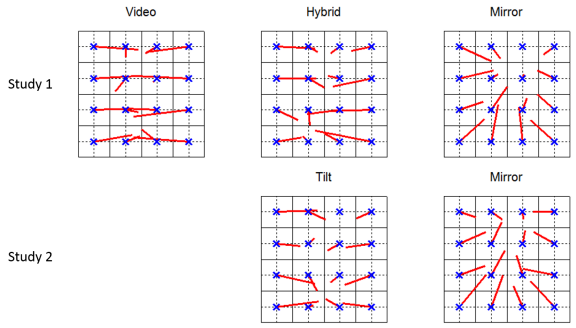

Figure 10. Gaze Estimation: Bias over different locations. The blue cross is the leader's true gaze direction and the red line indicates the follower's gaze estimation bias.

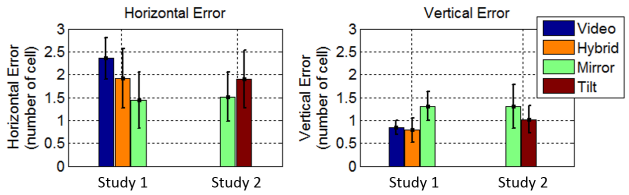

Figure 11. Horizontal and Vertical Error of gaze estimation game.

For Study 2 , the paired Student's t-test revealed that the Mirror was significantly better than Tilt on horizontal error ($p = .027$) and significantly worse on vertical error ($p = .033$).

Discussion of Gaze Estimation Game

In Study 1 (top row of Figure 10), the Video and Hybrid conditions have large horizontal bias and small vertical bias. The Mirror condition is opposite - small horizontal bias and large vertical bias. For the Video and Hybrid conditions, the horizontal and vertical biases are not equal, because the person-board relationship is preserved vertically, but not horizontally. Hence, it is difficult for the follower to estimate the horizontal gaze direction. Some subjects talked about this in the interview - *"It was pretty easy to tell up and down but was harder to pick up the column for the Video and Hybrid conditions."* For the Mirror condition, the person-board relationship is preserved both vertically and horizontally. As expected, the bias in each direction has similar magnitude, and the horizontal bias is less than in the Video and Hybrid conditions. However, it is surprising that the Mirror condition has significantly more vertical bias than the Video and Hybrid conditions. This is likely due to the poor video quality in the Mirror condition, especially around the eye area, such that the eye ball direction cannot be seen as well. Hence the participants rely more on the head direction to estimate the gaze direction. Thus, if the highlighted cell is directly in front of the partner (eye balls at neutral position), the bias is small. The bias increases when the highlight cell moves toward the boundaries, since the eye ball movement (other than head movement) contributes more to the eye gaze [19]. This can be confirmed in Figure 10. The blocks on the top-middle have small bias because that is the average head position of the remote partner. The blocks on the left/right/down have larger bias and the bias increases as the cell moves further.

Unexpectedly, the Hybrid condition is better than the Video condition on horizontal gaze bias in spite of having the same video quality. We conjecture that the leader may have learned implicitly from the reference (i.e., follower's pointing) to convey better his gaze direction to the follower.

In Study 2 (bottom row of Figure 10), the Mirror condition has a pattern similar to that of Study 1 . Like the Mirror condition, the Tilt board condition preserves both the horizontal and vertical person-board relationship, but from a different perspective. Since the Tilt board provides a side view, the camera position for rendering is relatively close to the Kinect camera position for capturing. Thus the Tilt board has a better video quality than the Mirror condition. In consequence, the Tilt Board has less vertical bias than the Mirror condition. However, the horizontal bias for the Tilt Board is more than for the Mirror. This is because it is more difficult to estimate horizontal eye gaze direction from a side view than from a frontal view, which the Mirror condition has.

Result of Symbol Matching Game

Figure 12 shows the follower's average response time in the symbol matching game. The response time is defined as the difference between the time the leader selects a symbol and the time the follower clicks the correctly matched symbol.

A Repeated Measures ANOVA revealed a significant main effect on the response time in Study 1 , $F_{2,22} = 7.99, p = .002$. Post-hoc pairwise comparison (with bonferonni corrections) revealed that Mirror was significantly faster than Video ($p = .005$). The Hybrid is between the Video and the Mirror. For Study 2 , the Mirror and the Tilt are very similar.

Discussion of Symbol Matching Game

In Study 1 , the Mirror condition has a better capability than the Video condition to transmit leader's gesture direction, intention and attention. In Study 2 , the Tilt board condition has a nearly equal capability with the Mirror condition. Thus, the Mirror and Tilt board conditions significantly (and Hybrid slightly) outperform the Video condition because rendering the leader's arm in the former conditions helps the follower anticipate the symbol that the leader is about to touch. We observed that many users understood the challenges of the Video condition and prepared themselves by concentration. Some other users had to scan through all shapes to identify the missing symbol. The subjects also discussed this in the interview - *"In the Video condition, you do not know which symbol is about to disappear, but when you can see where the hand was in Hybrid and Mirror conditions, you can anticipate which symbol will be selected."* Hence, touch and pointing positions are quite important for the immersive telepresence visualization.

Questionnaire

The participants are asked to rank the conditions with respect to the questions in Figure 13 and Table 1. In Study 1 , the subjects ranked the three conditions (Video, Hybrid, Mirror) from the worst to the best. In the Study 2 , the subjects pick the better condition from either Mirror and Tilt Board.

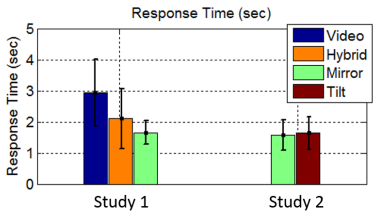

Figure 12. Response time of symbol matching game.

Table 1. Result of Friedman and pairwise Wilcoxon Signed Ranks Tests on the participant's ranking in Study 1 . "*" indicates the significance.

Questions	Friedman Tests		Wilcoxon: Video/Hybrid		Wilcoxon: Video/Mirror	
	$\chi_{12,2}$	p	Z	p	Z	p
being together	10.17	.006*	−3.28	.001*	−2.43	.015*
enjoy experience	8.00	.018*	−2.81	.005*	−2.077	.038
video was useful	12.50	.002*	−3.22	.001*	−2.67	.008*
video quality	6.50	.039*	−1.90	.058	−.58	.564
eye contact	3.50	.174	−1.73	.083	−1.57	.117
ideas conveyed	15.17	.001*	−3.18	.001*	−2.91	.004*
where to look	3.50	.174	−1.73	.083	−1.57	.117
where to touch	18.00	.001*	−3.15	.002*	−3.15	.002*
read agreement	7.17	.028*	−2.80	.005*	−1.71	.087

Figure 13 shows the ranking results. In Study 1 , most subjects prefer either Hybrid or Mirror for all questions except the question about video quality. Table 1 shows the statistic test results. Significant main effects were revealed for seven questions (except the *"eye contact"* and *"see where the partner was looking"*). Within these seven questions, pairwise comparison revealed that both Hybrid and Mirror conditions were ranked significantly better than the Video for four questions (*"being together," "video was useful," "ideas were conveyed," "see where the partner was about to touch"*) and the Hybrid condition was also ranked significantly better than the Video for another two questions (*"enjoy experience,"*, *"read partner's agreement level"*). The main effect ($p < .05$ for Friedman test) and pairwise significance ($p < .0167$ for Wilcoxon signed ranks test) is marked by "*". There is no significance between the Hybrid and Mirror for any question.

In Study 2 , we observe the diversity of the subjects' preferences over all questions except *"video quality"*, on which the Tilt outperforms the Mirror significantly ($\chi_{20,1} = 9.80, p = .002$). Other than that question, Mirror and Tilt are very close.

Subjects in Study 2 also rated if the variations, i.e., head tracking for the Mirror and optional board rotation for the Tilt board, improved the condition on a 7 point scale (disagree=1, agree=7). Subjects mildly agreed on these two variations (head tracking 4.6/7 and board rotation 4.95/7).

Feedback

ImmerseBoard, including Hybrid, Mirror and Tilt Board conditions, received very positive feedback from the subjects. All subjects were excited about using ImmerseBoard because *"this is very cool, new experience that could be very useful in my profession," "this is impressive for remote collaboration," "we were naturally trying to help each other."*

The subjects also explained why they enjoyed ImmerseBoard, such as *"It was a good experience to see the video of my partner, to see his reaction, where he is looking and what he is doing," "I was able to predict where my partner is about to touch, and this would be great especially for co-workers."*. They also like the simple setup - *"The setup is so simple that I can easily fit this to my office."*

We also received negative feedback mostly on the video quality, particularly for the Mirror condition, like *"Video quality of partner was not that great, particularly face and eyes," "The video quality was not great. If it were better the experience would be much improved. It was difficult to see the eyes with current video quality."*

For the Hybrid condition, the participants liked the arm extension as the reference as they said, *"The hybrid condition is my favorite because I can see a clear cut of her and her arm, and I know where she is about to touch," "The video was decent, and you can see where your partner was writing and the part he was pointing."* The negative feedback included *"It is hard to tell where my partner was looking at horizontally"* and *"The extended arm sometimes made me distracted especially when the hand moves fast."*

For the Mirror condition, the subjects enjoyed the fluid experience, especially for rapid interactions such as brainstorming and collaborations - *"I definitely like the Mirror condition for the collaboration purpose, I was standing right with the partner and interacting closely," "It feels like both of us were physically there," "It was easy to see where my partner was looking and pointing and it was a little more precise."* Not surprisingly, the subjects did not like the video quality - *"I did not like the video quality in the Mirror, I really want to see where my partner's eyes were going."* Also, some participants were concerned about the overlapping of the remote user and the writing, *"My partner's body was blocked by what we drew on the board."* Actually, the overlapping causes difficulty seeing the writing if participant's outer clothes and the writing are same color.

For the Tilt condition, subjects felt that it was natural and realistic especially for the teaching scenarios - *"I like the Tilt board so much because it is more realistic, we normally work on one side of the board," "It was more natural especially for teaching or presentation."* However, some subjects were concerned that the perspective introduces imprecision as they commented - *"Compared to the Mirror, the Tilt board was more imprecise to see where my partner was pointing and where she was going to select," "Because of the perspective, the shapes look skewed as it should not be, and it made more difficulty to mentally register which shapes they were."*

Finally, the subjects discussed their preferences based on applications. In general, the subjects preferred the Hybrid and Tilt board conditions for teaching and presentation, while preferring the Mirror for close interaction and collaborations. (e.g. brainstorming).

DISCUSSION

We now summarize what we learned from the study. First, participants quickly got used to the ImmerseBoard and preferred the three immersive conditions (Hybrid, Mirror, Tilt) over the Video condition since the immersive conditions provided them better ability to estimate their remote partners' eye gaze direction, gesture direction, and intention, making the remote collaboration more natural. Second, the participants enjoyed the 3D immersion (Mirror and Tilt) as it is more natural and realistic, and would show even more preference if the video quality were improved. Third, it is important to provide different conditions for participants to choose from, as they have diverse preferences and their preferences are also task-dependent. Finally, the participants felt the setup or form factor is so simple that they could envision using it in their office or home.

Figure 13. Questionnaire: Ranking Results.

CONCLUSION AND FUTURE WORK

We introduced ImmerseBoard, which combines a large touch display with an RGBD camera to give remote collaborators an immersive experience as if they were writing side by side on a physical whiteboard or mirror. In addition to designing and implementing the system, we conducted a detailed user study involving 32 subjects performing several tasks. The tasks included a teaching task, as well as tasks to assess awareness of eye gaze direction and gestural attention/intention, all reflecting important aspects of real-world collaboration. Subjects were quantitatively better at estimating their remote partners' gesture direction and intention, and level of agreement, which translated qualitatively into a heightened sense of being together and a more enjoyable experience. However, the results also revealed limitations due to the 3D image quality from Kinect. In the future, we plan to improve the ImmerseBoard in several directions. First and foremost, we will improve the image quality by using high resolution sensors on both sides of the PPI screen. In addition, we will apply human body models for 3D reconstruction. We also plan to investigate the integration of three conditions (Hybrid, Mirror and Tilt) to provide the best experience for users based on applications and user's preference.

ACKNOWLEDGMENTS

We thank Kori Inkpen and Sasa Junuzovic for their support.

REFERENCES

1. Apperley, M., McLeod, L., Masoodian, M., Paine, L., Phillips, M., Rogers, B., and Thomson, K. Use of video shadow for small group interaction awareness on a large interactive display surface. AUIC '03 (2003), 81–90.
2. Birnholtz, J. P., Grossman, T., Mak, C., and Balakrishnan, R. An exploratory study of input configuration and group process in a negotiation task using a large display. CHI '07, 91–100.
3. Diamant, E. I., Fussell, S. R., and Lo, F.-L. Collaborating across cultural and technological boundaries: team culture and information use in a map navigation task. IWIC '09, 175–184.
4. Doucette, A., Gutwin, C., Mandryk, R. L., Nacenta, M., and Sharma, S. Sometimes when we touch: how arm embodiments change reaching and collaboration on digital tables. CSCW '13, 193–202.
5. Elrod, S., Bruce, R., Gold, R., Goldberg, D., Halasz, F., Janssen, W., Lee, D., McCall, K., Pedersen, E., Pier, K., Tang, J., and Welch, B. Liveboard: a large interactive display supporting group meetings, presentations, and remote collaboration. CHI '92, 599–607.
6. Genest, A. M., Gutwin, C., Tang, A., Kalyn, M., and Ivkovic, Z. Kinectarms: a toolkit for capturing and displaying arm embodiments in distributed tabletop groupware. CSCW '13, 157–166.
7. Ishii, H., and Kobayashi, M. Clearboard: a seamless medium for shared drawing and conversation with eye contact. CHI '92, 525–532.
8. Jouppi, N. P., Iyer, S., Thomas, S., and Slayden, A. Bireality: mutually-immersive telepresence. MULTIMEDIA '04, 860–867.
9. Junuzovic, S., Inkpen, K., Blank, T., and Gupta, A. Illumishare: sharing any surface. CHI '12, 1919–1928.
10. Kendon, A. Some functions of gaze-direction in social interaction. *Acta psychologica 26* (1967), 22–63.
11. Khan, A., Matejka, J., Fitzmaurice, G., and Kurtenbach, G. Spotlight: directing users' attention on large displays. CHI '05, 791–798.
12. Kim, K., Bolton, J., Girouard, A., Cooperstock, J., and Vertegaal, R. Telehuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. CHI '12, 2531–2540.
13. Liu, S., Chou, P. A., Zhang, C., Zhang, Z., and Chen, C. W. Virtual view reconstruction using temporal information. *IEEE ICME 2012*, 115–120.
14. Misawa, K., Ishiguro, Y., and Rekimoto, J. Livemask: a telepresence surrogate system with a face-shaped screen for supporting nonverbal communication. AVI '12, 394–397.
15. Moezzi, S. Immersive telepresence. *MultiMedia, IEEE 4*, 1 (1997), 17–17.
16. Morikawa, O., and Maesako, T. Hypermirror: toward pleasant-to-use video mediated communication system. CSCW '98, 149–158.
17. Mulligan, J., Zabulis, X., Kelshikar, N., and Daniilidis, K. Stereo-based environment scanning for immersive telepresence. *Circuits and Systems for Video Technology, IEEE Transactions on 14*, 3 (2004), 304–320.
18. Nescher, T., and Kunz, A. An interactive whiteboard for immersive telecollaboration. *The Visual Computer 27*, 4 (2011), 311–320.
19. Proudlock FA, Shekhar H, G. I. Coordination of eye and head movements during reading. *Invest Ophthalmol Vis Sci. 44*, 7 (2003), 2991–2998.
20. Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., and Fuchs, H. The office of the future: a unified approach to image-based modeling and spatially immersive displays. SIGGRAPH '98, 179–188.
21. Rekimoto, J. A multiple device approach for supporting whiteboard-based interactions. CHI '98, 344–351.
22. Roussel, N. Experiences in the design of the well, a group communication device for teleconviviality. MULTIMEDIA '02, 146–152.
23. Streitz, N. A., Geissler, J., Holmer, T., Konomi, S., Müller-Tomfelde, C., Reischl, W., Rexroth, P., Seitz, P., and Steinmetz, R. i-land: an interactive landscape for creativity and innovation. CHI '99, 120–127.
24. Tan, K.-H., Robinson, I., Samadani, R., Lee, B., Gelb, D., Vorbau, A., Culbertson, B., and Apostolopoulos, J. Connectboard: A remote collaboration system that supports gaze-aware interaction and sharing. MMSP '09, 1–6.
25. Tang, A., Pahud, M., Inkpen, K., Benko, H., Tang, J. C., and Buxton, B. Three's company: understanding communication channels in three-way distributed collaboration. CSCW '10, 271–280.
26. Tang, J. C., and Minneman, S. Videowhiteboard: video shadows to support remote collaboration. CHI '91, 315–322.
27. Tseng, B. L., Shae, Z.-Y., Leung, W. H., and Chen, T. Immersive whiteboards in a networked collaborative environment. In *IEEE Multimedia and Expo* (2001).
28. Uchihashi, S., and Tanzawa, T. Mixing remote locations using shared screen as virtual stage. MM '11, 1265–1268.
29. Zhang, C., Cai, Q., Chou, P., Zhang, Z., and Martin-Brualla, R. Viewport: A distributed, immersive teleconferencing system with infrared dot pattern. *MultiMedia, IEEE 20*, 1 (2013), 17–27.
30. Zillner, J., Rhemann, C., Izadi, S., and Haller, M. 3d-board: A whole-body remote collaborative whiteboard. UIST '14, 471–479.