PRESERVING GAZE DIRECTION IN TELECONFERENCING USING A CAMERA ARRAY AND A SPHERICAL DISPLAY

Ye Pan, Anthony Steed

Department of Computer Science
University College London, United Kingdom

ABSTRACT

The movement of human gaze is very important in face to face conversation. Some of the quality of that movement is lost in videoconferencing because the participants look at a single planar image of the remote person. We use an array of cameras to capture a remote user, and then display video of that person on a spherical display. We compare the spherical display to a face to face setting and a planar display. We demonstrate the effectiveness of the camera array and spherical display system in that it allows observers to accurately judge where the remote user is placing their gaze.

Index Terms — Non-planar displays, camera arrays, human gaze

1. INTRODUCTION

Video teleconferencing systems are now very widely spread, but their inability to represent the full intention of eye-gaze of participants is well-known (see e.g. [1], [2]). In particular although local users see a remote user's head move, it is difficult to judge the direction accurately. This can lead to problems sharing socially useful information such as attention targets, conversational turn-taking indicators etc.

As reviewed in the next section, a variety of multiple view systems have been built, though the majority use a planar display or a virtual reality system. We propose to use non-planar displays, in particular a spherical display as this type of display provides the same angle of view from all directions. Because cameras are now becoming very cheap, we further propose to use a camera array to capture the remote user, so that we can select an appropriate video of them to show. We can then evaluate whether local users can accurately determine where the user is looking.

We run an experiment to demonstrate that the camera array plus sphere display can convey gaze relatively accurately. This demonstration and results thus motivate the further study of novel display configurations and the supporting camera and networking infrastructure for them.

2. BACKGROUND

2.1. Display systems

A variety of video systems have been developed to improve several aspects of conversations. These include MAJIC [3], Hydra [4], GAZE-2 [5], and MultiView [1]. Current immersive systems, such as, Im.point [6] also can replicate a life-like face to face conversation. The most related previous display work to ours is SphereAvatar [7] which also uses a spherical display, but which uses synthetic avatar head.

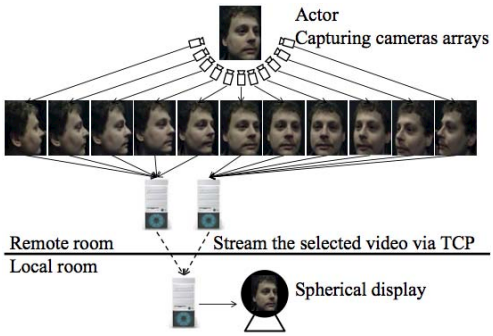


Figure 1. Diagram of the directional spherical video conferencing system.

2.2. Camera arrays

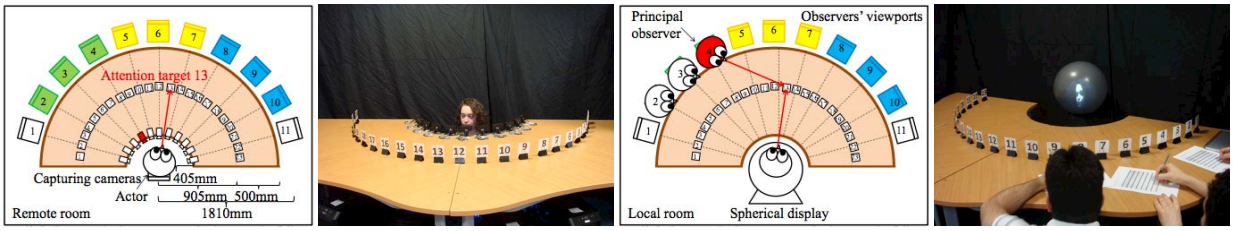
A scene can be captured by a set of cameras and the video streams can be interpolated to achieve free viewpoint video [6]. When the camera density is very high, view generation is simply by selecting the closest camera image. When the camera density is moderately high, view generation needs some processing. When the camera density is low, intermediate views can be generated by detecting geometry in the scene. In our experiment, the views are moderately dense, but we are not currently doing view interpolation.

2.3. Gaze evaluation

Detecting the gaze direction of a person is important for human-computer interaction applications in video conferencing or shared collaborative workspaces [2]. Nguyen et al. [1] 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. 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 work is using this framework in experiment design, but the camera and display configuration is very different.

3. SYSTEM DESIGN

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 eleven capturing cameras controlled by two PCs. In the *local*



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

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 renderer PC. Streaming is done using TCP.

3.1. Semicircular camera arrays

In the remote room, eleven low-cost PlayStation® Eye USB digital cameras are mounted on a half annular table with an inner radius of 405mm at every 15°, as illustrated in Figure 2(a). The cameras are set to the 56° field of view setting. The cameras capture at 30 Hz at 320×240 pixel resolution, which provide qualified videos while saving bandwidth for real time streaming.

We manually adjust the cameras to look at the point above the centre of the half annular table. We then use Camera Calibration Toolbox for Matlab® to locate the cameras' positions and orientations accurately. These positions and orientations are used in the rendering process, see below.

3.2. Directional spherical screen

In the local room, a spherical display is located at the centre of a half annular table which is the same size as the one in the remote room. Eleven observer viewports set around the half annular table with a radius of 1810mm at every 15° which exactly line up with each camera in the remote room as depicted in Figure 2(c). 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 on to the inside of the sphere.

The presentation of the remote participant onto the sphere is done in four main stages shown in Figure 3.

First, a sphere acts as a proxy geometry of 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"[8]. According to the observer's viewport, the video captured by 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 in to an environment map. The idea of storing environment maps as cube maps is proposed by Greene where six subimages representing the six different faces of a cube[9]. We render the scene in to an environment map using six cameras positioned outside the cube at the position

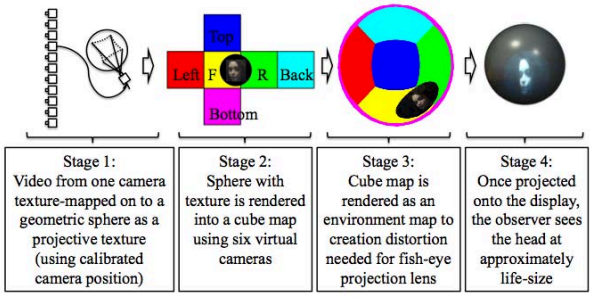


Figure 3. Illustrating stages of the rendering pipeline.

of the observer's eye. Each of the six facets of cube map is thus rendered using the non-symmetric view volumes. The resulting cube map looks as if the head is outside looking in, but once reflected in the environment mapping, it gives the illusion that the head is situated within the spherical display.

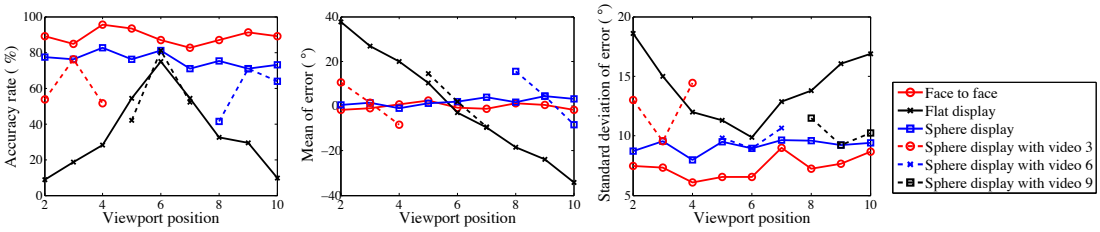
Then, we draw a 3D sphere using an environment map. Environment mapping proposed by Blinn and Newell simulates the reflectance of a surface, by using the reflected eye vector as a lookup into the texture rather than a simple texture coordinate[10]. We render a sphere with the environment map as its texture in order to generate a 2D distorted image, that is suitable for projection through a fish eye lens [7].

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

3.3. Hypotheses

The purpose of the experiment is to demonstrate that the combination of a spherical display and a camera array (*sphere display* condition) can better represent the actor's gaze than can a single camera view shown on a flat display (*flat display* condition). We also include a benchmark where the actor can be viewed directly (*face to face* condition).

We measure the effectiveness of the displays by measuring the ability of observers to accurately judge which target the actor is gazing at. We make the following two hypotheses: firstly, both face to face and sphere display will demonstrate higher levels of accuracy than flat display when the observers are in varied positions. We further expect face to face to be better than sphere display. Secondly, for both sphere display and flat display, we expect that if the observer is not sat in the same direction as the camera that is observing the actor, the accuracy will be worse than



(a) Accuracy rate (b) Mean of error (c) Standard deviation of error

Figure 4. Result for analysing the actual targets and perceived targets in different treatment conditions.

if the camera chosen for the display is aligned with the observer's position.

3.4. Method

3.4.1. Participants

60 participants, students and staff at University College London, were recruited to take part as observers in our user study. 20 groups of three were used for testing and each group experienced one of three different conditions (sphere display, flat display, face to face). Eight further participants were actors in these experiment: four acted were recorded on video for the sphere and flat condition and four acted in 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, for the purposes of control of the experimental stimuli, in the sphere and flat display conditions, video was recorded to disk and replayed, so that the cameras and display could be configured in the same way as in the face to face configuration.

3.4.2. Apparatus and materials

We use a half annulus table with the larger semicircle of radius 1405mm and the smaller one of radius 405mm see Figure 2. Horizontally, 23 gaze target cards were placed in a semicircle of radius 905mm, every 7.5° on the table. When capturing video, the target cards with even numbers lined up with cameras and observer's viewports. We ensured that the vertical alignment of the eye gaze of the actor, the eye level of observers, eye level of the video of the actor on the spherical or flat display, capturing cameras, attention target cards were the same. This ensured equivalence in stimuli alignment and apparent size between the two display conditions and the face to face condition.

For the two display conditions we recorded the actor's performance. The actor sits at the centre position of the half annulus table and his or her head is captured by 11 video cameras. The actor listens 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, giving 46 targets in the audio instruction and thus in the recorded videos. Four participants were actors, and thus four sets of 11 video were generated.

3.4.3. 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 & 4 or 5,6 & 7 or 8, 9 & 10, in a balanced random order.

For each trial, each observer was given a sheet of paper with an empty grid with 46 squares. In all three conditions, 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 then write this in the relevant grid square.

For the face to face condition, the three observers and actor were in the same room. The actor sat at the centre position of 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 (see Section 3.4.2). 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. 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 actor video three times; on each repetition the group of three observers move to another one of the positions and each observer was the principal observer on one of these occasions. For the flat display condition, the three observers observed the pre-recorded video on the flat display. The video was always that from position six, simulating a simple web-cam set up where the observers might be looking obliquely at the screen, and the actor looking obliquely at the camera. The experiment took about 20 minutes.

3.5. Result

The results of experiment are presented in Figure 4. 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 centre of conferencing table. Therefore, $\alpha = p \times 15°$.

The primary measurement in our results is the accuracy rate in perceiving the attention target: the observer is accurate if they successfully identify the correct target. We then define error(ϵ_i) to be the difference between the actual target number(t_{ai}) and the observer perceived attention target number(t_{oi}) converted to degrees, based on attention targets being 7.5° apart from each other. Thus, $\epsilon_i = (t_{ai} - t_{oi}) \times 7.5°$.

Each observer indicates 46 target positions in each trial. Each observer does three trials. There are 12 observers in the face to face condition (four groups of three). There are nine observer seating positions. Thus, there are 184($46 \times 3 \times 12/9$) rating events in each seating position. Similarly there are 184 rating events in each seating position for the flat display. For the sphere display, there are 36 observers (twelve groups of three) but only one of the group is in the principal position. Thus, there are also

$184(46 \times 3 \times (36/3)/9)$ rating events for principal observers in each of the nine observer seating positions. However in the analysis below, we include some data from the secondary observers. In particular, for seating positions 3, 6 and 9 we analyse the 184 rating events for the observer sat on their left and 184 rating events for the observer on their right. This gives us a view of how important it is to use the correct video for the observer position.

The result of accuracy rate in different conditions is shown in Figure 4(a). For the 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 sat in position 6, they still see the video from 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 flat display condition. The fact that the accuracy doesn't 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 flat display.

When we consider the secondary positions in the sphere display (the three "three point hat" graphs in Figure 4(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 principle observer seated immediately to their right. This pattern is repeated for all secondary observers.

The difference between face to face performance and sphere display performance is likely due to video quality. We note that for observer position 6 on the 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 flat display, but it has the advantage that it has the same apparent size in the different observer positions.

Next, we analysed the mean of error and standard deviation of error for the actual targets and observer perceived targets in different treatment conditions in Figure 4(b) and Figure 4(c). For face to face meeting and sphere display, the observer position has no significant effect on the mean of error which is around 0°. The standard deviation of the error for the sphere display is higher, but there is no systematic bias, indicating that the observers are generally finding it harder to determine gaze.

In contrast, for the flat display, the mean of error varies linearly according to viewport position. We utilised the first-order Matlab® Polyfit function to generate the coefficients of the polynomial to simulate a curve to fit the data and found the relationship between the error of mean and angle of viewport position: $\sigma(\epsilon_i) = -0.6\alpha + 54.27^\circ = 0.6 \times (90^\circ - \alpha) + 0.27^\circ$.

4. DISCUSSION & CONCLUSIONS

We have presented a novel display system for video conferencing. The highlights of this system are as follow: firstly, the spherical display offers a 360° view whereas flat displays are only visible from the front; secondly by using a surrounding camera array we allow principal observers to accurately tell where the visitor is looking from multiple observing positions. It is notable that with the sphere display, for the principal observer positions we see no systematic bias in the error of angle, so that we can reproduce accurate judgement of gaze direction at all angles.

The linear model of error in the flat display condition is interesting in that it suggests that the observer's judgement of gaze angle from front is only 60% of what it should be; therefore for the flat display the observer perceives the actor to be looking more directly straight out of the display. This effect appears very reliable and this means that it may be possible to model and thus predict the distortion. We might be able to correct for this distortion in some display configurations.

There are several routes for development. 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. Further, we will investigate less constrained positioning of the cameras and different eye-lines. Also, the display could be made for multiple viewers. As noted, although the experiment necessarily used recorded data, the system will run in a live, automatic camera switching mode and thus we will investigate how users utilize movement to control the video.

5. REFERENCES

[1] D. Nguyen and J. Canny, "Multiview: spatially faithful group video conferencing," in *SIGCHI*, Portland, Oregon, USA, 2005, pp. 799–808.

[2] D. Roberts, R. Wolff, J. Rae, A. Steed, R. Aspin, M. McIntyre, A. Pena, O. Oyekoya, and W. Steptoe, "Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together," in *VR*, Washington, DC, USA, 2009, pp. 135–142.

[3] K.I. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita, "Multiparty videoconferencing at virtual social distance: Majic design," in *CSCW*, Chapel Hill, North Carolina, USA, 1994, pp. 385–393.

[4] A. Sellen, B. Buxton, and J. Arnott, "Using spatial cues to improve videoconferencing," in *SIGCHI*, Monterey, California, USA, 1992, pp. 651–652.

[5] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "Gaze-2: conveying eye contact in group video conferencing using eye-controlled camera direction," in *SIGCHI*, Ft. Lauderdale, Florida, USA, 2003, pp. 521–528.

[6] O. Schreer, P. Kauff, and T. Sikora, *3D videocommunication*, Wiley Online Library, 2005.

[7] O. Oyekoya, W. Steptoe, and A. Steed, "Sphereavatar: A situated display to represent a remote collaborator," in *SIGCHI*, Austin, Texas, USA, 2012, pp. 2551–2560.

[8] M. Segal, C. Korobkin, R. Van Widenfelt, J. Foran, and P. Haeberli, "Fast shadows and lighting effects using texture mapping," in *ACM SIGGRAPH Computer Graphics*, Jul. 1992, vol. 26, pp. 249–252.

[9] N. Greene, "Environment mapping and other applications of world projections," *Computer Graphics and Applications, IEEE*, vol. 6, pp. 21 –29, Nov. 1986.

[10] J.F. Blinn and M.E. Newell, "Texture and reflection in computer generated images," *Commun. ACM*, vol. 19, pp. 542–547, Oct. 1976.