TeleHuman: Effects of 3D Perspective on Gaze and Pose Estimation with a Life-size Cylindrical Telepresence Pod

Kibum Kim[1], John Bolton[1], Audrey Girouard[1,2], Jeremy Cooperstock[3] and Roel Vertegaal[1]

[1] Human Media Lab
Queen's University
Kingston, ON, K7L 3N6
Canada
{*kibum, bolton, roel*}*@cs.queensu.ca*

[2] School of Information Technology
Carleton University
Ottawa, ON, K1S 5B6
Canada
audrey_girouard@carleton.ca

[3] Centre for Intelligent Machines
McGill University
Montreal, QC, H3A 2A7
Canada
jer@cim.mcgill.ca

ABSTRACT

In this paper, we present TeleHuman, a cylindrical 3D display portal for life-size human telepresence. The TeleHuman 3D videoconferencing system supports 360º motion parallax as the viewer moves around the cylinder and optionally, stereoscopic 3D display of the remote person. We evaluated the effect of perspective cues on the conveyance of nonverbal cues in two experiments using a one-way telecommunication version of the system. The first experiment focused on how well the system preserves gaze and hand pointing cues. The second experiment evaluated how well the system conveys 3D body postural information. We compared 3 perspective conditions: a conventional 2D view, a 2D view with 360º motion parallax, and a stereoscopic view with 360º motion parallax. Results suggest the combined presence of motion parallax and stereoscopic cues significantly improved the accuracy with which participants were able to assess gaze and hand pointing cues, and to instruct others on 3D body poses. The inclusion of motion parallax and stereoscopic cues also led to significant increases in the sense of social presence and telepresence reported by participants.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Design, Human Factors, Teleconference.

Keywords: Telepresence, cylindrical display, organic user interfaces, 3D video, videoconference, motion parallax.

INTRODUCTION

Current videoconferencing systems range from the popular, low-end, small displays of Skype and FaceTime to expensive, large-screen business systems such as Cisco TelePresence and Polycom RealPresence, the latter of which can support life-size display. However, all of these systems suffer limitations in their ability to support important nonverbal communication cues such as eye contact, 3D spatial reasoning, and movement of interlocutors. The effect of these cues on remote communication may be difficult to measure, and may not affect typical parameters, such as task performance [33]. However, we believe that differences in user experience of telecommunication versus face-to-face communication may be attributed to subtle violations of such nonverbal communication [31].

Since the Talking Heads system [20], researchers have worked on preserving cues in telecommunication to enhance human telepresence [3]. However, very few systems approach the richness of direct face-to-face communication. Most only preserve a partial set of visual cues or suffer from costly and complex implementations [9]. One approach has been the use of animated 3D avatars of users [8] and head-mounted 3D virtual reality systems [34]. In such systems, a 3D model of the user is produced once, then animated in real time by *measuring* the user's behavior. Since only animation parameters are transmitted in real time, these systems typically require little bandwidth. However, they do so at a cost in realism that results in an Uncanny Valley effect [19].

While recent advances in 3D avatar systems offer highly realistic renditions [1], we believe there are significant advantages to using 3D video instead. Video-based systems differ from avatar systems in that they capture a realistic 3D video model of the user every frame, which is then broadcast and rendered in real time across the network [9]. This results in a highly realistic replication of behavioral cues, but at a cost of network bandwidth. The capturing and transmission of 3D video has, to date, required many special considerations in terms of camera placement and projection environment [9]. The associated requirements of such environments are prohibitive for the typical workplace.

TeleHuman

These observations motivated our development of TeleHuman, a 3D video-based conferencing system that provides the capabilities of 3D capture, transmission, and display in a lightweight, low-cost, low-bandwidth configuration. The system relies on 10 low-cost Microsoft

CHI'12, May 5–10, 2012, Austin, Texas, USA.

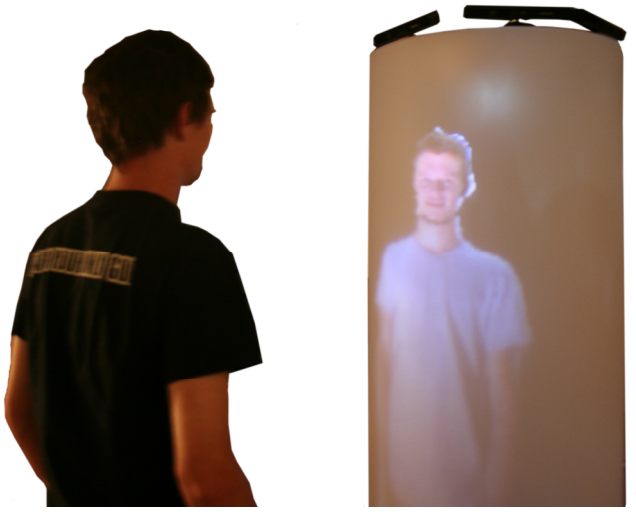

Figure 1. The TeleHuman system: local user (left) interacting with remote interlocutor (right) in 3D.

Kinects for capturing 360º 3D video models of the users. 3D models are efficiently broadcast over the network by adding a grayscale depth map frame to each frame of video. 3D video images are then synthesized locally through texture mapping, in response to actual viewer perspective. The 3D video models are rendered with perspective correction and stereoscopy on a life-sized cylindrical display, using an off-the-shelf 3D projector (see Figure 1).

Contribution

The chief contribution of TeleHuman is that it provides 360º motion parallax with stereoscopic live-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. There is evidence to suggest that motion parallax and stereoscopy play an important role in the experience of telepresence [25]. To evaluate how these factors might aid in the preservation of basic body orientation cues used in deixis [36] and in pose estimation tasks, we conducted two experiments. The first focused on how well the system preserves gaze directional and hand pointing cues. The second experiment evaluated how well the system conveys 3D body postural cues. For both tasks, the TeleHuman was tested in three different viewing conditions: conventional 2D, 2D + motion parallax, and motion parallax + stereoscopy. Results show the presence of both motion parallax and stereoscopic cues significantly improved the accuracy with which participants were able to assess gaze and hand pointing cues, and instruct others on 3D body posture. These cues also led to significant increases in the sense of telepresence reported by participants.

BACKGROUND

We will first review work from early studies in virtual telepresence systems, after which we review work on gaze awareness in video conference systems. Finally, we will discuss the use of 3D in telepresence systems, and review work on motion parallax.

Telepresence Systems

Research initiatives in electronic transmission of human telepresence trace back to as early as the late 1940s with Rosenthal's work on half-silvered mirrors to transmit eye contact during video broadcasts [30]. In the 1970s, Negroponte developed the Talking Heads project [23]. Driven by the US government's emergency procedures prohibiting the co-location of its highest-ranking five members, Talking Heads proposed a five-site system where each site was composed of one real person and four plastic heads mounted on gimbals that replicated user head orientation. Properly registered video was projected inside a life-size translucent mask in the exact shape of the face, making the physical mask appear animated with live images. However, the system was a mockup that, in practice, would have required head mounted cameras for appropriate registration of faces.

The BiReality system [12] 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 360º surround view of the remote location and the surrogate's head displayed a live video of the user's head from four sides. By providing a 360º surround environment for both locations, the user could perform all rotations locally by rotating his or her body. This preserved gaze and eye contact at the remote location. Although this system presented a life size tele-operated robotic surrogate, only the remote user's head image was rendered realistically. As implemented, the BiReality display was not responsive to viewer position, and thus, did not support motion parallax.

Gaze Direction

A lightweight approach to preserving gaze directional cues was provided by Hydra [31]. Hydra used multiple cameras, monitors, and speakers to support multiparty videoconferencing. It simulated a four-way round-table meeting by placing a camera, monitor, and speaker at the position of each remote participant, preserving both head orientation and eye contact cues. Although initial prototypes suffered from vertical parallax due to the spatial separation of the camera below the monitor, subsequent designs reduced this considerably by placing the camera directly above the display. Another limitation of Hydra was the use of small screens, which limited the size of remote participants. The size of the rendered interlocutor may indeed affect the sense of the social presence [4]. The MAJIC [26] and Videowhiteboard systems [32] projected life size images on semi-transparent surfaces by placing cameras behind the screen. However, these systems did not support 3D stereoscopic cues or motion parallax. The GAZE [33,36] groupware system provided integral support for conveying eye gaze cues using still images. Instead of using multiple video streams, GAZE measured where each participant looked by means of a desk-mounted eye-

tracking system. This technique presented a user with the unique view of each remote participant, emanating from a distinct location in space. Each persona rotated around its x and y axes in 3D space, thus simulating head movements. Later, motion video was added via the use of half-silvered mirrors in GAZE-2 [35].

3D Motion Parallax and Stereoscopy

A variety of technical solutions have been devised to explore the preservation of 3D depth cues and motion parallax. Harrison and Hudson presented a method for producing a simple pseudo-3D experience by providing motion parallax cues via head position tracking [10]. Their system required only a single traditional webcam at each end for both scene capture and the creation of head-coupled pseudo-3D views. This system utilized a 2D display that did not provide stereoscopic vision [39]. Some CAVE-like environments provide an immersive VR experience, providing motion parallax for a single user. They typically also require the use of shutter glasses, thus precluding the possibility of eye contact transmission. For example, Blue-C, an immersive projection and communication system [9,22], combines real-time 3D video capture and rendering from multiple cameras. Developing a novel combination of projection and acquisition hardware, it created photorealistic 3D video inlays of the user in real time [22]. The use of auto-stereoscopic display technologies [15,17,24] provides similar capabilities, but without the need for special eyewear and often, adding the ability to support multiple users simultaneously, each with their own perspective-correct view. However, these are restricted to specific optimal viewing zones, may result in significantly reduced resolution, and rely on a flat form factor.

We should note that the above examples all rely on planar screens, limiting the ability of users to walk around the display of a remote interlocutor as is, e.g., possible with LiteFast displays [16]. Another technology, swept-surface volumetric display [11], supports 3D display with motion parallax in a form factor often more suitable for this purpose, but recent examples have been too small to render a full human body at life size.

Empirical Work

Although the benefits of including motion parallax and stereoscopy in the presentation of graphic interfaces have been demonstrated [37], systematic evaluation of the impact of these factors in the context of task performance during video communication, specifically, in assessing pointing or poses of a remote interlocutor, is sparse. Böcker, Rundel and Mühlbach [6] compared videoconferencing systems that provide motion parallax and stereoscopic displays. While their results suggested some evidence for increased spatial presence and greater exploration of the scene, the studies did not evaluate effects on task performance. Subsequently, the provision of motion parallax was shown to generate larger head movements in users of video conferencing systems, suggesting that users do utilize such cues [5].

DESIGN RATIONALE

Our main consideration in the design of our capture and display system was to support 3D cues. These aid in the preservation of information related to head orientation pose, gaze, and overall body posture of a human interlocutor. In this context, we identified a number of relevant design attributes:

3D Cues – TeleHuman supports 3D both through optional use of stereoscopic shutter glasses and motion parallax. The latter results in a change of view and relative shifts of objects in the visual field due to changes in the observer's tracked position, allowing users to walk around and observe a virtually projected interlocutor from any angle.

Form Factor – Providing full 360º motion parallax required the use of a cylindrical form factor display [16] proportionate to the human body. Since this offers an unobstructed 360º field of view, it enables a user to explore different perspectives by natural physical movement.

Directional Cues – Being able to determine where users are looking or pointing has been shown to be an important cue in videoconferencing [34]. These cues can help regulate conversation flow, provide feedback for understanding, and improve deixis [13,20]. The use of 3D video models, as opposed to the direct display of a single 2D video camera output, facilitates preservation of eye contact. However, stereoscopy through shutter glasses inhibits estimation of eye orientation in bi-directional scenarios. We believed that motion parallax alone may suffice for estimation of gaze or pointing direction, as users are free to move to the location in which gaze and arm orientations align to point at the user [5].

Size – Prior work, such as Ultra-Videoconferencing [7] and that of Böcker et al. [4], suggests that to avoid misperceptions of social distance [2] and to aid in a sense of realism, preservation of body size is important [25]. This motivated the conveyance of life-size images in our design.

TELEHUMAN IMPLEMENTATION

Our implementation of TeleHuman revolved around the design of a cylindrical display coupled with 3D tracking and imaging. We first discuss the imaging hardware, after which we discuss software algorithms for capturing, relaying, and displaying live 3D video images.

TeleHuman Cylindrical 3D Display

Figure 2 shows the cylindrical display deployed in TeleHuman. The display consists of a 170 cm tall hollow cylinder with a diameter of 75 cm made of 6.3 mm thick acrylic. The cylinder was sandblasted inside and out to create a diffuse projection surface. The cylinder is mounted on top of a wooden base that holds the projector, giving the entire system a height of approximately 200 cm. These dimensions were chosen to allow for a range in size of remote participants. A DepthQ stereoscopic projector [14] is mounted at the bottom of each display, pointed upwards to reflect off a 46 cm hemispherical convex acrylic mirror.

Figure 2. TeleHuman hardware: a cylindrical display surface with 6 Kinects and a 3D projector inside its base.

This allows projections of images across the entire surface of the cylinder. The DepthQ projector has a resolution of 1280 *x* 720 pixels. However, since only a circular portion of this image can be displayed on the surface of the cylinder, the effective resolution is described by a 720 pixel diameter circle, or 407,150 pixels.

An Nvidia 3D Vision Kit [21] is used with the projector to create an active stereoscopic display. This kit provides an IR emitter that connects to a 3-pin sync port on our system's graphics card. Compatible shutter glasses are synced with the IR emitter and projected image, refreshing at 120 Hz. As a result, when viewing the display, a distinct image is shown to each eye, and disparity between these two images creates stereoscopy. By combining depth cues with perspective corrected motion parallax [37] the remote participant appears to be standing inside the cylinder.

User Tracking

We used Microsoft Kinect depth-sensitive cameras [18] to determine the location of users around the cylinder. Six Kinects are mounted on the top of the cylinder, pointed downwards (see Figure 2). These track the location of the user around the cylinder, and obtain frontal images. Four Kinects are located in a square around the cylinder, centered at approximately 2.5 m from its center. These obtain images from the side and back of the user. Images from the Kinects are accessed using OpenNI [29] drivers. Each camera provides a 640*x*480 pixel stream at 30 fps with both RGB and depth images. When a user approaches to within 2 m of the TeleHuman, the system starts tracking and broadcasting. The system tracks the location of users around the display until they step out of range. Each Kinect is connected to a PC, which sends the user's position via Open Sound Control [38], along with the user's RGB image and depth map to a Microsoft XNA application that controls the projection. The XNA application calculates the angle between the user and the cylinder and updates the displayed model accordingly. To maintain an appropriate frame rate, we use 1 PC per 2 Kinects, using a total of 5 PCs for preprocessing image data.

Live 3D Model Generation

In order to create a 3D representation of a user, depth values are used to position vertices in a 3D XNA application. Using the depth and RGB streams, the system calculates a four-channel image via OpenCV [28]. This image contains RGB information in the first three channels and depth information in the fourth channel. Images are then sent via a TCP connection with the XNA projection application running on a separate machine. Currently, our system sends images over a gigabit LAN connection, relying on the associated high network speeds to provide multiple live streams with low latency. Note that future versions will use more efficient UDP protocols.

Using the depth map, the XNA display application creates vertices corresponding to each pixel of the user. The depth value is used to determine the vertex locations along the *z* axis. Depth values are also used to remove the scene behind the user, via a basic depth threshold. Vertices are placed in a vertex buffer. The content of this buffer is read and rendered by the XNA application. Based on the distance of the viewer from the cylindrical display, the model is rendered such that the center of mass of the TeleHuman appears to be in the middle of the cylinder, which we treat as the origin. The RGB values from the input image are used to texturemap the resulting mesh model.

Motion Parallax and Projection Distortion

The view of a user on the cylinder is rendered from the perspective of a virtual camera targeted at his or her 3D model. The angular position of the user controls the angle with which this virtual camera looks at the 3D model of the interlocutor. As a user's position changes, the position of the camera changes accordingly, allowing him or her to view a motion parallax corrected perspective of the 3D video model of the other user. This camera view is rendered and stored as a texture. 3D information is preserved during this process allowing the texture to be viewed with stereoscopy. The projected image is rendered using Microsoft's XNA 4.0 framework. A custom distortion class was developed, creating a two-dimensional semi-circular object. The texture coordinates of this object

Figure 3. Textured 3D model with hemispherical distortion. When reflected off the convex mirror onto the cylinder, this produces a 3D model with proper proportions.

are modified to account for the distortions introduced by the hemispherical mirror and the cylindrical display surface. The distortion model is textured using the previously rendered camera view (Figure 3). When reflected off the hemispherical convex mirror, this creates an undistorted projection of the remote participant on the surface of the cylinder. When the user moves around the display, the distortion model ensures that the remote participant remains at the center of the user's field of view. As this projection changes based on user position, it creates a cylindrical Fish Tank VR view that preserves motion parallax [37]. Note that our approach does have the side effects of causing both resolution and brightness to drop off at lower elevations of the cylinder.

EMPIRICAL EVALUATION

We designed two experiments to evaluate effects of stereoscopy and 360º motion parallax on the preservation of nonverbal cues in our TeleHuman system. Our first experiment focused on how stereoscopy and motion parallax might aid in the preservation of basic body orientational cues. The second experiment focused on how stereoscopy and 360º motion parallax around the display might aid in conveying body postural cues.

EXPERIMENT 1: EFFECTS OF 3D PERSPECTIVE ON GAZE AND POINTING DIRECTION ESTIMATES

The first experiment was designed to gauge the effects of motion parallax and stereoscopy on judgment of eye gaze and hand pointing by a TeleHuman 3D video model.

Task

Participants were asked to indicate where a TeleHuman model was looking or pointing. To ensure equal conditions for all participants, we used a static prerecorded TeleHuman 3D video model in all conditions. We used a simplified, asymmetrical setup in which only one TeleHuman pod was used. At each position, participants were first asked if the TeleHuman was pointing or looking directly at them. If they answered negatively, they were asked to indicate where the TeleHuman was pointing or looking, with reference to a tape measure mounted on a

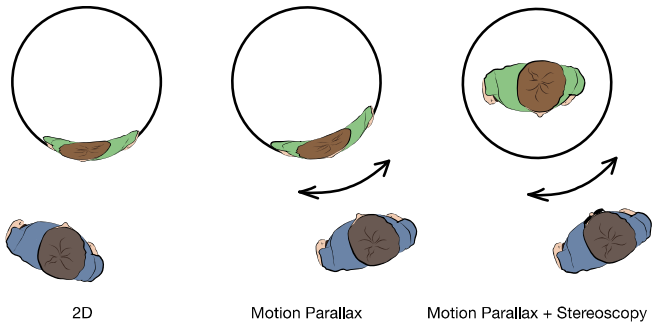

Figure 4. Top-view drawing of perspective conditions: conventional 2D (left), motion parallax (middle), motion parallax + stereoscopy (with glasses, right). In the case of motion parallax, the display would show the remote individual from a slightly side perspective. The black circle represents the cylinder, the person with a green shirt is the perception of the remote participant. The local user is wearing a blue shirt.

wall behind them. Next, participants were asked to move parallel to the wall until they were satisfied that the remote participant was looking or pointing straight at them, at which point we recorded their position.

Experiment Design

We used a within-subjects design in which we evaluated the effect of two fully factorial independent variables: *perspective* and *pointing cue.* To allow for a more realistic scenario, and a richer set of cues, we also varied the participant's location in front of the display: left, center, and right, and the TeleHuman's pointing angle: left, center and right, between conditions.

Perspective

The perspective factor consisted of three levels: *conventional 2D, motion parallax, motion parallax + stereoscopy* (see Figure 4). For the conventional condition, the TeleHuman was shown from the perspective of a front-facing camera, centered on the human. In the *motion parallax* condition, the TeleHuman was displayed with continuous perspective correction based on the location of the participant relative to the display. In the *motion parallax + stereoscopy* condition, participants additionally wore shutter glasses that provided them with a fully stereoscopic image of the TeleHuman, giving the impression that the human was *inside* the cylinder.

Pointing Cue

The pointing cue factor had three levels: *gaze, hand,* and *gaze + hand*. In the *gaze* condition, the TeleHuman indicated the pointing direction by both eye gaze and head orientation directed towards the same location on the wall. In the *hand* condition, the TeleHuman pointed at the target with their arm, hand and index finger. In this condition, the gaze of the TeleHuman was fixated directly to the center, unless the actual target was the center, in which case, gaze was oriented randomly to the left or right of the target. In the *gaze + hand* condition, the TeleHuman's arm, hand and index finger all pointed in the same direction as the eyes and head.

Setup and Procedure

Figure 4 shows a participant standing in front of the TeleHuman. The display was placed 2 m from a wall behind the participant. This wall showed a tape measure with markings at 5 cm intervals from left to right. To ensure presentation of consistent stimuli to all participants, we used a recorded still 3D image to constitute the pointing cues factor. These were rendered according to the perspective factor, as shown in Figure 4. For each condition, participants were asked to stand in between the display and a wall behind them, approximately 190 cm from the display and 10 cm from the wall. Participants experienced the perspective and pointing cue conditions from three locations, distributed between-conditions: directly in front of the cylindrical display, 45 cm to its left, and 45 cm to its right. In addition, in each condition, the TeleHuman pointed in a different angle, selected from left, center, or right. Note that while pointing targets were not visible within our display setup, targets could be projected in the environment in a real videoconferencing scenario.

Trials

Each participant carried out a total of 9 trials, by factorial combination of 3 perspectives (*2D*, *motion parallax*, *motion parallax + stereoscopy*) with 3 pointing cues (*gaze, hand, gaze+hand*). To allow for a richer set of cues, we also varied the locations of the participant (3 locations) and the directions of pointing between conditions (3 directions). We did not perform a fully factorial presentation as it would have led to 81 trials per participant. The order of presentation of conditions was counterbalanced using a Latin square. All participants were presented with the same set of stimuli, in different orders. The experimental session lasted one hour.

Participants

We recruited 14 participants (mean of 21 years old, 7 male), who were paid $15 for their participation. Three of the participants wore corrective glasses.

Measures

We determined the mean accuracy of pointing location through two measures: 1) *visual assessment*, where participants judged where the TeleHuman was pointing without moving from their initial location; and 2) *visual alignment*, where participants moved to the location at which the TeleHuman appeared to be pointing right at them. Visual assessment allowed us to determine any effects of a more stationary perspective on the accuracy of pointing direction estimates. We expected visual alignment to provide the most accurate method for determining where the TeleHuman pointed or looked, as it allowed users to align themselves such that the TeleHuman appeared to be looking or pointing directly at them. Each measure was calculated as the angular difference between reported viewing direction and the actual TeleHuman pointing direction.

Perspective	2D	Motion Parallax	Motion Parallax + Stereoscopy
Visual Assessment	15.3° (1.6)*	11.5° (1.5)	8.4° (1.2)*
Visual Alignment	21.6° (1.9)$^{\dagger, \ddagger}$	5.2° (.89)†	3.9° (.43)‡

Table 1. Angular mean difference between actual and reported target locations and standard error (s.e.) in degrees. There were significant differences, $^{*}p = 0.009$, $^{\dagger}p < 0.001$ and $^{\ddagger}p < 0.001$.

Questionnaire

To evaluate the degree of telepresence and social presence experienced, participants completed a seven-point Likert scale questionnaire after each *perspective* condition [25]. Telepresence was defined as the feeling of "being there", while social presence was defined as the perceived ability to connect with people through the medium. In the questionnaire, a 1 corresponded to *strongly agree* and 7 to *strongly disagree*.

Results

All results were analyzed using a within-subjects analysis of variance (ANOVA), evaluated at an alpha level of .05.

Pointing Location Estimation

Table 1 shows the accuracy of *pointing location* estimates for our two measures: visual assessment and visual alignment.

Visual Assessment

Results for visual assessment of pointing direction show a significant main effect of *perspective* on accuracy ($F(2,26)=6.35$, $p=0.006$), but no significant effect for *pointing cues* ($F(2,26)=1.92$, $p=0.17$). Bonferroni post-hoc tests showed that mean accuracy of visual assessment was 1.8 times higher in the *motion parallax + stereoscopy* condition than in the *conventional 2D* condition ($p=0.009$). However, there were no significant differences between other conditions.

Visual Alignment

Results for visual alignment show a significant main effect for *perspective* ($F(2,26)=66.51$, $p<0.001$), but not for *pointing cues* ($F(2,26)=0.88$, $p=0.425$). Post-hoc pairwise Bonferroni corrected comparisons of the *perspective* conditions show that mean accuracy was significantly greater in the *motion parallax* condition ($p<0.001$) and in the *motion parallax + stereoscopy* condition ($p<0.001$), compared to the *conventional 2D* condition. There was no significant difference between the *motion parallax* and *motion parallax + stereoscopy* conditions ($p=0.71$).

Questionnaire

Table 2 summarizes the answers to each question for each of the three perspective conditions presented. A Friedman test indicated that there were significant differences between perspective conditions in S1 "*It was as if I was facing the partner in the same room*" ($\chi^2(2)=6.69$,

Statements	2D	Motion Parallax	Motion Parallax + Stereoscopy
It was as if I was facing the partner in the same room. (S1)[*]	4.21 (2.0)	3.21 (1.8)	3.14 (2.0)
My partner seemed a real person. (S2)[†]	4.43 (2.3)	3.86 (2.0)	3.36 (2.2)
I felt immersed in the environment. (T1)[‡]	4.07 (1.9)	3.14 (2.1)	2.64 (1.8)
I felt surrounded by the environment. (T2)[+]	4.00 (2.1)	3.21 (1.9)	2.50 (1.4)

Table 2. Means and standard errors (s.e.) for social presence (S) and telepresence (T) scores. Lower scores indicate stronger agreement. There were significant differences between perspective conditions, [*]p = 0.035, [†]p = 0.011, [‡]p < 0.001 and [+]p < 0.001.

p=0.035), S2 "*My partner seemed a real person*" ($\chi^2(2)=9.05$, *p=0.011*), T1 "*I felt immersed in the environment*" ($\chi^2(2)=15.37$, *p<0.001*) and T2 "*I felt surrounded by the environment*" ($\chi^2(2)=16.06$, *p<0.001*).

Wilcoxon Signed-Rank post-hoc analysis for social presence showed significant differences in rankings between the *motion parallax* and *conventional 2D* perspectives (Z=-2.22, *p=0.026* for S1, Z=-1.99, *p=0.046* for S2) and between the *motion parallax + stereoscopy* and *conventional 2D* perspectives (Z=-2.70, *p=0.007* for S1, Z=-2.41, *p=0.016* for S2). However, we found no significant differences between the *motion parallax* and the *motion parallax + stereoscopy* conditions.

For the degree of *telepresence*, there was a significant difference between the *motion parallax* and *conventional 2D* perspectives (Z=-2.32, *p=0.020* for T1, Z=-2.37, *p=0.018* for T2), and between the *motion parallax + stereoscopy* condition (Z=-2.65, *p=0.008* for T1, Z=-2.99, *p=0.003* for T2) and the *conventional 2D* condition. However, there were no significant differences between *motion parallax* and *motion parallax + stereoscopy* conditions.

EXPERIMENT 2: EFFECTS OF PERSPECTIVE CUES ON COMMUNICATION OF 3D BODY POSTURAL CUES

In the second experiment, we examined whether support for a 360° life-size stereoscopic view with motion parallax improved the ability to convey the body pose of a remote person on the TeleHuman.

Task

A remote instructor, displayed on the TeleHuman, first positioned herself in one of the predetermined yoga poses (see Figure 5), one per condition. The remote instructor was blind to the conditions. At that point, the main participant ("coach") instructed a co-located partner ("poser") to reproduce the pose as accurately as possible, within a 3 minute time limit. The reason for using a poser, rather than having the coach assume the pose him or herself

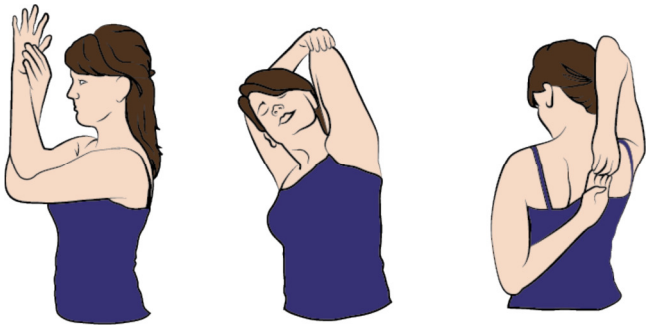

Figure 5. Sample Yoga stances used in Experiment 2.

is that this allowed the coach to walk freely around the display, as well as around the poser. Participants were asked to walk around the TeleHuman to examine the pose, and around the poser to examine the result, in all conditions. Note that while participants were allowed to ask the instructor to rotate to show her back conventional 2D conditions, none did, as this would have interfered with her ability to perform the pose.

Experiment Design

We used a within-subject experiment design to evaluate the effects of the *perspective* factor only, as per the first experiment (see Figure 4).

Setup and Procedure

The coach and the poser were co-located in the same room as the TeleHuman system; but only the coach could see the TeleHuman system. The instructor was in a separate room, and displayed using a live 3D 360º video model on the TeleHuman system. We used an asymmetrical version of the system that allowed for full 360º motion parallax, in which the coach could see and hear the instructor as represented by the TeleHuman, but the instructor could not see the coach. The instructor was not allowed to interfere with the directions of the coach to the poser. Once the coach was satisfied with the poser's posture, the instructor would go to the poser's room to evaluate the poser's stance, while the coach filled out a questionnaire.

We used pairs of participants, unfamiliar with yoga, alternating as coach and poser. To alleviate learning effects, a different yoga pose was used for every condition between pairs of participants, for a total of six yoga poses. All yoga poses, preselected by the yoga instructor, were of the same intermediate level of difficulty as judged by the instructor, and focused on upper body positioning (Figure 5). All poses had limb elements positioned on the back, front and sides of the instructor. The choice of yoga pose was randomly assigned to each coach and condition, and no feedback was provided by the instructor to the poser about the quality of any poses. The three visual perspective conditions were counter-balanced for each coach. The poser was never instructed on the perspective level at hand.

Participants

Eleven of the fourteen participants from the first experiment took part in the second experiment (mean of 22

Perspective	2D	Motion Parallax	Motion Parallax + Stereoscopy
Similarity Score	4.5 (0.71)*	5.5 (0.79)	7.1 (0.55)†

Table 3. Mean pose similarity score and standard error (s.e.) on a scale from 0 to 10 by yoga instructor, per condition. There were significant differences, *p = 0.03 and †p = 0.04.

years old, 7 male). They were paid a further $15 for their participation.

Measures

The instructor evaluated the similarity between her pose and that of the poser on a scale from 0 to 10 (10 meaning perfectly identical). In this process, she took into account limb angles and orientations, as well as overall posture. After each condition, coaches completed the same questionnaire administered in the first experiment, which evaluated the degree of telepresence and social presence experienced.

Results

We used a within-subjects ANOVA to evaluate differences between conditions, at an alpha level of .05.

Posture Similarity Scores

Table 3 shows the mean pose similarity score and standard error for each perspective condition. Results show that posture similarity scores were significantly different between perspective conditions ($F(2,20)=4.224$, $p=0.03$). Post-hoc tests using Bonferroni correction show that scores in the *motion parallax + stereoscopy* condition were significantly different from scores in the *conventional 2D* condition ($p=0.04$).

Questionnaire

Table 4 summarizes the mean scores for each question, per perspective condition. A Friedman test indicated that there were significant differences between perspective conditions for all social presence ratings (S1, same room $\chi^2(2)=16.06$, $p=0.001$), (S2, real person $\chi^2(2)=12.87$, $p=0.002$), and (S3 acquaintance $\chi^2(2)=11.29$, $p=0.004$). Differences between perspective conditions were also significant for all telepresence ratings (T1, immersion $\chi^2(2)=8.63$, $p=0.013$), (T2 surrounding $\chi^2(2)=12.65$, $p=0.002$), and (T3, involvement $\chi^2(2)=14.4$, $p=0.001$).

Wilcoxon Signed-Rank post-hoc analysis for social presence and telepresence ratings showed significant differences in rankings between the *motion parallax* and *conventional 2D* conditions (Z=-2.83, $p=0.005$ for S1, Z=-2.54, $p=0.011$ for S2, Z=-2.55, $p=0.011$ for S3, Z=-2.85, $p=0.004$ for T1, Z=-2.54, $p=0.011$ for T2, Z=-2.55, $p=0.011$ for T3) and between the *motion parallax + stereoscopy* and *conventional 2D* conditions (Z=-2.69, $p=0.007$ for S1, Z=-2.55, $p=0.011$ for S2, Z=-2.36, $p=0.018$ for S3, Z=-2.54, $p=0.011$ for T1, Z=-2.06, $p=0.040$ for T2, Z=-2.56, $p=0.011$ for T3). However, there were no significant differences between *motion parallax* and the *motion parallax + stereoscopy* conditions.

Statements	2D	Motion Parallax	Motion Parallax + Stereoscopy
*It was as if I was facing the partner in the same room. (S1)**	4.82 (1.1)	2.91 (1.1)	3.00 (1.3)
My partner seemed a real person. (S2)†	4.36 (1.5)	2.82 (0.9)	2.82 (1.0)
I could get to know someone that I only met through this system. (S3)‡	4.55 (1.4)	3.18 (1.2)	3.45 (1.0)
I felt immersed in the environment. (T1)+	4.45 (1.8)	2.82 (1.6)	3.09 (1.4)
I felt surrounded by the environment. (T2)†	5.18 (1.5)	3.55 (1.6)	3.45 (1.4)
*The experience was involving. (T3)**	3.64 (1.4)	2.00 (0.6)	2.27 (0.8)

Table 4. Mean agreement and standard errors (s.e.) with social presence and telepresence statements. Lower scores indicate stronger agreement. There were significant differences between perspective conditions, *p = 0.001, †p = 0.002, ‡p = 0.004, and +p = 0.013.

DISCUSSION

We now present a discussion of results from our two experiments.

Effects of 3D Perspective on Pointing Cue Assessment

Results from our first experiment confirmed a strong effect of perspective on the accuracy of assessment of remote pointing cues. Motion parallax + stereoscopy increased the accuracy of angular judgment by a factor of 1.8 over traditional 2D conditions in cases where participants were stationary. As expected, motion parallax alone, in this situation, was limited, and thus, the addition of stereoscopy was important. When participants were allowed to move, motion parallax was shown to provide the dominant effect, with participants achieving four times higher accuracy on average in angular judgment of remote pointing cues as compared to 2D conditions. In this case, stereoscopy appeared to provide little additional benefit. Note that the type of pointing cue: *gaze, hand only,* or *gaze + hand*, had no significant effect on accuracy measures.

Qualitative measures support the above analysis. Social presence rankings were significantly higher in conditions where motion parallax cues were supported, with no significant additional effect for motion parallax augmented by stereoscopy. As for the degree of telepresence or immersion, the combined effect of motion parallax and stereoscopy was critical for obtaining significant differences from 2D conditions.

Stereoscopy therefore appears to be beneficial for judgment of pointing angle when motion parallax cannot be exploited. However, this comes at the cost of preventing

reciprocal gaze awareness if shutter glasses are deployed. Motion parallax, even in the absence of a stereoscopic display, may, however, suffice for preservation of social presence or pointing cues.

Effects of 3D Perspective on Body Pose Assessment

Results for our second experiment, in which we evaluated the effects of perspective cues on preservation of postural cues, were in line with those from Experiment 1. The presence of *motion parallax + stereoscopy* cues increased the accuracy of pose scores by a factor of 1.6 over conventional 2D conditions. These results suggest that both motion parallax and stereoscopy needed to be present in order to judge and convey poses accurately. Surprisingly, the presence of motion parallax cues alone only marginally improved scores. This was likely due to the fact that while motion parallax allowed users to see the sides and back of poses, stereoscopy helped improve their judgment of the relative angles of the limbs.

Qualitative measures indicate little additional effect of the presence of stereoscopic cues. Social presence rankings were significantly higher in conditions where *motion parallax* or *motion parallax + stereoscopy* were supported. As for the degree of telepresence, rankings were significantly higher in cases where *motion parallax* or *motion parallax + stereoscopy* were supported. However, there appeared to be little additional effect of the presence of stereoscopic cues over *motion parallax* only. While the presence of stereoscopy did not significantly affect qualitative measures, we can conclude that in this task both motion parallax and stereoscopy were required.

LIMITATIONS, APPLICATIONS & FUTURE DIRECTIONS

Our first study was limited by the fact that the TeleHuman was a static 3D image, and communication was not reciprocal. Although this permitted us to evaluate the effect of stereoscopy on pointing cue assessment, it necessitated an artificial communication condition in which the shutter glasses had no detrimental effect on perception of eye contact. There is an obvious tradeoff between supporting eye contact between interlocutors and presentation of a stereoscopic display requiring the use of shutter glasses. However, other display technologies, such as autostereoscopic and volumetric displays do support glasses-free stereo viewing. We hope to conduct future experiments to evaluate the added benefit that such technologies might offer in terms of eye contact perception with TeleHuman. Note that participants in our study did not ask the instructor to rotate in the 2D condition. There may be cases in which such rotation would provide adequate information to complete a 3D pose task. To avoid introducing confounding factors, we did not specifically compare results with traditional 2D flat display conditions. However, we believe that the results of our 2D conditions would generalize to such conditions.

Future Application Scenarios

The TeleHuman system has potential applications in a number of areas where 2D displays may limit the users' viewpoints. One example is in remote sports instruction. As Experiment 2 demonstrates, examination of the mechanics of limb movement may benefit from the ability to review movement and posture from any angle. For example, this may be helpful in teaching golfers to improve their swing. Applications also exist in telemedicine and remote medical instruction, for which the benefits of arbitrary view control were demonstrated previously in the context of surgical training [27]. TeleHuman could similarly offer doctors the ability to examine remote patients from any angle, but at full scale. This may be particularly beneficial for orthopedic or postural conditions, where the patient cannot reorient herself for a side view. Finally, applications exist in gaming, as the ability to render a 3D gaming character or another online gamer in a 360º view allows for a more immersive gaming experience in first-person shooter scenarios.

Support for Multiparty Videoconferencing

In the near future, we hope to leverage TeleHuman for *multiparty* teleconferencing scenarios. To support such experimentation, we will be replacing the current TCP communication layer with a UDP-based alternative, suitable for low-latency interaction over larger distances. Support of a teleconference with n users requires n^2-n setups and, barring multicast support, a similar number of data streams. This entails significant bandwidth requirements for transmission of 3D video models. However, our design allows for such scaling without modifications to the TeleHuman hardware.

CONCLUSIONS

In this paper, we presented the TeleHuman system, a cylindrical display portal for life-size 3D human telepresence. The system transmits telepresence by conveying 3D video images of remote interlocutors in a way that preserves 360º motion parallax around the display, as well as stereoscopy. We empirically evaluated the effect of perspective on the user's accuracy in judging gaze, pointing direction, and body pose of a remote partner using an asymmetrical version of the system. Results for pointing directional cues suggest that the presence of stereoscopy is important in cases where the user remains relatively stationary. However, when users move their perspective significantly, motion parallax provides a dominant effect in improving the accuracy with which users were able to estimate the angle of pointing cues. As for pose estimation, the presence of both 360º motion parallax cues and stereoscopic cues appeared necessary to significantly increase accuracy. Both motion parallax and stereoscopy appear important in providing users with a sense of social presence and telepresence. We conclude that we recommend inclusion of both motion parallax and stereoscopic cues in video conferencing systems that support the kind of tasks used in our evaluation, with the caveat that tools such as shutter glasses, which obstruct views of the remote participants eyes, are most likely not recommendable for bi-directional communication systems.

ACKNOWLEDGEMENTS

This work was supported by grants by the Natural Sciences and Engineering Council of Canada (NSERC) and the Ontario Research Fund. We also thank Autodesk Research.

REFERENCES

1. Alexander, O., Lambeth, W., and Debevec, P. Creating a Photoreal Digital Actor : The Digital Emily Project Previous Efforts at Photoreal Digital Humans. *Proc. SIGGRAPH*, (2009), 1-15.
2. Argyle, M. and Ingham, R. Gaze, mutual gaze and proximity. *Semiotica 6*, 1 (1972), 32-49.
3. Buxton, B. Telepresence: integrating shared task and person spaces. *Proc. GI*, (1992), 123-129.
4. Böcker, M. and Mühlbach, L. Communicative Presence in Videocommunications. *Proc. Human Factors and Ergonomics Society*, (1993), 249-253.
5. Böcker, M., Blohm, W., and Mühlbach, L. Anthropometric data on horizontal head movements in videocommunications. *Proc. CHI*, (1996), 95-96.
6. Böcker, M., Rundel, D., and Mühlbach, L. On the Reproduction of Motion Parallax in Videocommunications. *Proc. HFES*, (1995), 198-20.
7. Cooperstock, J.R. Multimodal Telepresence Systems: Supporting demanding collaborative human activities. *IEEE Signal Processing Magazine, Special Issue on Immersive Communications 28*, 1 (2011), 77-86.
8. Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., and Sasse, M.A. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Proc. CHI*, 5 (2003), 529.
9. Gross, M., Würmlin, S., Naef, M., et al. blue-c: a spatially immersive display and 3D video portal for telepresence. *ACM Transactions on Graphics 22*, 3 (2003), 819-827.
10. Harrison, C. and Hudson, S. Pseudo-3D Video Conferencing with a Generic Webcam. *Proc. IEEE International Symposium on Multimedia*, (2008), 236-241.
11. Jones, A., Lang, M., Fyffe, G., et al. Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Transactions on Graphics 28*, 3 (2009), 64:3-64:8.
12. Jouppi, N.P., Iyer, S., Thomas, S., and Slayden, A. BiReality: mutually-immersive telepresence. *Proc. MM*, (2004), 860-867.
13. Kendon, A. Some functions of gaze direction in social interaction. *Acta Psychologica 26*, 44 (1967), 22-63.
14. Lightspeed Design. DepthQ Stereoscopic Projector. http://www.depthq.com.
15. Lincoln, P., Nashel, A., Ilie, A., Towles, H., Welch, G., and Fuchs, H. Multi-view lenticular display for group teleconferencing. *Proc IMMERSCOM*, (2009).
16. Litefast. . http://www.litefast-display.com/.
17. Matusik, W. and Pfister, H. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics*, (2004), 814–824.
18. Microsoft Kinect. . http://www.xbox.com/kinect.
19. Mori, M. The Uncanny Valley. *Energy 7*, 4 (1970), 33-35.
20. Morikawa, O. and Maesako, T. HyperMirror: toward pleasant-to-use video mediated communication system. *Proc. CSCW*, (1998), 149-158.
21. NVidia. nVidia 3D Vision Kit. http://www.nvidia.com/object/3d-vision-main.html.
22. Naef, M., Lamboray, E., Staadt, O., and Gross, M. The blue-c distributed scene graph. *Proc. IPT/EGVE Workshop*, (2003), 125-133.
23. Negroponte, N. *Being Digital*. Vintage Books, New York, NY, USA, 1995.
24. Nguyen, D. and Canny, J. MultiView: spatially faithful group video conferencing. *Proc. CHI*, (2005), 799–808.
25. Nowak, K.L. and Biocca, F. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence Teleoperators and Virtual Environments 12*, 5 (2003), 481-494.
26. Okada, K.-I., Maeda, F., Ichikawaa, Y., and Matsushita, Y. Multiparty videoconferencing at virtual social distance: MAJIC design. *Proc. CSCW*, (1994), 385-393.
27. Olmos, A., Lachapelle, K., and Cooperstock, J.R. Multiple angle viewer for remote medical training. *Proc. International Workshop on Multimedia Technologies for Distance Leaning*, (2010), 19–24.
28. OpenCV. . http://opencv.willowgarage.com/.
29. OpenNI. . http://openni.org/.
30. Rosenthal, A. Two-way Television Communication Unit. (1947).
31. Sellen, A., Buxton, B., and Arnott, J. Using spatial cues to improve videoconferencing. *Proc. CHI*, (1992), 651-652.
32. Tang, J.C. and Minneman, S. VideoWhiteboard: video shadows to support remote collaboration. *Proc. CHI*, (1991), 315-322.
33. Vertegaal, R. and Ding, Y. Explaining Effects of Eye Gaze on Mediated Group Conversations : Amount or Synchronization ? *Proc. CSCW*, (2002), 278-285.
34. Vertegaal, R., Slagter, R., Der Veer, G. Van, and Nijholt, A. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. *Proc. CHI*, (2001), 301-308.
35. Vertegaal, R., Weevers, I., Sohn, C., and Cheung, C. GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction. *Proc. CHI*, (2003), 521–528.
36. Vertegaal, R. The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. *Proc. CHI*, (1999), 294-301.
37. Ware, C., Arthur, K., and Booth, K.S. Fish tank virtual reality. *Proc. CHI*, (1993), 37-42.
38. Wright, M. and Freed, A. Open Sound Control: A New Protocol for Communicating with Sound Synthesizers. *Proc. International Computer Music Conference*, (1997), 101-104.
39. Zhang, C., Yin, Z., and Florencio, D. Improving depth perception with motion parallax and its application in teleconferencing. *Proc. IEEE International Workshop on Multimedia Signal Processing*, (2009), 1-6.