A Gaze-preserving Situated Multiview Telepresence System

Ye Pan, Anthony Steed
Department of Computer Science
University College London, UK
y.pan, a.steed@cs.ucl.ac.uk

ABSTRACT

Gaze, attention, and eye contact are important aspects of face to face communication, but some subtleties can be lost in videoconferencing because participants look at a single planar image of the remote user. We propose a low-cost cylindrical videoconferencing system that preserves gaze direction by providing perspective-correct images for multiple viewpoints around a conference table. We accomplish this by using an array of cameras to capture a remote person, and an array of projectors to present the camera images onto a cylindrical screen. The cylindrical screen reflects each image to a narrow viewing zone. The use of such a situated display allows participants to see the remote person from multiple viewing directions. We compare our system to three alternative display configurations. We demonstrate the effectiveness of our system by showing it allows multiple participants to simultaneously tell where the remote person is placing their gaze.

Author Keywords

Non-planar displays; camera arrays; gaze

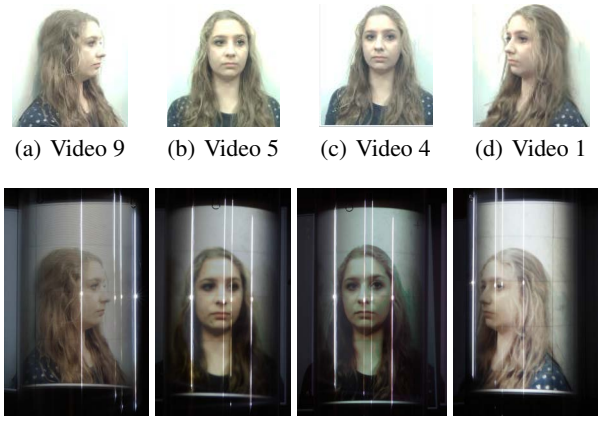
ACM Classification Keywords

H.4.3 Communications Applications: Computer conferencing, teleconferencing, and videoconferencing

INTRODUCTION

When a group of people communicate face to face, numerous cues of attention, eye contact, and gaze direction provide important additional channels of information, such as attention targets, conversational turn-taking indicators [1]. However, those non-verbal cues can be lost in traditional teleconferencing systems [9].

A variety of systems have been developed to support gaze awareness in group video conferencing, though the majority use a 2D planar display [4]. We propose to use a cylindrical display which provides the same angle of view from all directions. We further propose to use a camera array to surround the remote person horizontally, capturing unique and perspective-correct videos for each potential observer's viewing direction (see Figure 1(a) to Figure 1(d)). A projector array is arranged in the same manner as the camera array around the cylindrical display, which allows each observer to see different views simultaneously (see Figure 1(e) to Figure 1(h)).



(a) Video 9 (b) Video 5 (c) Video 4 (d) Video 1

(e) View 9 (f) View 5 (g) View 4 (h) View 1

Figure 1. The remote person is gazing at target 5. The top row is four videos simultaneously captured from four different cameras. The bottom row is four photos of the same display from four different perspectives. See Figure 2 for camera, target and viewpoint numbers.

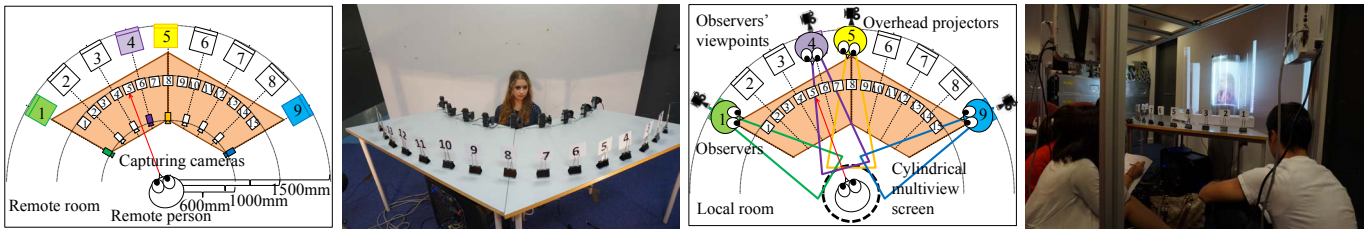
We evaluate the effectiveness of our system by measuring the ability of observers to accurately judge which target the remote person is gazing at. We run an experiment to demonstrate that our system can convey gaze relatively accurately, especially for observers viewing from off-center angles. This demonstration and results thus motivate the further study of novel display configurations and the supporting camera and networking infrastructure for them.

BACKGROUND

Many systems have achieved accurate reproduction of gaze direction in group video conferencing, including MAJIC [5], Hydra [10], GAZE-2 [11], Animatronic shader lamps avatars [3] and 3-d live [8]. These systems support correct gaze direction when used with one participant per site. Several current multiview display systems use a single display and a filter method or a lenticular separation method to produce different views. These methods divide the resolution of a display among multiple views so that each view has only N/K pixels, where N is the pixels of the full display and K is the number of views. MultiView [4] supports K full-resolution views. However, those planar displays are visible from the front only. Current situated displays, such as, TeleHuman [2] and SphereAvatar [6], are viewable from 360°. These systems achieved maintaining accurate gaze by providing a perspective correct image via a single user's head

CHI'14, April 26–May 1, 2014, Toronto, Canada.

http://dx.doi.org/10.1145/2556288.2557320



(a) Remote room schematic layout (b) Photo of the remote person (c) Local room schematic layout (d) Photo taken behind observer 1

Figure 2. Experiment setup: In the remote room, a camera array is used to capture unique and correct perspectives of the remote person gazing at the target 5. In the local room, a cylindrical multiview display is used to allow each observer to view their respective perspectives simultaneously. One of observers seating in viewpoint 1, only sees the video captured by camera 1.

position tracking. Eventually only a mono or stereo image is presented on the display, thus they are currently developed for a single observer. We propose a very low-cost telepresence system that combines the features of both multiview displays and situated displays discussed above.

SYSTEM DESIGN

The goal of our system is to allow multiple *observers* to perceive the gaze of a *remote person* accurately. That is observers can each see a unique and perspective-correct image from their viewing directions simultaneously. Each camera is linked to the corresponding projector to stream real-time video using TCP. The cylindrical screen ensures that each projected image will only be seen by an observer who is in the viewing zone for that projector. Also, using available off-the-shelf components allows our system to be built at a low cost. The cost for a three person multiview display would be approximately $1000, for a nine person multiview display would be approximately $2920.

Semicircular camera array construction

In the remote room, nine PlayStation® Eye USB digital cameras are vertically mounted on an angled table at radius of 600mm every 15°, as illustrated in Figure 2(a) and Figure 2(b). We manually adjust the cameras to look at the point above the center of the angled table. We then use Camera Calibration Toolbox for Matlab® to locate the cameras' positions and orientations accurately. The accurate positions and orientations of the cameras are used in the arrangement of projectors, so that accurate projecting of video can be done. The cameras are set to the 56° field of view setting. The cameras capture at 30 Hz at 640×480 pixel resolution. We arrange each camera vertically in order to make full use of the pixel resolution to represent the remote person's head.

Cylindrical multiview screen design

In the local room, the cylindrical screen is located at the center of an angled table which is the same size as the one in the remote room. We designed a cylindrical screen 32 cm in diameter and 70cm in height. The size is small enough to situate almost anywhere in a room. This display is visible from all directions, whereas flat displays are only visible from the front. The radius of curvature of the screen is similar to a real convex face to avoid the TV-screen-turn effect [1]. Using a cylindrical screen surface significantly simplifies projecting correct vertical perspective to observers at different heights and distances from the display.

The screen's main function is to reflect the image produced by a projector only to an observer in a very specific viewing zone. The idea of creating multiview screen for video conferencing is proposed by Nguyen et al. [4], where the screen's optics carefully retro reflects the light in the direction of the projector but diffuses it vertically, allowing viewers to see the image from any positions above or below the projector. Our screen consists of a retroreflective layer around the cylinder, with a one-dimensional diffuser layer 6mm above. Experimentation was conducted with different retroreflective materials, leading to the decision to use "white number plate reflective" from ORALITE® ($10), because it has a strong retroreflective characteristic, minimal reflective properties and good diffusive properties to reduce glare effects. A 1D lenslets-based lenticular sheet is used as the one-dimensional diffuser. The lines of the lenticular sheet placed horizontally to provide vertical diffusion. A 6mm or more physical spacing between retroreflective layer and lenticular sheet allows the light to mix vertically. The smooth side of the lenticular sheet is facing the observes and projectors. The 40 lenticules per inch (LPI) sheet with 49° viewing angle from Pacur® ($30) is chosen for two reasons: the thin thickness (0.838mm) of this sheet allows it easily to wrap around the cylinder; we only require a modest amount of vertical diffusion. More diffusion would hurt the brightness of the image.

Semicircular projector arrays construction

Nine projectors and observer viewpoints were set around the half annular table with a radius of 1500mm at every 15° which exactly line up with each camera in the remote room as depicted in Figure 2(c) and Figure 2(d). We vertically mounted each projector at a height of 1800mm, allowing an observer to sit under a projector. We use Projector-Camera Calibration Toolbox® to align the projectors' positions and orientations accurately. Each projector projecting a unique image on the part of the cylinder at the same horizontal level, but there are some overlap between images that projected by different projectors. The cylindrical multiview screen controls diffusion and produces relatively narrow viewing zones above, below, and slightly to the sides of a light source. Therefore, a observer sitting under the bottom of a projector sees only the image from that projector. We used NEC® NP110 projectors with resolutions of 800×600 pixels.

EXPERIMENT

The purpose of the experiment was to demonstrate that our cylinder multiview system can better represent the remote person's gaze for multiple observers. We measured the effectiveness of the displays by measuring the ability of multiple observers to accurately judge which target the remote person was gazing at.

We compared four display conditions. *Cylinder multiview multi-video* condition was our system discussed above, which could support correct viewing for multiple viewpoints around a conference table (see Figure 3(a)). *Cylinder multiview single-video* condition was identical to the cylinder multiview multi-video condition, except that only the center camera was used for capturing the remote person (see Figure 3(b)). All projectors project this video, instead of projecting unique perspective-correct videos. Thus, observers would perceive the gaze direction as if they were standing straight in front. This condition should show the benefit of using camera array. *Cylinder diffuse single-video* condition used a curved diffuse white projection screen. Only the center camera and projector were used (see Figure 3(c)). This condition mimicked Tele-Human [2], which developed for a single user; other users can view the display but will see a distorted view. *Flat diffuse single-video* condition used a conventional 2D flat screen, instead of 3D cylinder surface. This condition mimicked the commonly found Mona Lisa gaze effect, which occurs when 3D objects are rendered in 2D, causing the gaze perception of all in a room to be the same (see Figure 3(d)). Image quality remained the same in all conditions. We expect that viewers in cylinder multiview single-video condition and flat diffuse single-video condition will see much more incorrect targets compared to those in cylinder multiview multi-video condition. We further expect the cylinder diffuse single condition to lie between these two in performance, as the 3D cylindrical surface eliminates Mona Lisa gaze effect but viewers only could see part of head in some extreme viewpoints.

We explored four observers' viewpoints (1, 4, 5 & 9). We included viewpoint 5 where the observer at the center position as a benchmark; viewpoint 1 and 9 where observers sat at two extreme viewing angles; and viewpoint 4 where the observe sat right next to observer 5. In the cylinder multiview multi-video condition, we expect a similar level of error for observer perceiving targets at all viewpoints. For the other three display conditions, we expect the level of error will increase symmetrically as the viewpoint diverges horizontally from the central position.

Method

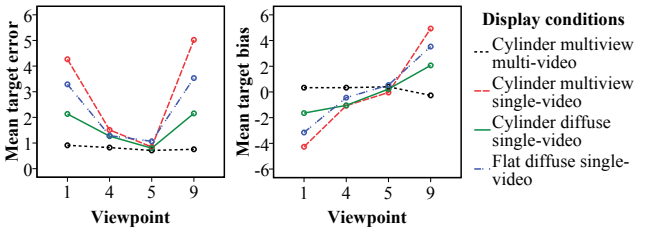
Participants

48 participants, students at University College London, were recruited to take part as observers in our user study. All participants had normal or corrected to normal eye sight. One further participant was an remote person recorded on video.

Apparatus and materials

We video-recorded the remote persons' head movements (see Figure 2(a)). The remote person sat at the center position of the table and his or her head is captured by 4 cameras simultaneously. The remote person listened to an audio recording that instructed to turn his or her head to look at the targets. A new target was given every 10 seconds. The targets were randomly ordered, each one was gazed at only once, amounting to 15 targets in the audio instruction and thus in the recorded videos. One set of 4 videos were generated.

Figure 3. Photos of display conditions taken from viewpoint 1: when the remote person gazing at the target 10, observers perceive different targets in four display conditions.



(a) Mean target error (b) Mean target bias

Figure 4. The mean target error and mean target bias for each display conditions and viewpoints.

Procedure

12 groups of four were used for testing, and each group experienced one of four different display conditions with each observer sat at one of the four viewpoints (see Figure 2(c)). Each observer was given a sheet of paper with an empty grid of 15 squares. The video of the remote person 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. Then, observers would then judge which target (1-15) the remote person was gazing at and then write this in the relevant grid square. The experiment took about 5 minutes. Participants received chocolates as compensation.

Result

The primary measurement in our results was the level of error in perceiving targets. We defined target error (ϵ_i) to be the absolute value of difference between the observer's perceived target number (t_{oi}) and the actual target number (t_{ai}): $\epsilon_i = |t_{oi} - t_{ai}|$. Figure 4(a) shows the target error at the four viewpoints in four display conditions. The line of the cylinder multiview multi-video condition shows that it achieved the lowest mean target error. The means were very similar across the four viewpoints, indicating that the viewpoint had little impact in this display conditions. At the extreme viewpoints (1 and 9), the means were significantly below that of the other three display conditions. In addition, the graph shows that the central viewpoint had the lowest mean target error, where four display conditions all had perspective-correct video; the mean target error increased symmetrically as the viewpoint diverges

from the central position for cylinder multiview single-video condition, cylinder diffuse single-video condition and flat diffuse single-video condition. This is expected as when the observer did not sit in viewpoint 5, those display conditions still used the video from camera 5.

A 4 display conditions $\times$ 4 viewpoints $\times$ 15 target positions mixed design ANOVA was conducted on the target error, with display condition and viewpoints as two between-subjects factors and target positions as a within-subjects factor. Mean target error differed significantly across the four display conditions, $F(3, 32) = 32.167, p < .001$. Tukey post-hoc tests revealed significant mean differences between each of the display conditions. The cylinder multiview multi-video condition ($M = .800, 95\%\ CI\ [.473, 1.127]$) gave significantly lower mean target error than the cylinder diffuse single-video condition ($M = 1.589, 95\%\ CI\ [1.262, 1.916]), p = .008$, the cylinder multiview single-video condition ($M = 2.911, 95\%\ CI\ [2.584, 3.238]), p < .001$, and the flat diffuse single-video condition ($M = 2.294, 95\%\ CI\ [1.968, 2.621]), p < .001$. This supports the primary hypothesis. Results also revealed a significant main effect of viewpoints, $F(3, 32) = 39.448, p < .001$. Tukey post-hoc comparisons indicated the mean target error at viewpoint 5 ($M = .856, 95\% CI\ [.529, 1.182]$) is significantly lower than viewpoint 1 ($M = 2.65, 95\%\ CI\ [2.323, 2.977]), p < .001$ and viewpoint 9 ($M = 2.867, 95\%\ CI\ [2.54, 3.194]), p < .001$, which supports the second hypothesis; however, it did not significantly differ from viewpoint 4 ($M = 1.222, 95\%\ CI\ [.895, 1.549]), p > .05$, which is expected as the seat position only slightly diverges from the front. The mean target error at viewpoint 1 did not significantly differ from viewpoint 9, $p > .05$, which is also expected as the viewing angles of viewpoint 1 and 9 are equal only opposite in direction. The display conditions $\times$ viewpoints interaction was significant, $F(9, 32) = 7.277, p < .001$, indicating that mean target error due to viewpoints were different in four display conditions.

We further investigated whether there was leftward bias or rightward bias in perceiving targets in different display conditions. We defined target bias (β_i) to be the difference between the observer's perceived target number (t_{oi}) and the actual target number (t_{ai}): $\beta_i = t_{oi} - t_{ai}$. Figure 4(b) shows the target bias at four viewpoints in four display conditions. Positive values indicated leftward biases whereas negative values indicated rightward bias. For the cylinder multiview multi-video condition, the mean target bias did not change substantially across different viewpoints. This further supports the hypothesis. By contrast, for the other three display conditions, the biases depended on the observers' viewpoints. For the flat diffuse single-video condition, the biases of four viewpoint in this study nicely fit in the previous work [7] that is the mean target bias varies linearly according to seat position. The graph also shows that the bias of cylinder diffuse single-video condition is less than flat diffuse single-video condition. This parallels the previous finding [1] that biases occur differently while observing convex, flat and concave surfaces.

CONCLUSION

We presented a novel display system for video conferencing. The highlights of this system are as follows. Firstly, the cylindrical display offers a 360° view whereas flat displays are only visible from the front. Secondly by using a camera array, a projector array and a multiview screen, we are able to transmit the remote person to multiple observers gathered around the cylindrical display, maintaining accurate cues of gaze direction. A similar cylindrical multiview display could also use a very dense projector array covering 360°, thus supporting a large number of viewpoints from any directions without introducing crosstalk and reducing resolution. Our current system is used for asymmetric conversations, such as a teaching scenario. Systems using similar principles could be configured to support symmetric conversations, by arranging camera arrays that are denser but further from the users. As cameras and projectors are now becoming very cheap, the low cost and ease of setup make this an interesting platform for next generation video conferencing.

REFERENCES

1. Anstis, S. M., Mayhew, J. W., and Morley, T. The perception of where a face or television'portrait'is looking. *The American journal of psychology* (1969).
2. Kim, K., Bolton, J., Girouard, A., Cooperstock, J., and Vertegaal, R. Telehuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod. In *SIGCHI*, ACM (2012), 2531–2540.
3. Lincoln, P., Welch, G., Nashel, A., Ilie, A., and Fuchs, H. Animatronic shader lamps avatars. *VR* (2011).
4. Nguyen, D., and Canny, J. Multiview: spatially faithful group video conferencing. In *SIGCHI* (2005), 799–808.
5. Okada, K., Maeda, F., Ichikawaa, Y., and Matsushita, Y. Multiparty videoconferencing at virtual social distance: Majic design. In *CSCW* (1994), 385–393.
6. Oyekoya, O., Steptoe, W., and Steed, A. Sphereavatar: A situated display to represent a remote collaborator. In *SIGCHI* (Austin, Texas, USA, 2012), 2551–2560.
7. Pan, Y., and Steed, A. Preserving gaze direction in teleconferencing using a camera array and a spherical display. In *3DTV-CON*, IEEE (2012), 1–4.
8. Prince, S., Cheok, A. D., Farbiz, F., Williamson, T., Johnson, N., Billinghurst, M., and Kato, H. 3-d live: real time interaction for mixed reality. In *CSCW* (2002).
9. Schreer, O., Kauff, P., and Sikora, T. *3D videocommunication*. Wiley Online Library, 2005.
10. Sellen, A., Buxton, B., and Arnott, J. Using spatial cues to improve videoconferencing. In *SIGCHI* (Monterey, California, USA, 1992), 651–652.
11. Vertegaal, R., Weevers, I., Sohn, C., and Cheung, C. Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In *SIGCHI* (Ft. Lauderdale, Florida, USA, 2003), 521–528.