Ye Pan*
Oyewole Oyekoya
Anthony Steed
Department of Computer Science
University College London
Gower Street, London WC1E 6BT

A Surround Video Capture and Presentation System for Preservation of Eye-Gaze in Teleconferencing Applications

Abstract

We propose a new video conferencing system that uses an array of cameras to capture a remote user and then show the video of that person on a spherical display. This telepresence system has two key advantages: (i) it can capture a near-correct image for any potential observer viewing direction because the cameras surround the user horizontally; and (ii) with view-dependent graphical representation on the spherical display, it is possible to tell where the remote user is looking from any viewpoint, whereas flat displays are visible only from the front. As a result, the display can more faithfully represent the gaze of the remote user. We evaluate this system by measuring the ability of observers to accurately judge which targets the actor is gazing at in two experiments. Results from the first experiment demonstrate the effectiveness of the camera array and spherical display system, in that it allows observers at multiple observing positions to accurately tell at which targets the remote user is looking. The second experiment further compared a spherical display with a planar display and provided detailed reasons for the improvement of our system in conveying gaze. We found two linear models for predicting the distortion introduced by misalignment of capturing cameras and the observer's viewing angles in video conferencing systems. Those models might be able to enable a correction for this distortion in future display configurations.

1 Introduction

Traditional video conferencing setups often distort the appearance of human gaze due to screen and camera orientation disparities and other asymmetries. For example, even when looking directly at their collaboration, the video of users may appear as if their gaze is up, down, or to the sides. These distortions can lead to problems sharing socially useful information such as attention targets, conversational turn-taking indicators, etc. A variety of teleconferencing systems have been successfully built for reproducing eye-gaze, though the majority use a planar display (Nguyen & Canny, 2005). However, these planar displays are visible from the front only.

Presence, *Vol. 24, No. 1, Winter 2015,* 24–43
doi:10.1162/PRES_a_00213

*Correspondence to y.pan@cs.ucl.ac.uk; w.oyekoya@cs.ucl.ac.uk; a.steed@cs.ucl.ac.uk.

We propose the use of non-planar displays, in particular a spherical display to represent a remote user. We use a spherical display that is small enough to situate almost anywhere in a room and that is visible from all directions (Benko, Wilson, & Balakrishnan, 2008; Oyekoya, Steptoe, & Steed, 2012). Also, compared to volumetric displays (Yendo, Fujii, Tanimoto, & Panahpour Tehrani, 2010), robotics (Desai, Tsui, Yanco, & Uhlik, 2011) and animatronic shader lamp avatars (Lincoln et al., 2009), it is technically quite simple to build and similar systems can be constructed inexpensively. Furthermore, as cameras are now becoming very cheap, we propose the use of a camera array to capture the remote user, so that we can select an appropriate video of the user to show.

We compare the effectiveness of both spherical and flat displays by measuring the ability of observers to accurately judge which target a user is gazing at. Experiment 1, as a pilot study, demonstrated that the camera array plus sphere display can convey gaze relatively accurately (Pan & Steed, 2012). Experiment 2 compared observers' performance in different flat and spherical display conditions further, by modeling systematic biases and investigating the influence of seat and target positions.

In the following sections, we review related work and present the software and hardware components needed to implement our spherical display system. This is followed by two experimental evaluations of our system. Finally, we present discussions of the results, implications for future designs, conclusions and future work.

2 Related Work

2.1 Display System

2.1.1 Flat Display. Eye-gaze distortions in video conferencing are mainly caused by two factors: parallax shift effect and collapsed viewer effect (Norris, Schnädelbach, & Qiu, 2012; Nguyen & Canny, 2005). The parallax shift effect occurs due to the placement of the video camera, which tends to be perched on top of a monitor display in traditional videoconferencing systems. This causes the user's eye direction to be different from the video camera's capturing direction. Note that the parallax shift effect can occur both horizontally and vertically. The collapsed viewer effect is where all remote participants share the same virtual viewing position of the local scene. This happens in group-to-group video communication systems. For example, if a participant is looking directly at the capturing camera in a remote room, all the viewers in the local room will feel that the remote participant is looking at them. In this section, we review different types of flat displays that can reproduce eye-gaze in different two-way conversation, three-way or N-way conversation, and group-to-group conversation scenarios.

In a two-way conversation, where only two participants at different geographical locations join the video communication, there are various methods of producing a correct gaze direction (Bohannon, Herbert, Pelz, & Rantanen, 2013). Using a half-silvered mirror (Arai, Kuriki, & Sakai, 1992; Acker & Levitt, 1987), embedding a camera in the center of display (Adalgeirsson & Breazeal, 2010), or using a transparent display could allow a video camera to capture a participant's correct gaze direction without blocking the image on the screen. However, once participants are moving or not sitting in front of the display, the parallax shift effect will occur.

For three-way or N-way conversations, more than two participants at different places link up in the conversation. Apart from considering parallax shift effect, the structure of the three-way or N-way communication network is also an essential issue. Round-table and SVTE (shared virtual table environment) are basic schemes to build a three-way or N-way communication network (Schreer, Kauff, & Sikora, 2005). For the round-table scheme, N parties are virtually grouped around a round table while cameras are placed in all lines of sight. The Hydra system (Sellen, Buxton, & Arnott, 1992) placed several hydra units in front of a local user to present the video of remote users. For MAJIC (Okada, Maeda, Ichikawaa, & Matsushita, 1994), at each site of this system, a large semi-transparent curved screen was mounted behind a normal computer terminal. In the MONJUnoCHIE system (Aoki et al., 1999), a special semi-transparent display based on holographic optical elements was used. However, the overall transmission bit is increased with the square of the connected sites; for example, N × (N–1) cameras are needed for N participants. Alternatively, the SVTE scheme manages to decrease the overall transmission bit by integrating generic 3D representations of the conferees into a shared virtual environment (Atzpadin, Kauff, & Schreer, 2004; Mulligan, Isler, & Daniilidis, 2002).

Group-to-group conversation means that multiple users are collocated with an instance of the teleconferencing system. The two-way conversation systems that we discussed could be used for communication between two groups of users, if the viewing distance between the two users is sufficiently large. The Telepresence Wall (Buchner, 2006) is an example of a display used to support two groups at two sites. The GAZE-2 (Vertegaal, Weevers, Sohn, & Cheung, 2003) uses an eye-controlled camera direction to ensure parallax-free transmission of eye contact. GColl (Slovák, Troubil, & Holub, 2009) supported mutual-gaze as well as partial-gaze awareness for all participants with modest technical requirements; for example, notebooks with two USB cameras for each user. In Jones et al. (2009), a one-to-many 3D video teleconferencing system is introduced to achieve multi-view conversation. The remote user's face is scanned in 3D at 30 Hz and transmitted in real time to an auto-stereoscopic horizontal parallax 3D display, displaying it over more than 180° in a field of view observable to multiple views. MultiView (Nguyen & Canny, 2005) accomplishes reproduction of eye-gaze in group-to-group conversation by capturing unique and correct perspectives for each participant using one of many cameras and simultaneously projecting each of them onto a directional screen that controls who sees which image.

As we discussed, many telepresence systems have been built to improve different videoconferencing scenarios, though the majority use planar displays. However, these planar displays are visible only from the front, and lack the 360° view offered by spherical displays. Also, in order to correct eye-gaze distortion, flat displays either use the half mirror which will reduce the video quality and increase the display complexity, or embed the camera in the center of display which will block display image. On the other hand, spherical displays project video from the bottom of the display, thus avoiding this problem. Nevertheless, previous research using flat displays for three-way or N-way conversation and group-to-group conversation scenarios are important for future development of spherical displays.

2.1.2 Virtual Reality Systems. Systems such as VIRTUE (virtual team user environment) (Schreer et al., 2002), im.point (immersive meeting point) (Tanger, Kauff, & Schreer, 2005), Blue-C (Gross et al., 2003), and office of the future (Raskar et al., 1998) are effective ways to simulate face-to-face conversations by applying the concept of a shared environment (Hindmarsh, Fraser, Heath, Benford, & Greenhalgh, 2000). These virtual reality systems mainly focus on how well observers can distinguish between objects according to the remote user's gaze direction. TELEPORT system (Gibbs, Arapis, & Breiteneder, 1999) and NTII (National Tele-Immersion Initiative) (Sadagic et al., 2001) utilized the SVTE concept described earlier to preserve gaze direction in three-way or N-way conversations. However, these systems need sophisticated equipment, such as complex display mountings, special tracking devices, etc. Also, they are currently too expensive to put into commercial use.

2.1.3 Telepresence Robots. Mobile telepresence robots, such as MeBot V4 (Adalgeirsson & Breazeal, 2010), PRoP (Paulos & Canny, 1998), Anybots'QB, and the VGo (Tsui, Desai, Yanco, & Uhlik, 2011), allow a remote user to control the robot's movement around a space while the user converses with other users in that space. These devices tend to have a built-in flat screen to display a video stream of the remote user. Using these telepresence robots, remote co-workers can wander the hallways and engage in impromptu interactions, increasing opportunities for connection in the workplace (Lee & Takayama, 2011). Since mobility is the characteristic that differentiates mobile telepresence robots from video conferencing technologies, we could potentially integrate a spherical display into a robotic platform. Compared to mobile telepresence robots, the 360° visibility of the spherical display would be a distinct advantage.

Humanoid robotics focus more on better conveyance of a person's remote physical presence. Geminiod HI-1 (Sakamoto, Kanda, Ono, Ishiguro, & Hagita, 2007) was developed to closely resemble a specific human. Animatronic Shader Lamps Avatars (SLA) (Lincoln et al., 2009) use the technique where an image of an object is projected onto a screen whose shape physically matches the object. It uses cameras and projectors to capture and map the dynamic motion and appearance of a real person onto a humanoid animatronic model. These humanoid robots can potentially be used to represent specific visitors at a destination but they are limited in terms of their flexibility in representing other teleoperators. In comparison, the sphere display offers flexibility compared to humanoid robotics as it is not constrained to a single head size or shape.

2.1.4 Situated Display. There are different kinds of non-flat display surfaces (Ten Koppel, Bailly, Müller, & Walter, 2012), particularly, situated displays, such as spherical displays and tubular displays, which are small enough to situate almost anywhere in a room and which are visible from all directions. The BiReality system (Jouppi, Iyer, Thomas, & Slayden, 2004) consisted of a display cube at a user's location and a surrogate in a remote location. Both the remote participant and the user appeared life-size to each other. The display cube provided a complete 360° surround view of the remote location and the surrogate's head displayed a live video of the user's head from four sides. Horizontal gaze is best preserved for the user as seen by remote participants when the user is looking into the cameras in the corner of the display cube, and is sloppier when the user is looking at the center of a screen. SphereAvatar (Oyekoya et al., 2012) represents a remote user as an avatar on a spherical display which is able to accurately convey head-gaze. Our system extends the work of SphereAvatar (Oyekoya et al., 2012). We surround a camera array to reproduce real-time video of a remote participant instead of an avatar in order to improve reproduction fidelity and preserve eye-gaze. TeleHuman (Kim, Bolton, Girouard, Cooperstock, & Vertegaal, 2012) provides 360° motion parallax with stereoscopic life-sized 3D images of users, using a lightweight approach. Motion parallax is provided via perspective correction that adjusts views as users move around the display. Stereoscopy is provided through shutter glasses worn by the user. The system uses ten Microsoft Kinects for capturing 3D video models of the user in 360°. Telehuman is a reconstruction system, whereas we focus on spatial video transmission.

2.2 Arrangement of Capturing Cameras

The arrangement of capturing cameras in video conferencing can be divided into three categories: fixed camera, moving camera, and dense camera array. For the fixed camera approach, it is common to place the video camera close to the image of the other person's eye to avoid a parallax shift effect. However, it limits the user to a specific position. For a moving camera, the camera's position changes according to the direction given by the user's eye. One of the representative examples of a moving camera system is a telepresence robot, discussed earlier. The last type of arrangement is the dense camera array. A scene is captured by a set of cameras and optionally manipulated by view interpolation as discussed below. When the camera density is very high, view generation is done simply by selecting a camera image or by collecting pixels from the camera image, such as the NHK system (Arai, Okui, Yamashita, & Okano, 2006) or 1D integral image 3D display system (Hirayama, 2009). When the camera density is moderately high, view generation needs some processing such as FTV (Free viewpoint TV) (Tanimoto, 2006), Bird's eye view system (Sekitoh, 2001), Light field camera system (Wilburn, Smulski, Lee, & Horowitz, 2001), Surface light field camera system (Chen, Bouguet, Chu, & Grzeszczuk, 2002), EyeVision, 3D-TV (Matusik & Pfister, 2004), and Free-viewpoint video player (Eisemann et al., 2008). When the camera density is low, intermediate views can be generated by detecting a model in the scene, such as 3D room (Saito, Baba, Kimura, Vedula, & Kanade, 1999) and 3D Video (Matsuyama & Takai, 2002). In our system, the number of views is moderately high. We arrange cameras in a semicircle and thus the system is able to capture the views from multiple horizontal directions.

2.3 Gaze Evaluation

2.3.1 Object-Focused Gaze Awareness and Mutual Gaze.

Detecting the gaze direction of a person is important for human–computer interaction applications in video conferencing or shared collaborative work spaces (Leyrer, Linkenauger, Bülthoff, Kloos, & Mohler, 2011). Eye-gaze includes object-focused gaze awareness and mutual-gaze. Object-focused gaze awareness means that if the partner is gazing at an object in the shared workspace, the viewer can tell what the object is. Mutual-gaze is knowing whether someone is looking at you. This is more commonly known as *eye contact* and has some well documented functions in regulating conversation. In an object-focused gaze awareness situation (Roberts et al., 2009), while a remote partner (attention source) fixed his or her gaze, the local participant (observer) was asked: “Which object is being looked at?” In the mutual-gaze situation, the local participant was asked: “Are you being looked at?” Technically, gaze awareness may be achieved through physical alignment of cameras and displays to enable natural lines of sight operating within Chen’s offset threshold (Chen, 2002) of 1° horizontal and 5° vertical. While the 360° and multi-view capabilities of a spherical projection are novel, it is not clear whether observers can interpret gaze direction on closely spaced target objects. The direction of a person’s gaze is one feature that is relevant in judging objects of interest in an environment.

2.3.2 Perception of Head and Eye-Gaze Direction.

Early work indicates that gaze direction may be perceived by both the direction in which the head is oriented and the eyes’ position relative to the head (Gibson & Pick, 1963). Other research has focused on studies in which the eyes and the head were counter-rotated to varying degrees while maintaining fixation on the subject (Gibson & Pick, 1963; Anstis, Mayhew, & Morley, 1969). These studies consistently showed an interaction between eye and head position in the perception of gaze direction. Gibson and Pick (1963) examined three head-gaze conditions: head to front, left, and right. In each condition, an observer at a distance of 2 m gazed at seven positions in a prearranged random order, each 0.1 m apart on a wall behind participants. Participants made “yes” or “no” judgments of whether they felt they were being looked at. The frequency distributions of “yes” judgments showed a head-turn effect such that when the target’s head was rotated in one direction, participants’ judgment tended to perceive gaze to be rotated in the opposite direction. In addition to the three head-gaze conditions, Anstis et al. (1969) investigated three orientations of a TV screen. They found three effects. First was a similar effect to the head-turn effect. Second was a TV-screen-turn effect where the apparent displacement of the perceived direction was in the same direction as the turn of the screen. Third was an overestimation of the deviation of the looker’s gaze from the straight ahead. They suggested that the convex curvature of the screen probably caused the TV-screen-turn effect. Overestimation was found to increase with the complexity of the viewing condition. Overall, these studies suggest that observers may be constructing a mental line based on the head orientation before judging the eye direction relative to the head. Despite the importance of the head as an attentional cue, there has been relatively little research on the perception of its orientation. Troje and Siebeck (1998) quantified accuracy of head orientation discrimination under varying illumination conditions with the eyes pointing directly forward. Discrimination was shown to be most accurate within the ±15° range of forward gaze directions but was markedly poorer at 30° head rotation. This was also observed by Wilson et al., who also found that changes in head orientation could be perceived even when the internal features of the head or the outline head contour is removed, suggesting that the deviation of nose angle may be a likely cue (Wilson, Wilkinson, Lin, & Castillo, 2000).

The SphereAvatar (Oyekoya et al., 2012) presents an evaluation of the use of spherical displays for representing remote participants, but focuses on head-gaze and employs the static eye-gaze condition. In this paper, we investigate the relationship between human-gaze (i.e., both head-gaze and eye-gaze) and observer's viewing positions in different display configurations.

2.3.3 Evaluation Frameworks. Nguyen and Canny (2005) proposed a framework for evaluation with three variables: attention source, attention target, and observer. The attention source is a person who provides attention to the attention target. In our system, we refer to the attention source as the "actor." The attention target is an object which could be a person or anything else that receives attention from the source. The observer is the person who is trying to understand the presented information about attention including its source, its target, and any attached meaning.

Our experiments are designed to evaluate the system's accuracy of preserving gaze directions, based on previous quantitative design reviewed earlier. However, the camera and display configuration are very different. Moreover, we also look into different influencing factors, such as the seat positions and target positions.

3 Multiple View-Port Spherical Video Conferencing System

The goal of our system is to allow local users to perceive the eye-gaze of a remote user accurately. Figure 1 depicts the system design. A remote user (the *actor* in the *remote room*) is captured by 11 capturing cameras controlled by two PCs. In the *local room*, a single PC renders video on a spherical display which is seen by a local user, the *observer*. Depending on the observer's position, the most appropriate camera feed is streamed from one of the two camera controller PCs to the rendering PC. Streaming is done using TCP.

3.1 Semicircular Camera Arrays

In the remote room, 11 low-cost PlayStation® Eye USB digital cameras are mounted on a half annular table with an inner radius of 405 mm at every 15°, as illustrated in Figure 2. The cameras are set to a 56° field of view setting. The cameras capture at 30 Hz with 320 × 240 pixel resolution.

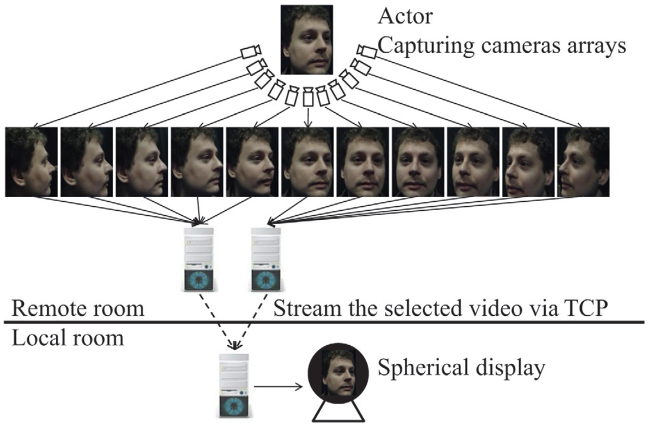


Figure 1. *Diagram of the directional spherical video conferencing system.*

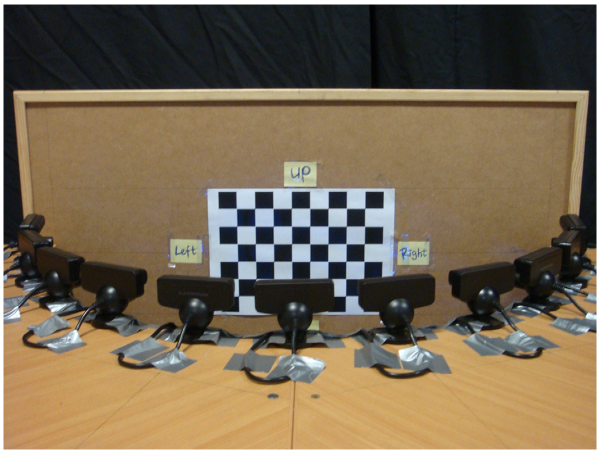


Figure 2. *Camera calibration setup.*

We manually adjust the cameras to look at the point above the center of the half annular table. We used Zhang's camera calibration method which involves showing all of the cameras a flat checkerboard target in at least two different orientations (Zhang, 2000). We then use the Camera Calibration Toolbox for MATLAB to compute the cameras' positions and orientations accurately (in Figures 2 and 3). The computed positions and orientations of the cameras are used in the rendering process, so that the video can be rendered accurately.

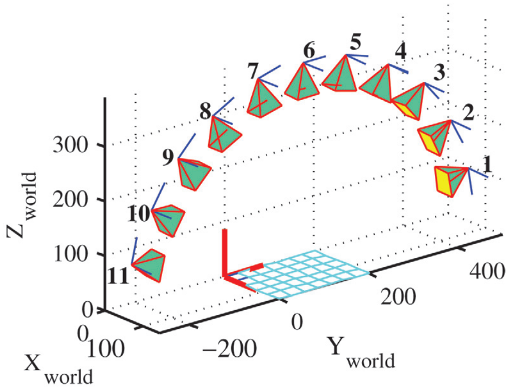


Figure 3. *Camera calibration result.*

3.2 Directional Spherical Screen

In the local room, a spherical display was located at the center of a half annular table which is the same size as the one in the remote room. Eleven observer viewports were set around the half annular table with a radius of 1810 mm at every 15° which exactly line up with each camera in the remote room. The spherical display is the commercially available Magic Planet display by Global Imagination®. The Magic Planet is a projection display device with a 16″ sphere-shaped surface and an internal fisheye lens to project imagery onto the inside of the sphere.

The presentation of the remote participant onto the sphere is done in four main stages as shown in Figure 4. First, a sphere acts as a proxy geometry of a human head, onto which the video images are displayed using projected texture mapping (PTM). PTM is a method of texture mapping described by Segal that allows the texture image to be projected onto the scene as if by a “slide projector” (Segal, Korobkin, Van Widenfelt, Foran, & Haeberli, 1992). According to the observer’s viewport, the video captured by the corresponding capturing camera is selected. This video is projected onto the polyhedron, which is approximately human-head size. This ensures that the capturing camera, the “slide projector,” and the observer’s eye are in close alignment.

Next, we render this proxy geometry onto an environment map. The idea of storing environment maps as cube maps is proposed by Greene where six sub-images represent the six different faces of a cube (Greene, 1986). We render the scene onto an environment map using six cameras positioned outside the cube at the position of the observer’s eye. Each of the six facets of the cube map is thus rendered using the non-symmetric view volumes.

Then, we draw a 3D sphere using an environment map. Environment mapping proposed by Blinn and Newell (1976) simulates the reflectance of a surface, by using the reflected eye vector as a lookup into the texture rather than a simple texture coordinate. We render a sphere with the environment map as its texture in order to generate a 2D distorted image that is suitable for projection through a fish-eye lens (Oyekoya et al., 2012).

Finally, the projected light travels through the bottom of the sphere, allowing the sphere to be completely illuminated except for the area immediately around the lens itself and achieving 360° horizontal visibility. The observer sees the head approximately life-sized.

4 Independent Variable

4.1 Display Modes

The display mode variable consists of five display types: Face to face (*Face*), sphere display (*Sphere*), fixed single video flat display (*Fixed single flat*), fixed multiple video flat display (*Fixed multiple flat*), and free multiple video flat display (*Free multiple flat*). We ensured that the vertical alignment of the eye-gaze of the actor, the eye level of observers, the eye level of the video of the actor on the spherical or fixed single video flat display, the capturing cameras, and the attention target cards were the same. This ensured equivalence in stimuli alignment and apparent size between the four display conditions and the face-to-face condition. Note that although the system as designed and built is a real-time collaborative system that can connect a remote room to a local room, video was recorded to disk and replayed for the purposes of control of the experimental stimuli in these four display conditions.

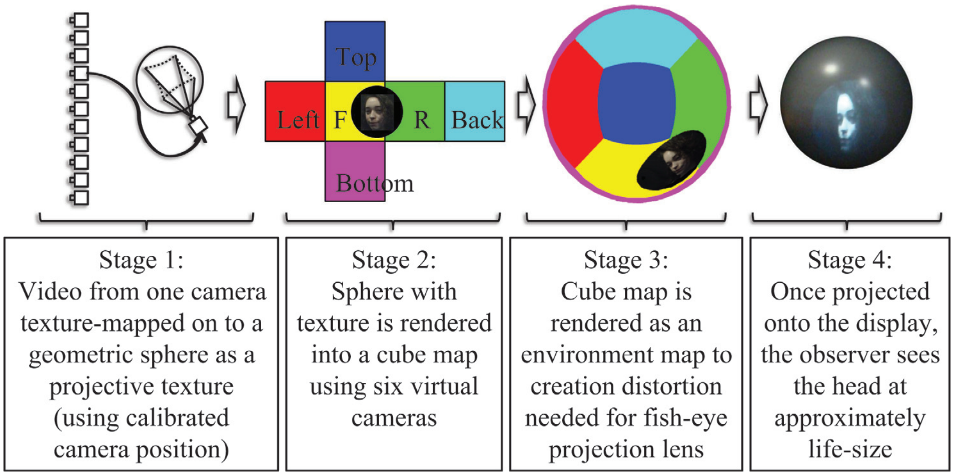

Top
Left
F
R
Back
Bottom
Stage 1: Video from one camera texture-mapped on to a geometric sphere as a projective texture (using calibrated camera position)
Stage 2: Sphere with texture is rendered into a cube map using six virtual cameras
Stage 3: Cube map is rendered as an environment map to creation distortion needed for fish-eye projection lens
Stage 4: Once projected onto the display, the observer sees the head at approximately life-size

Figure 4. *Illustrating stages of the rendering pipeline. Note that in the cube map and the 2D distorted image, the colored backgrounds (stage 2) represent six different faces of a cube and are actually rendered as black (stage 4).*

- Face to face

  Figure 6(a) shows the face-to-face condition, in which the observer and actor were in the same room. The actor sat at the center position of the table and the observer sat on the outside. The actor was wearing small headphones listening to the same audio instruction as was used when recording the videos for the display conditions.
- Fixed single video flat display

  The spatial arrangement of this condition was almost identical to the sphere display condition. In this case, the conventional flat display and only the center camera (lined up with position 6) were used, as presented in Figure 6(b). Image quality remained the same. This condition mimicked the commonly found distorted video conferencing system where the actor is not always lined up with the capturing camera, and the observer is not always lined up with the display screen.
- Sphere display

  In Figure 6(c), the observer observed the prerecorded video on the sphere display. Hence, the actor and the observer achieved the line-of-sight effect.
- Fixed multiple video flat display

  This condition was similar to the fixed single video flat display condition except all the capturing cameras were used, as shown in Figure 6(d). According to the observer's position, the proper video is selected. The actor is always lined up with the capturing camera, but the observer might be looking obliquely at the screen.
- Free multiple video flat display

  This condition was alike to the fixed multiple video flat display condition except the flat display was rotated based on the observer's position, allowing the observer to look directly at the screen, as shown in Figure 6(e). Hence, the actor and the observer achieved the line-of-sight effect.

4.2 Seating Positions

We define the participants' seating positions at 30°, 45°, 60°, 75°, 90°, 105°, 120°, 135°, and 150° relative to the display. Therefore there were nine levels of categorical variables of seating positions. The distance between participant and display remained constant.

4.3 Target Numbers

Twenty-three numbered target cards were placed on the semicircular table from 15° to 165° at every 7.5°. Therefore, there were 23 levels of categorical variables of target numbers. The distance from target position to participant and display remained constant.

5 Experiment 1

The purpose of the first experiment was to demonstrate that the combination of a spherical display and a camera array can better represent the actor's gaze than a fixed single video flat display. We measured the effectiveness of the displays by measuring the ability of observers to accurately judge which target the actor was gazing at for three display modes, as presented in Figures 6(a), 6(b), and 6(c). Also, we investigated the situation when the observer was not seated in the same direction as the camera that was observing the actor. We formed two hypotheses.

H1. We expect that both face-to-face and sphere display will demonstrate higher levels of accuracy (the observers are accurate if they successfully identify the correct target) than fixed single video flat display when the observers are in varied positions. We further expect face to face to be better than sphere display.

H2. For both sphere display and fixed single video flat display, we expect that if the observer is not seated in the same direction as the camera that is observing the actor, the accuracy will be worse than if the camera chosen for the display is aligned with the observer's position.

5.1 Method

5.1.1 Participants. Sixty participants, students and staff at University College London, were recruited to take part as observers in our user study. Twenty groups of three were used for testing and each group experienced one of three different conditions (sphere display, fixed single video flat display, or face to face). Eight more participants were actors in these experiments: four actors were recorded on video for the sphere and fixed single video flat condition and four acted in the face-to-face condition.

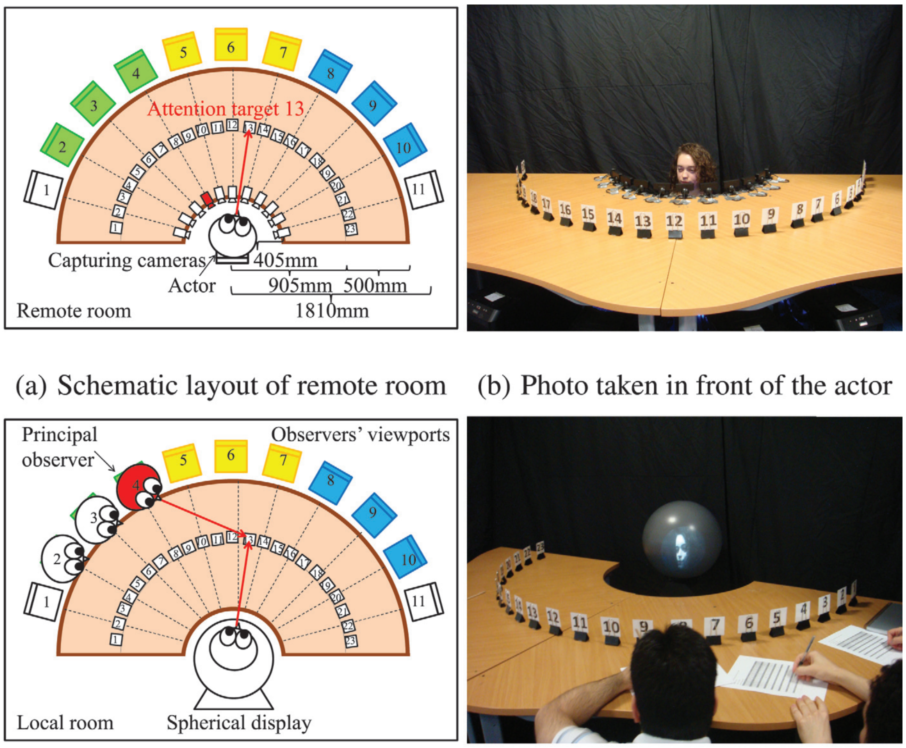
5.1.2 Apparatus and Materials. For the two display conditions, we video-recorded the actors' head movements, presented in Figures 5(a) and 5(b). The actor sits at the center position of the half annulus table and his or her head is captured by 11 video cameras. The actors listen to an audio recording that instructs them to look at the gaze target cards. A new target is given every 10 seconds. The targets are randomly ordered, and each one is gazed at twice, amounting to 46 targets in the audio instruction and thus in the recorded videos. Four participants were actors, and thus four sets of 11 videos were generated.

5.1.3 Procedure. Nine different positions for observers were investigated. Observers took part in groups of three. In all conditions, the group performed three trials. On each trial, the group would sit in positions 2, 3, and 4; or 5, 6, and 7; or 8, 9, and 10.

For each trial, each observer was given a sheet of paper with an empty grid of 46 squares. In all three conditions, the actor or the video of the actor reoriented to a new target card every 10 seconds. At the same time, an audio prompt to the observers instructed them that this was a new target. They would then judge which target (1–23) the actor was gazing at and then write this in the relevant grid square.

For the face-to-face condition, the three observers and the actor were in the same room. The actor sat at the center position of the table and the three observers sat on the outside. The actor was wearing small headphones listening to the same audio instruction as was used when recording the videos for the display conditions. The actor performed the sequence of gazes three times. On each repetition, the group of three observers moved to another one of the group positions.

For the sphere display condition, the three observers observed the pre-recorded video on the sphere display, presented in Figures 5(c) and 5(d). For each group position, one of the observers was the principal observer. The video corresponding to the principal observer's position was shown on the display. Each group saw the



(a) Schematic layout of remote room (b) Photo taken in front of the actor

(c) Schematic layout of local room (d) Photo taken behind observer 4

Figure 5. *Example of experiment setup: The actor gazes at the target card 13 captured by semicircular camera arrays in the remote room. Since the principal observer is sitting in viewport 4, the video captured by camera 4 is presented on the sphere display, which lines up with observer 4.*

actor's video three times. On each repetition, the group of three observers exchanged positions; hence, each observer became a principal observer at least once.

For the fixed single video flat display condition, the three observers observed the pre-recorded video on the fixed single video flat display. The video was always from camera position six, simulating a simple webcam setup where the observers might be looking obliquely at the screen, and the actor looking obliquely at the camera.

The experiment took about 20 minutes.

5.2 Results

A summary of the results of experiment are presented in Figure 7. In each figure, the horizontal axis indicates viewport position (p) from 2 to 10. The angle of viewport position (α) in degrees is from 30° to 150° at every 15° relative to the center of conferencing table.

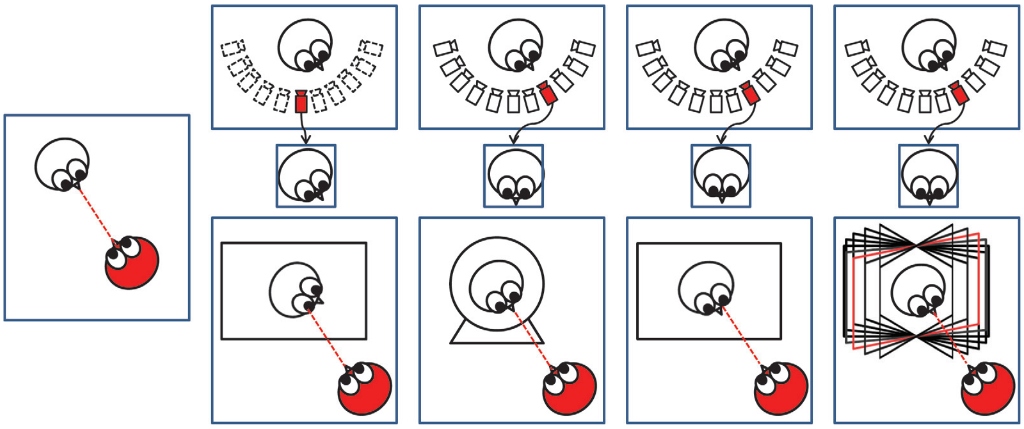
$$\alpha = p \times 15^\circ. \qquad (1)$$

The primary measurement in our results is the accuracy rate in perceiving the attention target. The accuracy rate is the percentage of accurate prediction over total prediction.

We then define systematic bias (β_i) to be the difference between the actual target number (t_{ai}) and the observer's perceived attention target number (t_{oi}) converted to degrees, based on attention targets being 7.5° apart from each other.

$$\beta_i = (t_{ai} - t_{oi}) \times 7.5^\circ. \qquad (2)$$

Each observer indicates 46 target positions in each of three trials. There are 12 observers in the face-to-face



(a) Face to face (b) Fixed single video flat display (c) Sphere display (d) Fixed multiple video flat display (e) Free multiple video flat display

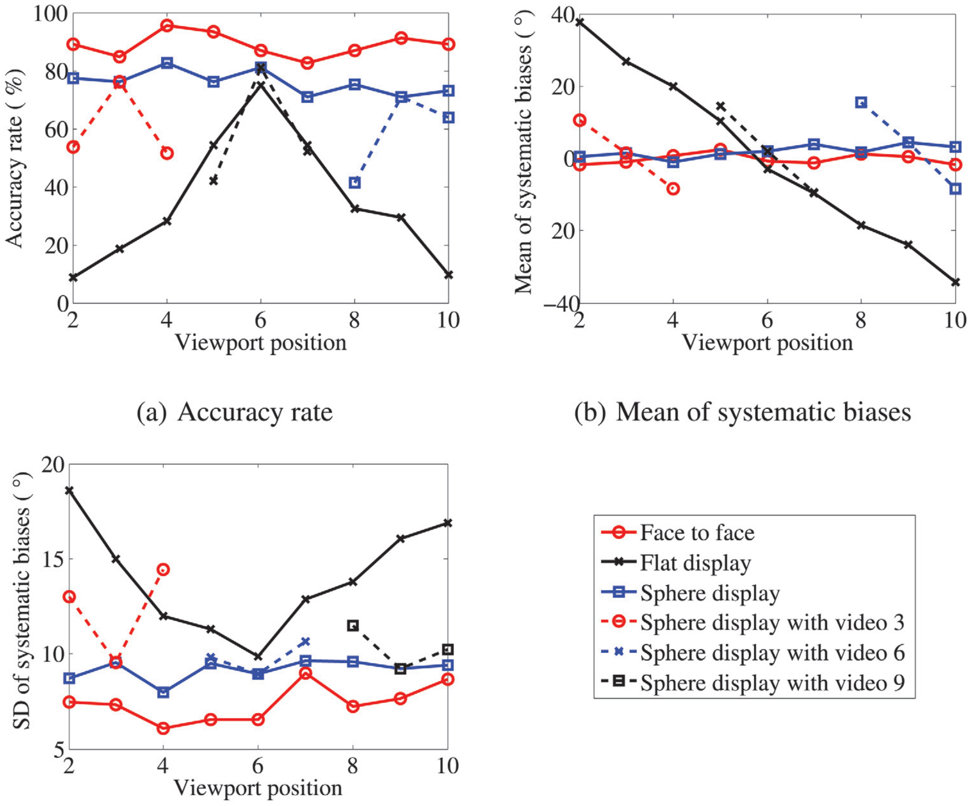
Figure 6. *Five levels of categorical variable media representation. The observer (in red) is seated at viewport 4, therefore camera 4 (in red) is enabled. Top row: capturing actor in the remote room; middle row: captured video for transmission; bottom row: view of screen showing actor's gaze direction in the local room. The dashed red line is the actual actor's gaze direction.*

condition (four groups of three) and nine observer seat positions. Thus, there are 184 (46 × 3 × 12/9) rating events in each seating position. Similarly, there are 184 rating events in each seating position for the fixed single video flat display. For the sphere display, there are 36 observers (12 groups of three) but only one of the group is in the principal position. Thus, there are also 184 (46 × 3 × (36/3)/9) rating events for principal observers in each of the nine observer seating positions. However, in the following analysis, we include some data from the secondary observers. In particular, for seating positions 3, 6, and 9, we analyze the 184 rating events for the observer seated on the left and 184 rating events for the observer on the right. This gives us a view of how important it is to use the correct video for the observer position.

5.2.1 Accuracy Rate. The result of accuracy rate in different conditions is shown in Figure 7(a). For the fixed single video flat display, with the observer at the central viewport, the accuracy rate is 75%. However, the accuracy rate drops off symmetrically as the observer position diverges from the central position. This is expected, as when the observer is not seated in position 6; he or she will still see the video taken from the camera at position 6.

The results for face-to-face and sphere display are not affected by viewport position and the average accuracy rates are 89% and 76%, respectively. The average accuracy rate of sphere display is slightly lower than face to face, but similar to the observer sitting at the central position in the fixed single video flat display condition. The fact that the accuracy does not vary with observer position for the sphere display when considering the principal observer supports the primary hypothesis. The performance of the sphere display at the extreme positions (2 and 10) is significantly above that of the fixed single video flat display.

When we consider the secondary positions in the sphere display—the three "three point hat" graphs in Figure 7(a)—we see that it is very important that the camera selected be aligned with the observer position. Considering the principal observer at position 3, we see that the observer in position 2, observing the video from position 3, has a performance of under 54% compared to the accuracy of almost 76% for the principal observer seated immediately to his or her right. This pattern is repeated for all secondary observers.



(a) Accuracy rate

(b) Mean of systematic biases

(c) Standard deviation of systematic biases

Figure 7. *Result for analyzing the actual targets and perceived targets in different treatment conditions.*

The difference between face-to-face performance and sphere display performance may be due to video quality. We note that for observer position 6 on the fixed single video flat display, the ideal situation for this position, the accuracy is very similar to the sphere display at this position. This indicates that the sphere display is no worse than the fixed single video flat display, but it has the advantage that it has the same apparent size in the different observer positions.

5.2.2 Mean of Systematic Biases and Standard Deviation.

Next, we analyzed the mean and standard deviation of systematic biases for the actual targets and observers' perceived targets in different treatment conditions, as shown in Figures 7(b) and 9(c). For the face-to-face and sphere display conditions, the observer position has no significant effect on the mean of systematic biases, which is around 0°. The standard deviation of the systematic biases for the sphere display is higher, but there are no systematic biases, indicating that the observers are generally finding it harder to determine gaze.

In contrast, for the fixed single video flat display, the mean of systematic biases varies linearly according to viewport position. We used the first-order MATLAB Polyfit function to generate the coefficients of the polynomial to simulate a curve to fit the data and found a relationship between the systematic biases of mean and angle of viewport position:

$$\begin{aligned}\sigma(\beta_i) &= -0.6\alpha + 54.27^\circ \\ &= 0.6 \times (90^\circ - \alpha) + 0.27^\circ. \qquad (3)\end{aligned}$$

The linear model of systematic biases in the fixed single video flat display condition is interesting in that it suggests that the observer's judgment of gaze angle from the front is only 60% of what it should be. Therefore, for the fixed single video flat display, the observer perceives the actor to be looking more directly straight out of the display.

6 Experiment 2

In the second experiment, we introduced two more display modes, shown in Figures 6(d) and 6(e). We compare the sphere display with fixed multiple video flat display and free multiple video flat display to demonstrate the improvement of representing the actor's gaze by using the camera array and the spherical display simultaneously. In addition, we used the mixed design analysis of variance (ANOVA) as a more reliable statistical analysis to further investigate factors influencing the observers in perceiving targets in different conditions. We specifically form the three hypotheses.

H3. We explore the level of error with which observers can discriminate the actor's gaze orientation for all five display modes. Specifically we measure the ability of participants to identify which set of targets the actor appears to be gazing toward. Given the five display modes, we expect that the level of error of observers' performance would follow the trend below:

$$\begin{aligned} &\textit{Face} < \textit{Sphere} < \textit{Free multiple flat} \\ &\quad < \textit{Fixed multiple flat} < \textit{Fixed single flat}. \qquad (4) \end{aligned}$$

H4. We then explore the influence of seat position. We expect that face-to-face, sphere display, and free multiple video flat displays will show a similar level of error for all seat positions. However, the level of error will increase symmetrically as the observer position diverges from the central position for fixed multiple video flat display and fixed single video flat display.

H5. We further explore the influence factor of target position. We expect that face-to-face, sphere display, and free multiple video flat displays will show a similar level of error while observing all numbered targets. However, there should be systematic biases for fixed multiple video flat display and fixed single video flat display.

6.1 Method

6.1.1 Participants. Forty participants, students and staff at University College London, were recruited to take part as observers in our user study. Each participant judged only one of five display modes, a between-subjects design. But a within-subjects design was employed for the two factors of nine seating positions (2–10) and the 23 target numbers (1–23). We randomly mixed the seating positions and target numbers in order to reduce any confounding influence of the orderings such as learning effects or fatigue.

Two further participants were actors in this experiment: one actor was recorded on video for four video display conditions and the other one acted in the face-to-face condition.

6.1.2 Apparatus and Materials. For the four display conditions we recorded the actor's performance. The actor sits at the center position of the half annular table and his or her head is captured by 11 video cameras. The actor listens to an audio recording that instructs him or her to look at the gaze target cards. A new target is given every 10 seconds. The targets are randomly ordered, giving 23 targets in the audio instruction and thus in the recorded videos. A set of 11 videos was generated.

6.1.3 Procedure. The experiment took about 30 minutes for each participant. Upon arrival, each participant was assigned to one of five treatment conditions. Eight observers were investigated for each treatment condition.

Nine different positions for each observer were investigated. Observers were initially seated in one of the nine positions in a counterbalanced random order. For each trial, each observer was given a sheet of paper with an empty grid with 23 squares. Every 10 seconds, the actor reoriented to a new target card. At the same time, an audio prompt to the observers instructed them that this was a new target. They would then judge which target (1–23) the actor was gazing at and write this in the relevant grid square. After each trial, the session was paused to allow the participants to change seating position accordingly.

6.2 Results

6.2.1 Level of Error.

The primary measurement in our results is the level of error in perceiving the attention target. We define error (ε_i) to be the absolute value of difference between the actual target number (t_{ai}) and the observer's perceived attention target number (t_{oi}) converted to degrees, based on attention targets being 7.5° apart from each other.

$$\varepsilon_i = |t_{ai} - t_{oi}| \times 7.5^\circ. \qquad (5)$$

The dependent variable data (ε_i) were entered into a mixed design analysis of variance (ANOVA) with the three factors of display condition, seating position, and target position. We used Mauchly's test of sphericity to validate our repeated measures factor ANOVAs, thus ensuring that variances for each set of difference scores were equal. Mauchly's test indicated that the assumption of sphericity had not been violated.

Results reveal that there was a significant main effect of display condition, $F(4, 8279) = 684.842, p < 0.01$, and post-hoc Tukey tests revealed significant mean differences between each of all those displays. The face to face ($M = 5.104$) achieved the lowest level of error, followed by sphere display ($M = 6.916$), free multiple video flat display ($M = 8.262$), fixed multiple video flat display ($M = 10.375$), and then fixed single video flat display ($M = 21.162$). See Figure 8. This supports the third hypothesis.

While this absolute level of error is a good basic measure, it effectively accumulates the positive and negative systematic biases. In order to get a more detailed view of effectiveness of different displays in perceiving the attention target, whether there are left or right systematic biases, how seat position varies, and the target position variable effect, the result of different display conditions must be taken into account.

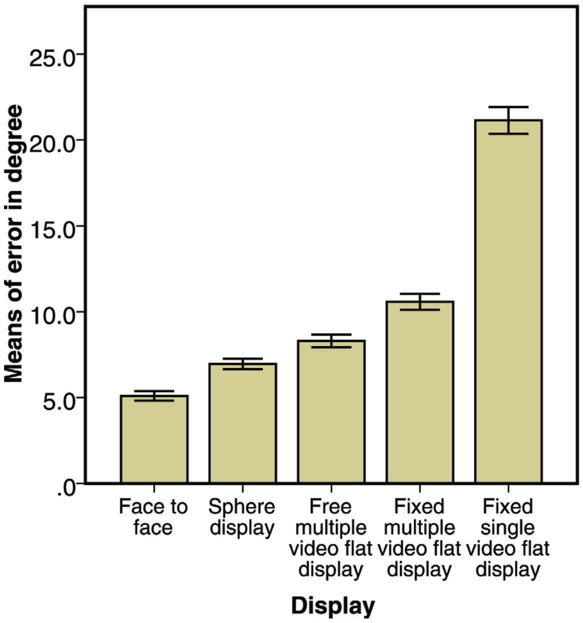


Figure 8. *Bars show estimated marginal means of error in different treatment conditions; error bars show 95% CI of the means.*

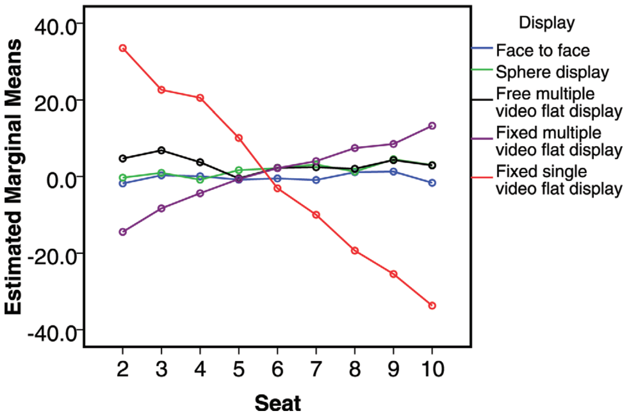


Figure 9. *2-way interaction: estimated marginal means of biases in degree.*

6.2.2 Systematic Biases.

Similarly, we then look into systematic bias (β_i), which is defined in Experiment 1. First, we look into two-way interaction. Figure 9 shows the average systematic bias of different seat positions under five different display conditions. For face-to-face, sphere display and free multiple videos flat display, the average systematic bias curves roughly around 0° and does not change over different seating positions. Moreover, the face-to-face condition is the most stable and the closest approximate to 0°, followed by sphere display and then the free multiple video flat display. By contrast, the average systematic bias varies linearly according to seat position for fixed multiple video display and fixed single display. The absolute value of systemic bias is the error defined earlier. The lines of fixed multiple video flat display and fixed single video flat display are symmetric about $seat = 6$. Therefore, seat variable has an effect only for fixed multiple video and fixed single video conditions. This supports the fourth hypothesis.

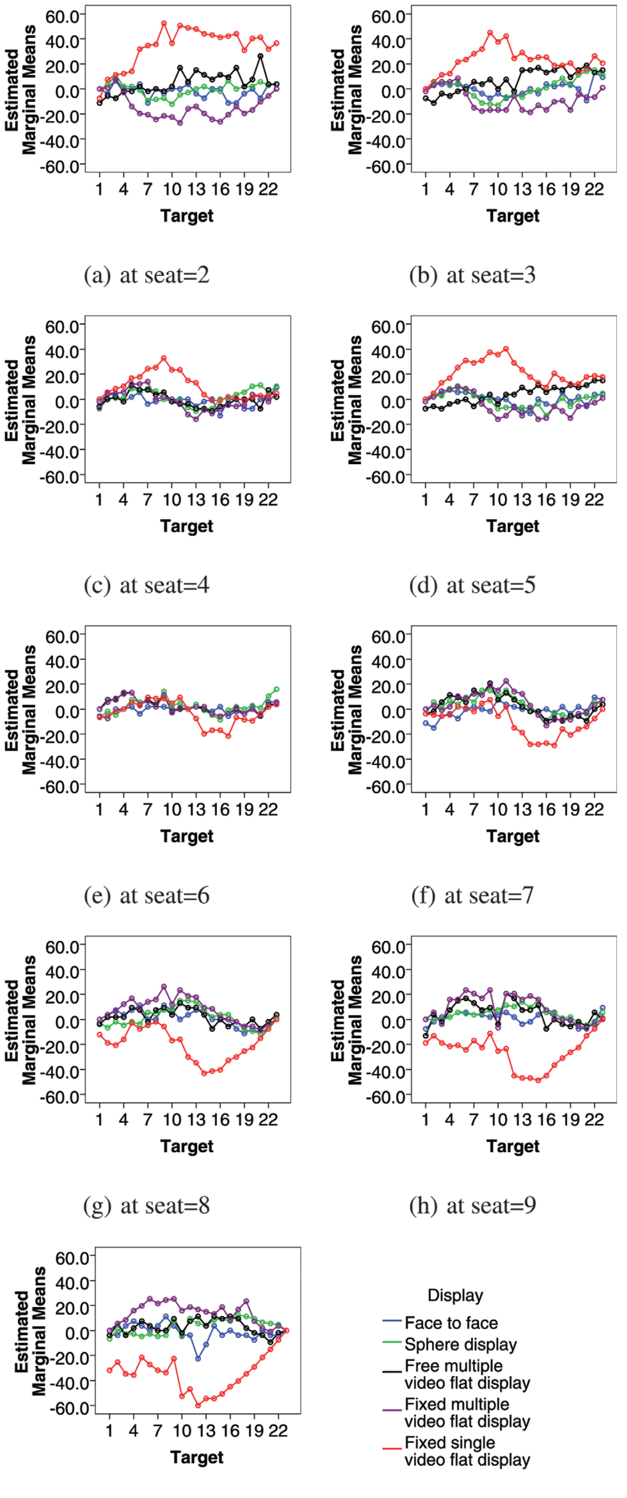
We conduct a three-way interaction to investigate whether the seat × display is the same for all targets. We use the estimated marginal means to interpret the three-way interaction (see Figure 10). For face-to-face, sphere display, and free multiple videos flat display, the average systematic bias curves are basically around 0° with slight fluctuations among different target positions.

However, for the fixed multiple videos flat display and the fixed single flat display, the average systematic bias varies over different target positions. The fixed single video display has more biases compared to the fixed multiple videos display. This supports the fifth hypothesis.

Interestingly, Figure 10 shows that the curves can be modified into symmetrical parts for each pair of seat positions 2 and 10; 3 and 9; 4 and 8; and 5 and 7, which are symmetrically arranged on both sides of the center seat position 6. For seat position 6, the curve itself is symmetry relative to point $(12, 0)$.

6.2.3 Linear Regression for Systematic Biases.

As we discussed in the previous section, the mean of systematic biases varies linearly according to viewport position for the fixed single video flat display (red line in Figure 9) and multiple video flat display (purple line in Figure 9) conditions. However, the mean of systematic biases are sloped in opposite directions in those two conditions.



(a) at seat=2

(b) at seat=3

(c) at seat=4

(d) at seat=5

(e) at seat=6

(f) at seat=7

(g) at seat=8

(h) at seat=9

(i) at seat=10

Figure 10. *3-way interaction: estimated marginal means of biases in degree.*

A simple regression was carried out to ascertain if the angle of viewport position (α) can predict the systematic biases of fixed single video flat display ($\beta_{fixed\ single\ flat}$). A strong correlation was found between the angle of viewport position and the systematic biases of fixed single video flat display, $r = .831$; and the regression model predicted 69% of the variance. The model was a good fit for the data, $F(1, 1654) = 3685.526$, and $p < .001$. The linear regression model is presented in Equation 6, $b = -.57$, $t(1654) = -60.709$, and $p < .001$. This further confirms the result in Equation 3 in the first experiment.

Similarly, standard simple regression analysis was conducted to evaluate how well the angle of viewport position (α) predicted the systematic biases of fixed multiple video flat display ($\beta_{fixed\ multiple\ flat}$). The angle of viewport position was significantly related to the systematic biases of fixed multiple video flat display, $F(1, 1654) = 814.257$, and $p < .001$. The correlation coefficient was $r = .574$, indicating that approximately 33% of the variance of the systematic biases of fixed multiple video flat display can be accounted for by angle of viewport position. The regression equation for predicting the systematic biases of fixed multiple video flat display was shown in Equation 7, $b = .221$, $t(1654) = 28.535$, and $p < .001$.

$$\begin{aligned}\beta_{fixed\ single\ flat}(\alpha) &= -0.57\alpha + 50.804^\circ \\ &= 0.57 \times (90^\circ - \alpha) - 0.496^\circ. \quad (6)\\ \beta_{fixed\ multiple\ flat}(\alpha) &= 0.211\alpha - 18.13^\circ \\ &= -0.211 \times (90^\circ - \alpha) + 0.86^\circ. (7)\end{aligned}$$

7 Discussion

7.1 Camera Arrays vs. Single Camera

The line of fixed single video flat display has a higher slope value compared to fixed multiple video flat display (in Figure 9). This indicates a steeper incline and higher systematic biases. In some extreme cases, such as in seat positions 2 and 10, the observer had more difficulty in perceiving targets in fixed single video flat display. The fixed multiple video display improves the system's ability to represent the actor's gaze, by lining up the capturing cameras using camera arrays.

7.2 Directional Projection

The gradient of line indicates systematic biases in fixed multiple video flat display (see Figure 9); however, the line is always stable around 0° for the free multiple video flat display. The observer can perceive targets better in free multiple video flat display, particularly when seat position is further apart from the center. The free multiple video flat display improves the system's ability to present the actor's gaze, by lining up the projector (screen).

7.3 Sphere vs. Free Multiple Video Flat Display

Figure 8 shows that the level of error in the sphere display is only slightly lower than the free multiple video flat display condition. However, in free multiple video flat display, we have to manually rotate the flat display for each viewport position for each observer, which is impossible for practical video conferencing.

Previous findings (Anstis et al., 1969; Pan & Steed, 2014) suggested that biases occur differently while observing convex, flat, and concave surfaces. For this spherical display, we plan to further explore this finding, with our next step being to collect data for more viewing angles.

7.4 Video Quality

The higher level of error in Figure 8 and larger fluctuation around 0° in Figure 9 in sphere display compared to face to face shows that observers can better perceive the actor's attention target in face to face. This suggests that there is more work to be done on the quality of representation of gaze with such displays.

7.5 Seat Position

From this discussion, the seat position has a linear effect on the fixed single flat display and fixed multiple video display. Observers could interpret the direction of actor-gaze of the sphere display more accurately than the free multiple video flat display and similarly to the face-to-face condition for all seat positions.

7.6 Linear Model for Predicting Distortion

The study by Roberts et al. (2009) found that the correct viewing of the sides of the face is important for the interpretation of gaze. Large errors in estimation coincided with either the face being viewed from the wrong perspective or unevenly lit. This is in line with our results, where we modeled the systematic biases for two flat display configurations. We found the negative linear correlation between the angle of viewport position and the systematic biases of the fixed single video flat display in Equation 6, and the positive linear correlation between the angle of viewport position and the systematic biases of the fixed multiple video flat display in Equation 7, respectively. This indicates that the fixed single video flat display is biased in the opposite direction to the fixed multiple video flat display condition (see Figure 6). While the biases may have been caused by incorrect viewing angles in both conditions, the single capturing angle of the fixed single video condition may have caused the bias to be in the opposite direction. Also, this effect appears very reliable and this means that it may be possible to model and thus predict the distortion.

8 Limitations, Applications and Future Directions

Our spherical display could be used in a teaching scenario or a telesurgery application where a remote person instructs a local user. The local user could perceive precise spatial information from any viewpoint in the room whereas flat displays are visible only from the front. Our current system is used for asymmetric conversations; however, systems using similar principles could be configured to support symmetric conversations, by arranging camera arrays that are denser but further from the users. Also, our spherical display could be mounted on a robot to include haptic (hands or body) or mobility capabilities. Current telepresence robots generally use flat screens, with a webcam view of the remote participant. This webcam view could be rendered on to a spherical display and oriented, independent of the robot base, to face in any direction. This would support more rapid head movement than turning the base itself. This could help in social situations where attention needs to be directed quickly. The direction of this surface video view could be driven in multiple ways, including following the eye or head direction of the remote person.

An interesting question is the potential support for multiple viewers. The evaluation of the secondary positions in the first experiment, the three “three point hat” graphs in Figure 7(a), demonstrated that the gaze cues are preserved only for the principal observer. This is because the position of the observer is needed in order to render the head correctly for that perspective. The spherical display could be made for multiple viewers. The inflated display mode of SphereAvatar (Oyekoya et al., 2012) supported multiple viewers in avatar-mediated teleconferencing. For video-mediated teleconferencing, we could project the whole head by using the similar idea proposed by Jones et al. (2009) in the One-to-Many 3D Video Teleconferencing System.

In this system, the video texture is projected on a sphere. An alternative approach would have been to project onto an ellipsoid or a more “head-shaped” object than a sphere; however, this would have worked for head rotations around the vertical axis while the projection would be severely distorted for rotations around other axes. In addition, we will investigate novel rendering methods to avoid the steep drop in accuracy when the observer is not aligned with the cameras by interpolating between videos. Furthermore, we will investigate less constrained positioning of the cameras and different eye-lines. As noted, although the experiment used recorded data, the system can run in a live, automatic camera switching mode and thus we will investigate how users utilize movement to control the video.

9 Conclusions

We developed a novel spherical display system in order to give an observer some of the advantages of meeting face to face without the disadvantages of traveling. It offers a 360° view whereas flat display is visible only from the front. By using a surrounding camera array, we allow principal observers to accurately tell where the actor is looking, from multiple observing positions at all angles. The captured video is projected from the bottom of a spherical display, which successfully maintains the gaze fidelity without reducing the quality of the video and complexity of the display system, compared to using a half mirror. This motivates further development of video conferencing systems that exploit multiple cameras and non-planar displays. We are the first to compare situated display and flat display in preserving gaze direction. We have demonstrated that the sphere display preserves the accuracy of observing the actor's gaze direction, even at extreme seat positions. This may be due to the ability of sphere displays to produce a correct view. Furthermore, we proposed two linear models for predicting the spatial distortion introduced by misalignment of capturing cameras and observer's viewing angles. Therefore, we might be able to correct for this distortion in future display configurations. We designed a situated system that concentrates on the broader goal of teleconferencing by simulating face-to-face conversation versus just eye-gaze alone. Upcoming experiments will focus on case studies of more complex teleconferencing scenarios.

Acknowledgments

The authors would like to acknowledge the FP7 EU collaborative project BEAMING (248620) and the China Scholarship Council. We thank the UCL VECG group members for their general support and the anonymous reviewers for their feedback.

References

Acker, S. R., & Levitt, S. R. (1987). Designing videoconference facilities for improved eye contact. *Journal of Broadcasting & Electronic Media*, *31*(2), 181–191.

Adalgeirsson, S., & Breazeal, C. (2010). MeBot: A robotic platform for socially embodied presence. *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction* (pp. 15–22).

Anstis, S., Mayhew, J., & Morley, T. (1969). The perception of where a face or television "portrait" is looking. *The American Journal of Psychology*, *82*(4), 474–489.

Aoki, T., Widoyo, K., Sakamoto, N., Suzuki, K., Saburi, T., & Yasuda, H. (1999). Monjunochie System: Videoconference system with eye contact for decision making. *Institute of Electronics, Information and Communication Engineers Technical Report*, *98*(552), 9–14.

Arai, H., Kuriki, H., & Sakai, S. (1992). New eye-contact technique for video phones. *SID International Symposium Digest of Technical Papers* (pp. 149–152).

Arai, J., Okui, M., Yamashita, T., & Okano, F. (2006). Integral three-dimensional television using a 2000-scanning-line video system. *Applied Optics*, *45*(8), 1704–1712.

Atzpadin, N., Kauff, P., & Schreer, O. (2004). Stereo analysis by hybrid recursive matching for real-time immersive video conferencing. *IEEE Transactions on Circuits and Systems for Video Technology*, *14*(3), 321–334.

Benko, H., Wilson, A. D., & Balakrishnan, R. (2008). Sphere: Multi-touch interactions on a spherical display. *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (pp. 77–86).

Blinn, J., & Newell, M. (1976). Texture and reflection in computer generated images. *Communications of the ACM*, *19*(10), 542–547.

Bohannon, L. S., Herbert, A. M., Pelz, J. B., & Rantanen, E. M. (2013). Eye contact and video-mediated communication: A review. *Displays*, *34*(2), 177–185.

Buchner, G. (2006, May 23). *Interactive audio-visual system.* Google Patents. (US Patent 7,048,386)

Chen, M. (2002). Leveraging the asymmetric sensitivity of eye contact for videoconference. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 49–56).

Chen, W., Bouguet, J., Chu, M., & Grzeszczuk, R. (2002). Light field mapping: Efficient representation and hardware rendering of surface light fields. *ACM Transactions on Graphics*, *21*(3), 447–456.

Desai, M., Tsui, K. M., Yanco, H. A., & Uhlik, C. (2011). Essential features of telepresence robots. *Technologies for Practical Robot Applications* (pp. 15–20).

Eisemann, M., De Decker, B., Magnor, M., Bekaert, P., De Aguiar, E., Ahmed, N. . . . Sellent, A. (2008).

Floating textures. *Computer Graphics Forum* (Vol. 27, pp. 409–418).

Gibbs, S., Arapis, C., & Breiteneder, C. (1999). Teleport–towards immersive copresence. *Multimedia Systems*, *7*(3), 214–221.

Gibson, J., & Pick, A. (1963). Perception of another person's looking behavior. *The American Journal of Psychology*, *76*(3), 386–394.

Greene, N. (1986). Environment mapping and other applications of world projections. *Computer Graphics and Applications*, *6*(11), 21–29.

Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., . . . Staadt, O. (2003). Blue-c: A spatially immersive display and 3D video portal for telepresence. *Proceedings of the SIGGRAPH Conference on Computer Graphics and Interactive Techniques* (pp. 819–827).

Hindmarsh, J., Fraser, M., Heath, C., Benford, S., & Greenhalgh, C. (2000). Object-focused interaction in collaborative virtual environments. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(4), 477–509.

Hirayama, Y. (2009). One-dimensional integral imaging 3D display systems. *Proceedings of the 3rd International Universal Communication Symposium* (pp. 141–145).

Jones, A., Lang, M., Fyffe, G., Yu, X., Busch, J., McDowall, I., . . . Debevec, P. (2009). Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Transactions on Graphics (TOG) 28*(3), 64.

Jouppi, N., Iyer, S., Thomas, S., & Slayden, A. (2004). BiReality: Mutually-immersive telepresence. *Proceedings of the 12th Annual ACM International Conference on Multimedia* (pp. 860–867).

Kim, K., Bolton, J., Girouard, A., Cooperstock, J., & Vertegaal, R. (2012). TeleHuman: Effects of 3D perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2531–2540).

Lee, M. K., & Takayama, L. (2011). Now, I have a body: Uses and social norms for mobile remote presence in the workplace. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 33–42).

Leyrer, M., Linkenauger, S. A., Bülthoff, H. H., Kloos, U., & Mohler, B. (2011). The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments. *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization* (pp. 67–74).

Lincoln, P., Welch, G., Nashel, A., Ilie, A., Fuchs, H., et al. (2009). Animatronic Shader Lamps Avatars. *Proceedings of 8th IEEE International Symposium on Mixed and Augmented Reality* (pp. 27–33).

Matsuyama, T., & Takai, T. (2002). Generation, visualization, and editing of 3D video. *Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission* (pp. 234–245).

Matusik, W., & Pfister, H. (2004). 3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics (TOG)*, *23*(3), 814–824.

Mulligan, J., Isler, V., & Daniilidis, K. (2002). Trinocular stereo: A real-time algorithm and its evaluation. *International Journal of Computer Vision*, *47*(1-3), 51–61.

Nguyen, D. & Canny, J. (2005). MultiView: Spatially faithful group video conferencing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 799–808).

Norris, J., Schnädelbach, H., & Qiu, G. (2012). CamBlend: An object focused collaboration tool. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 627–636).

Okada, K., Maeda, F., Ichikawaa, Y., & Matsushita, Y. (1994). Multiparty videoconferencing at virtual social distance: MAJIC design. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (pp. 385–393).

Oyekoya, O., Steptoe, W., & Steed, A. (2012). SphereAvatar: a situated display to represent a remote collaborator. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems* (pp. 2551–2560).

Pan, Y., & Steed, A. (2012). Preserving gaze direction in teleconferencing using a camera array and a spherical display. *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video* (pp. 1–4).

Pan, Y., & Steed, A. (2014). A gaze-preserving situated multiview telepresence system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2173–2176).

Paulos, E., & Canny, J. (1998). PROP: Personal roving presence. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 296–303).

Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L., & Fuchs, H. (1998). The office of the future: A unified approach to image-based modeling and spatially immersive displays. *Proceedings of the SIGGRAPH Conference on Computer Graphics and Interactive Techniques.*

Roberts, D., Wolff, R., Rae, J., Steed, A., Aspin, R., McIntyre, M., . . . Steptoe, W. (2009). Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together. *Virtual Reality Conference* (pp. 135–142).

Sadagic, A., Towles, H., Lanier, J., Fuchs, H., Dam, A. van, Daniilidis, K., . . . Zeleznik, B. (2001). *National tele-immersion initiative: Towards compelling tele-immersive collaborative environments.* Presentation Given at Medicine Meets Virtual Reality Conference.

Saito, H., Baba, S., Kimura, M., Vedula, S., & Kanade, T. (1999). Appearance-based virtual view generation of temporally-varying events from multi-camera images in the 3D room. *Proceedings of the Second International Conference on 3-D Digital Imaging and Modeling* (pp. 516–525).

Sakamoto, D., Kanda, T., Ono, T., Ishiguro, H., & Hagita, N. (2007). Android as a telecommunication medium with a human-like presence. *Proceedings of the 2nd ACM/IEEE International Conference on Human Robot Interaction* (pp. 193–200).

Schreer, O., Chang, K., Hendriks, E., Schraagen, J., Stone, J., Trucco, E., & Jewell, M. (2002). Virtual team user environments—A key application in telecommunication. *Proceedings of eBusiness and eWork* (pp. 916–923).

Schreer, O., Kauff, P., & Sikora, T. (2005). *3D videocommunication.* Wiley Online Library.

Segal, M., Korobkin, C., Van Widenfelt, R., Foran, J., & Haeberli, P. (1992). Fast shadows and lighting effects using texture mapping. *Proceedings of the SIGGRAPH Conference on Computer Graphics and Interactive Techniques.*

Sekitoh, M. (2001). Bird's eye view system for ITS. In *Intelligent Vehicle Symposium.*

Sellen, A., Buxton, B., & Arnott, J. (1992). Using spatial cues to improve videoconferencing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 651–652).

Slovák, P., Troubil, P., & Holub, P. (2009). GColl group-to-group videoconferencing system: Design and first experiences. *Collaborative Computing: Networking, Applications and Worksharing* (pp. 1–9).

Tanger, R., Kauff, P., & Schreer, O. (2005). Immersive meeting point. *Advances in Multimedia Information Processing—PCM 2004*, 89–96.

Tanimoto, M. (2006). Overview of free viewpoint television. *Signal Processing: Image Communication*, *21*(6), 454–461.

Ten Koppel, M., Bailly, G., Müller, J., & Walter, R. (2012). Chained displays: Configurations of public displays can be used to influence actor-, audience-, and passer-by behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 317–326).

Troje, N., & Siebeck, U. (1998). Illumination-induced apparent shift in orientation of human heads. *PERCEPTION-LONDON*, *27*, 671–680.

Tsui, K., Desai, M., Yanco, H., & Uhlik, C. (2011). Exploring use cases for telepresence robots. *Proceedings of the 6th International Conference on Human Robot Interaction* (pp. 11–18).

Vertegaal, R., Weevers, I., Sohn, C., & Cheung, C. (2003). GAZE-2: Conveying eye contact in group video conferencing using eye-controlled camera direction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 521–528).

Wilburn, B. S., Smulski, M., Lee, H.-H. K., & Horowitz, M. A. (2001). Light field video camera. *Electronic Imaging* (pp. 29–36).

Wilson, H., Wilkinson, F., Lin, L., & Castillo, M. (2000). Perception of head orientation. *Vision Research*, *40*(5), 459–472.

Yendo, T., Fujii, T., Tanimoto, M., & Panahpour Tehrani, M. (2010). The Seelinder: Cylindrical 3D display viewable from 360 degrees. *Journal of Visual Communication and Image Representation*, *21* (5–6), 586–594.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330–1334.