# Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System

Andrew Jones* Magnus Lang* Graham Fyffe* Xueming Yu* Jay Busch* Ian McDowall† Mark Bolas*‡ Paul Debevec*

* University of Southern California
Institute for Creative Technologies

† Fakespace Labs

‡ University of Southern California
School of Cinematic Arts

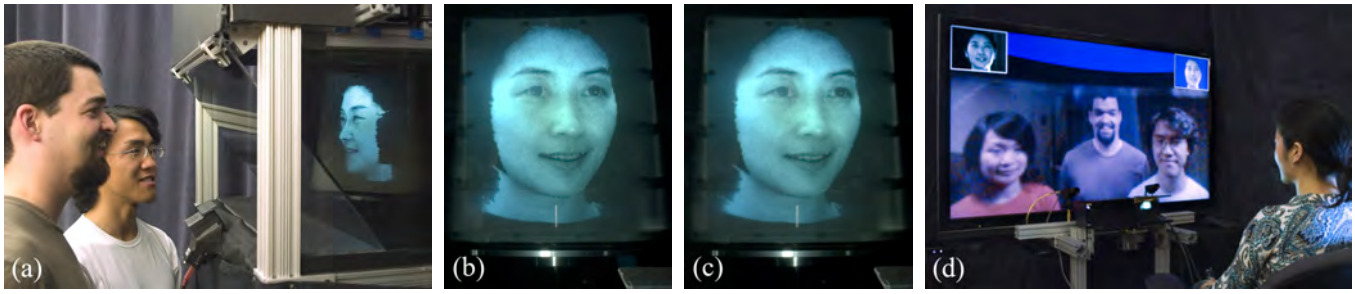



**Figure 1:** *(a) An audience interacts with a remote participant (RP) rendered in 3D on an autostereoscopic display. (b,c) A cross-fusable stereo pair where the RP appears life-size in correct perspective, able to make eye contact with the members of the audience. (d) The RP looks back at the audience via geometrically calibrated wide-angle 2D video while being scanned, transmitted, and rendered at 30Hz.*

## Abstract

We present a set of algorithms and an associated display system capable of producing correctly rendered eye contact between a three-dimensionally transmitted remote participant and a group of observers in a 3D teleconferencing system. The participant's face is scanned in 3D at 30Hz and transmitted in real time to an autostereoscopic horizontal-parallax 3D display, displaying him or her over more than a 180° field of view observable to multiple observers. To render the geometry with correct perspective, we create a fast vertex shader based on a 6D lookup table for projecting 3D scene vertices to a range of subject angles, heights, and distances. We generalize the projection mathematics to arbitrarily shaped display surfaces, which allows us to employ a curved concave display surface to focus the high speed imagery to individual observers. To achieve two-way eye contact, we capture 2D video from a cross-polarized camera reflected to the position of the virtual participant's eyes, and display this 2D video feed on a large screen in front of the real participant, replicating the viewpoint of their virtual self. To achieve correct vertical perspective, we further leverage this image to track the position of each audience member's eyes, allowing the 3D display to render correct vertical perspective for each of the viewers around the device. The result is a one-to-many 3D teleconferencing system able to reproduce the effects of gaze, attention, and eye contact generally missing in traditional teleconferencing systems.

## 1 Introduction

When people communicate in person, numerous cues of attention, eye contact, and gaze direction provide important additional channels of information [Argyle and Cook 1976], making in-person meetings more efficient and effective than telephone conversations and 2D teleconferences. However, with collaborative efforts increasingly spanning large distances and the economic and environmental impact of travel becoming increasingly burdensome, telecommunication techniques are becoming increasing prevalent. Thus, improving the breadth of information transmitted over a video teleconference is of significant interest.

The potential utility of three-dimensional video teleconferencing has been dramatized in movies such as *Forbidden Planet* and the *Star Wars* films. The films usually depict a single person transmitted three-dimensionally from a remote location to interact with a group of colleagues somewhere distant. The films depict accurate gaze and eye contact cues which enhance the dramatic content, but the technology is fictional. A recent demonstration by CNN showed television viewers the full body of a remote correspondent transmitted "holographically" to the news studio, appearing to making eye contact with the news anchor. However, the effect was performed with image compositing in postproduction and could only be seen by viewers at home; the anchor actually stared across empty space toward a traditional flat panel television [Rees 2008]. The Musion Eyeliner (http://www.eyeliner3d.com/) system claims holographic "3D" transmission of figures such as Price Charles and Richard Branson in life size to theater stages, but the transmission is simply 2D high definition video projected onto the stage using a Pepper's Ghost [Steinmeyer 1999] effect only viewable from the theater audience; the real on-stage participant must pretend to see the transmitted person from the correct perspective to help convince the audience of the effect. CISCO Systems' TelePresence systems use a controlled arrangement of high-definition video cameras and life-size video screens to produce the impression of multiple people from different locations sitting around a conference table, but the use of 2D video precludes the impression of accurate eye contact: when a participant looks into the camera, everyone seeing their video stream sees the participant looking toward them; when the participant looks away from the camera (for example, toward other participants in the meeting), no one sees the participant looking at them.

In this work, we develop a one-to-many teleconferencing system which uses a novel arrangement of 3D acquisition, transmission, and display technologies to achieve accurate reproduction of gaze direction and eye contact. We target the common application where

a single *remote participant* (RP) wishes to attend a larger meeting with an audience of local participants. In this system, the face of the RP is three-dimensionally scanned at interactive rates while watching a large screen showing an angularly correct view of the audience. The scanned RP's geometry is then shown on the 3D display to the audience. To achieve accurate eye contact in this 3D teleconferencing system, we make the following contributions:

1. We combine an adaptation of the real-time face scanning system of [Zhang and Huang 2006] with an evolved version of the 3D display system of [Jones et al. 2007], allowing life-sized 3D face transmission.
2. We reformulate a generalized multiple-center-of-projection rendering technique for accurately displaying three-dimensional imagery to arbitrary viewer positions by projecting onto anisotropic display surfaces. In particular, we generalize the technique to arbitrarily curved display surfaces, which allows using a concave display surface which significantly simplifies projecting correct vertical perspective to audience members at different heights and distances from the display. We achieve high-speed rendering with accurate conic intersection mathematics using a 6D vertex shader lookup table.
3. We project accurate, dynamic vertical parallax of the remote participant to multiple simultaneous viewers at different heights by interactively tracking the viewers' face positions in the teleconference video stream. This allows tracked vertical and autostereoscopic horizontal parallax to be simulated using a horizontal-parallax-only display, allowing accurate eye contact to be simulated.

## 2 Background and Related Work

Gaze, attention, and eye contact are important aspects of face to face communication [Argyle and Cook 1976]; they help create social cues for turn taking, establish a sense of engagement, and indicate the focus and meaning of conversation. Although eye contact sensitivity is asymmetric [Chen 2002] and special configurations can help experienced users determine which gaze directions signify mutual eye contact [Grayson and Monk 2003], it is still useful to develop systems that intrinsically support direct eye contact. Systems that support direct eye contact have elicited behaviors more similar to face to face conversation, allowing users to more quickly confirm the communications channel [Mukawa et al. 2005] and more easily develop trust in a group [Nguyen and Canny 2007].

Beamsplitters (e.g. [Quante and Muhlbach 1999]), teleprompter type configurations, and other hardware have been used to create direct eye contact in 2D video systems ([Rose and Clarke 1995] and [Grayson and Monk 2003] review many such systems). One notable use of such hardware has been in the film industry for documentary interviews with an enhanced dramatic connection to the audience [Morris 2004]. Other researchers have demonstrated software techniques for resynthesizing video imagery to enhance eye gaze, correcting for off axis camera and display placement [Yang and Zhang 2004; Jerald and Daily 2002; Gemmell et al. 2000; Ott et al. 1993; Liu et al. 1995]. Other telecollaboration systems that leverage eye gaze include Clearboard [Ishii et al. 1993], GAZE [Vertegaal 1999], Hydra [Sellen 1995], and Multiview [Nguyen and Canny 2005].

The design of our system is informed by several human factors studies. [Prussog et al. 1994] performed experiments demonstrating that the impression of telepresence is increased if the remote viewer is shown at natural size and/or stereoscopically. [Muhlbach et al. 1995] found that achieving more accurate eye contact angles improved participants' ability to recognize individually addressed nonverbal signals. [Chen 2002] reported that the perception of eye contact decreases below 90 percent if the horizontal contact angle is greater than $1^\circ$ or the vertical contact angle is greater than $5^\circ$. Accordingly, our design transmits the remote participant at natural size, autostereoscopically, with eye contact angles consistent with the tolerances recommended by [Chen 2002].

In order to achieve an autostereocopic experience across a wide field of view, it is necessary to record the participant in a manner that can be re-rendered from any point of view. Large camera arrays have been used to capture a subject's light field from all possible angles [Wilburn et al. 2005; Yang et al. 2002]. Our display generates 72 unique views over $180^\circ$, so a linear light field capture system would require 72 cameras for a single viewing height; [Matusik and Pfister 2004] showed a 3D TV system using 16 such cameras and projectors over approximately a $30^\circ$ field of view. However, several hundred cameras in a 2D array would be required to render sharp horizontal and vertical parallax over a wide field of view; [Taguchi et al. 2009] uses 64 cameras for a relatively narrow field of view. Furthermore, in two-way teleconferencing, it is difficult to distribute a large number of cameras without obstructing the participant's view of their own display.

An alternate approach is to render novel viewpoints based on captured 3D geometry (e.g. [Gross et al. 2003]). Many techniques exist for recovering geometry based on multi-camera stereo but few achieve real-time speeds. To make stereo matching more efficient, some real-time face scanning systems (e.g. [Raskar et al. 1998]) use active illumination to disambiguate geometry reconstruction. We use a similar phase unwrapping based approach to [Zhang and Huang 2006] based on a rapid series of projected sinusoid patterns. Although the system requires active illumination and can fail for fast moving scenes, it works well for facial conversations and requires only modest bandwidth and hardware.

## 3 System Overview

Our 3D teleconferencing system (Fig. 2) consists of a 3D scanning system to scan the *remote participant* (RP), a 3D display to display the RP, and 2D video link to allow the RP to see their *audience*.

**Real-time 3D Face Scanner** The face of the RP is scanned at 30Hz using a structured light scanning system based on the phase-unwrapping technique of [Zhang and Huang 2006]. The system uses a monochrome Point Grey Research *Grasshopper* camera capturing frames at 120Hz and a greyscale video projector with a frame rate of 120Hz. We determine the intrinsic and extrinsic calibration between the two using the calibration technique of [Zhang 2000]. Our four repeating patterns shown in Fig. 3 include the two 90-degree phase-shifted sinusoid patterns of [Zhang and Huang 2006], but instead of a fully-illuminated frame, we project a frame half-lit on the left followed by a frame half-lit on the right. We subtract one half-lit image from the other and detect the zero-crossings across scan lines to robustly identify the absolute 3D position of the pixels of the center of the face, allowing the phase unwrapping to begin with robust absolute coordinates for a vertical contour of seed pixels. Conveniently, the maximum of these two half-lit images provides a fully-illuminated texture map for the face, while the minimum of the images approximates the ambient light in the scene [Nayar et al. 2006]. We found that by subtracting ambient light from all frames, the geometry estimation process can be made to work in the presence of a moderate amount of ambient illumination. Generally, we found 120Hz capture to be relatively robust to artifacts resulting from temporal misalignment, though fast facial motion can produce waviness in the recovered geometry.

The result of the phase unwrapping algorithm is a depth map image for the face, which we transmit along with the facial texture images

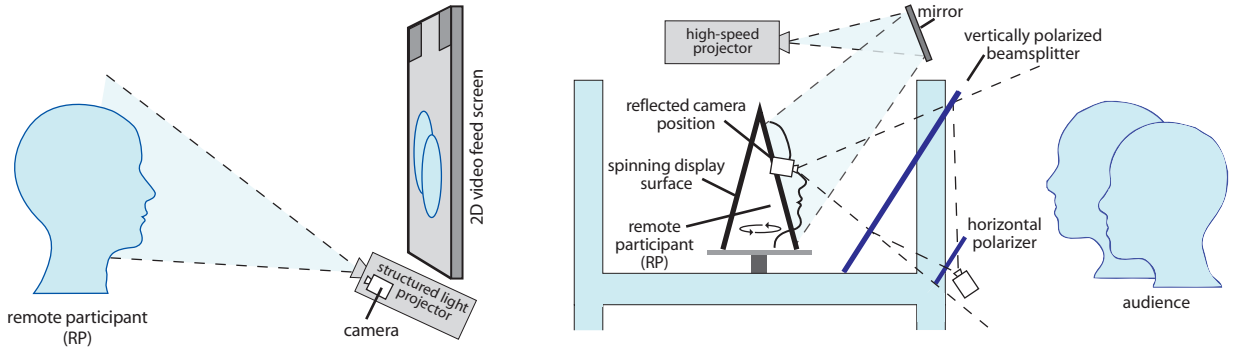
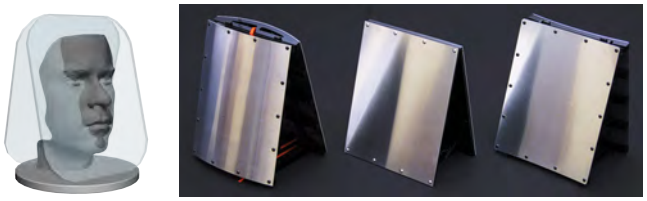



**Figure 2:** *(Left) The real-time 3D scanning system showing the structured light scanning system (120Hz video projector and camera) and large 2D video feed screen. (Right) The 3D Display apparatus showing the two-sided display surface, high-speed video projector, frontal beamsplitter, and 2D video and face tracking camera. Crossed polarizers prevent the video feed camera from seeing past the beamsplitter.*

**Figure 3:** *Patterns used for 3D face scanning, including two phase-shifted sinusoids and two half-lit images. The set of patterns is repeated at 30Hz.*



(a) (b)

**Figure 4:** *(a) A 3D face model shown intersecting the tent-shaped display surface. (b) Convex, flat, and concave display surfaces made from brushed aluminum sheet metal.*

at 30Hz to the display computer. The texture image is transmitted at the original $640 \times 480$ pixel resolution but we filter and downsample the depth map to $80 \times 60$ resolution. This downsampling is done to reduce the complexity of the polygonal mesh formed by the depth map, since it must be rendered at thousands of frames per second to the 3D display projector. While so far we have transferred this data only over a local area network, common image compression techniques (e.g. JPEG) would easily reduce the bandwidth to a similar amount used in commercial long-distance video chat systems. Real-time decimation and hole-filling algorithms could be used to improve the quality of the transmitted geometry.

**Autostereoscopic 3D Display** Our display is based on [Jones et al. 2007] with several key differences. The size, geometry, and material of the spinning display surface have been optimized for the display of a life-sized human face. The display surfaces (Fig. 4) are in the form of a two-sided tent shape with symmetrical sides made from thin $20cm \times 25cm$ sheets of brushed aluminum sheet metal. The brushed aluminum's high reflectivity and strongly anisotropic reflectance make it an inexpensive substitute for the holographic diffuser material of [Jones et al. 2007]. The two-sided shape provides two passes of a display surface to each viewer per full rotation, achieving a 30Hz visual update rate for 900 rpm rotation compared to the 15Hz update rate of [Jones et al. 2007]. The angle of the tent and the size of the surfaces were chosen to be consistent with the sloped shape of the human face as seen in Fig. 4(a). Instead of being placed directly above the display surface, the high-speed video projector projects onto the display surface from the front, inclined thirty degrees from the horizontal (two first-surface mirrors fold the projector's position into the top of the display.) As a result, the display has just somewhat more than a 180° field of view instead of the full 360° [Jones et al. 2007], but this omits only views of back of the head and allows nearly the full $1024 \times 768$ resolution of the projector to cover the display surface.

A monochrome MULE (Multi-Use Light Engine) high speed projector from Fakespace Labs projects 1-bit (black or white) frames at 4,320 frames per second using a specially encoded DVI video signal. Effectively, the display projects seventy-three unique views of the scene across a 180° field of view, yielding an angular view separation of 2.5 degrees. For a typical inter-pupillary distance of 65mm, this provides binocular stereo for viewing positions up to 1.5m away. Greyscale levels are simulated by a $4 \times 4$ ordered dither pattern implemented as a pixel shader. While the current system is monochrome, color could be achieved using multiple projectors with dichroic beamsplitters or a three-chip DLP projector.

**2D Video Feed** An 84° field of view 2D video feed allows the remote participant to view their audience interacting with their three-dimensional image. A beamsplitter (Fig. 2) is used to virtually place the camera near the position of the eyes of the 3D RP. The beamsplitter used is one of the protective Lexan transparent shields around the spinning mirror. We place linear polarizers with perpendicular polarization orientations on the camera lens and the Lexan shield to block light from the display from reaching the camera; a related technique appears in [Ishii et al. 1993]. Thus, the camera sees only light reflecting off the front of the polarizer on the shield. The camera thus sees a (slightly dim) view of the audience, easily made usable by ensuring the audience receives adequate illumination and using a wide f/stop on the camera lens.

The video from the aligned 3D display camera is transmitted to the computer performing 3D scanning of the RP. Key to our work is that the scanning computer also performs face tracking on this video stream so that the 3D display can render the correct vertical parallax of the virtual head to everyone in the audience. In addition, the scanning computer displays the video of the audience on a large LCD screen in front of the RP. The screen in our system is approximately 1.8 meters wide, one meter away from the RP, covering an 84° field of view. We calibrate both the camera and the projector's distortion parameters and field of view, and we texture-map a polygonal mesh with the transmitted images such that angles are consistent between rays captured by the camera and rays seen on the screen by the RP. Thus, the RP receives a view of the audience as if they were in the position of the virtual face. While the RP's view is not autostereoscopic, the screen is approximately at the typical distance of the audience members to the displayed RP so vergence is nearly correct.

## 4 Projecting 3D Vertices to the Display

To render 3D geometry to the display, we need to be able to project a 3D world-space vertex $(Q_x, Q_y, Q_z)$ to the appropriate pixel of the video projector $(p_u, p_v)$. Where this vertex should be drawn also depends on the current rotation angle of the mirror $\theta$ and the height and distance $(V_h, V_d)$ of the viewer who will observe that vertex. We formulate the projection so that one projection function works for any viewing azimuth $V_\psi$ around the display, that is, for a circle of potential viewing positions all at distance $V_d$ and height $V_h$ from the display. Thus, the projection function we desire is of the form $(Q_x, Q_y, Q_z, \theta, V_h, V_d) \mapsto (p_u, p_v)$.

To build this lookup table, we need to compute how a ray through a projector pixel will reflect off the anisotropic display surface and intersect some viewpoint at $(V_h, V_d, V_\psi)$. This problem was first addressed in [Jones et al. 2007], but with the simplifying assumption that rays reflect from the anisotropic display surface as vertically-aligned planar sheets of light allowing a real-time analytic solution to the projection. For their case, where the projector was directly above a display surface tilted at 45°, they argued a that a planar approximation is reasonably accurate. Unfortunately, our projector's off-axis relationship to the spinning display surface and and our desire to project onto arbitrarily curved display surfaces makes such an approximation unworkable.

Our brushed aluminum display surfaces reflect light as small cylindrical micro-facets aligned with the dominant axis of anisotropy $\vec{a}$. According to models of anisotropic reflection [Poulin and Fournier 1990; Kajiya and Kay 1989], a projector ray striking a cylindrical micro-facet will be specularly reflected with respect to all mirror angles perpendicular to the cylinder, forming a *cone* of light whose angle at the apex is equal to the angle of incidence (Fig. 5). Since our display surface rotates to oblique angles with respect to the incident projector rays, these cones have significant curvature for non-frontal viewing positions. Thus, for each pixel of the projector, we must intersect the reflected cones of light with the viewing circles $(V_h, V_d)$, which yields a *quartic* equation – a consequence of both cones and circles being quadratic.

To avoid implementing a real-time quartic equation solver, we built a six-dimensional lookup table evaluated across $(Q_x, Q_y, Q_z, \theta, V_h, V_d)$, solving the requisite circle/cone intersections using a GPU-accelerated numerical search.

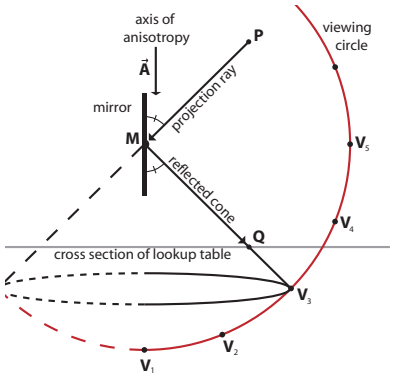



**Figure 5:** *A top view showing the anisotropic reflection of a projector ray into a cone. The angle formed between the ray and the axis of anisotropy is equal to the apex angle of the cone.*

**Building the Projection Lookup Table** We observe that the function $(Q_x, Q_y, Q_z, \theta, V_h, V_d) \mapsto (p_u, p_v)$ is smooth over all the dimensions, and thus can be approximated by a reasonably sparse lookup table ranging over each of the six arguments. We evaluate a lattice of points $Q$ over a $30cm^3$ volume (somewhat larger than the volume swept by the spinning display surface) in increments of $3.75cm$. We evaluate $\theta$ over 360° in increments of 1.25°. We evaluate $V_h, V_d$ over typical viewing heights ($-50cm$ to $+10cm$) and distances ($0.5m$ to $2.0m$) relative to the center of the display surface, in increments of $50cm$ and $10cm$ respectively.

For each input $(Q, \theta, V_h, V_d)$, we search for the viewing angle $V_\psi$ that coincides with some cone of reflected light, and the corresponding projector coordinates $(p_u, p_v)$ that produce the light ray that is reflected. This search lends itself well for parallelization on the GPU. We evaluate 2D cross-sections of the lookup table directly on the GPU, spanning the $(Q_x, Q_y)$ dimensions. We then iterate over $Q_z$ to fill the entire table.

To evaluate a 2D cross-section of the table, we first triangulate the mirror surface and store normal and tangent information at each vertex, as well as the $(p_u, p_v)$ projector coordinates that project onto each vertex. We then iterate $V_\psi$ over the possible viewing angles. We use a vertex shader to project the vertices onto the lookup table cross section as viewed from $V' = (V_h, V_d, V_\psi)$, using a frustum with four corner rays passing through the four corners of the lookup table cross section. We also compute and store $p_d$, the distance from the vertex to $V'$, in order to provide the GPU with depth values for z-buffer-based hidden surface removal. The GPU then rasterizes this projected mirror surface at discrete $(Qx, Qy)$ positions, corresponding to cells in our lookup table. We use a fragment shader to evaluate the discrepancy between the view ray and the cone of reflected light, using the interpolated position, normal, and tangent information at each rasterized sample, and the position of the viewer and projector. The discrepancy can be measured as $|\vec{I} \cdot \vec{a} - \vec{L} \cdot \vec{a}|$, where $\vec{I}$ is the direction towards the viewer, and $\vec{L}$ is the direction from the projector. We keep track of the smallest discrepancy value seen for each pixel by storing the value in the alpha channel of the frame buffer, and discard any sample with a higher value than the value already in the buffer. We then store the interpolated $(p_u, p_v, p_d)$ values in the RGB channels of the frame buffer. After iterating through all possible $V_\psi$, we copy the values in the frame buffer to the 2D cross-section of the lookup table.

Note that this lookup table (or LUT) computation does not assume that the display surface is flat and as a result, we can compute lookup tables for arbitrarily shaped mirrors. The output of the LUT is geometry mapped into projector UV coordinates. The geometry will appear warped in order to compensate for the mirror shape, mirror angle, and projector perspective and keystoning. To illustrate this warp, we applied the LUT to a vertical front-facing plane of geometry as shown in Fig. 6. The transform applied by the LUT is generally smooth, though curvature increases towards grazing angles. The mirror faces the front at 0°. At 30°, the image will reflect to viewers located 60° off-center. The resulting graphs are oriented so that 'up' corresponds to the top of the mirror and the black rectangle represents the extent of the projector frame.

**Evaluating the Lookup Table** We store our lookup table in the GPU memory as a series of 3D single-precision floating-point textures, each spanning $(Q_x, Q_y, Q_z)$ for a given $V_h$, $V_d$, and $\theta$. At the sampling densities we use, the lookup table for a particular mirror requires 24MB of memory, which requires little of the 1.5GB of memory of the display computer's nVIDIA graphics card. We currently use linear interpolation to evaluate the lookup table at intermediate vertex and viewing positions, and choose our sampling density to yield a sufficient approximation in this context. Our display's graphics card can perform 3D texture lookup with automatic trilinear interpolation. Thus, we need to perform just four texture lookups for the neighboring evaluated values of $V_h$ and $V_d$ in the vertex shader. Due to the synchronization of the rendered frames to the mirror rotation, the value of $\theta$ is always sampled exactly and requires no interpolation. If more than one viewer is present, we

select the $V_h$ and $V_d$ values per rendered frame by identifying the viewer closest to a ray originating from the center pixel of the projector and reflected off the rotated mirror. If the lookup table size were an issue, it is also possible to fit a higher-order polynomial to the LUT entries and store only the corresponding coefficients. On some current generation GPUs, this could provide a significant speedup as support for 3D texture lookups is not always fully optimized.

## 5 Flat and Curved 3D Display Surfaces

To demonstrate our projection technique, we designed and tested three different display surfaces – convex, flat, and concave – as seen in Fig. 4. These differently shaped surfaces offer different advantages and disadvantages, underscoring the utility of being able to project onto arbitrary surfaces. All display surfaces have the same 15°-from-vertical double-sided design and a surface area of 20*cm* wide by 25*cm* high of thin brushed aluminum sheet metal. The shape of each mirror is supported by a custom assembly of laser-cut plexiglass.

The flat display surface has the most similarities to the one used in [Jones et al. 2007], though it has a steeper angle to better match the shape of a face and two sides to effectively double the frame rate of the display. The diverging beam of the projector continues to diverge horizontally after reflection by the flat display surface, so that approximately a 20° wedge of the audience area observes some reflected pixels from the projector for any given mirror position as seen in Fig. 7(a). The flat mirror is the simplest to build and calibrate though other shapes can provide more useful optical properties.

The convex display surface is in the shape of a 40° cylindrical arc, curving ±20° over its 20*cm* of width. The convex curve spreads reflected light over 100° of the audience. The benefit of this mirror shape is that the line of reflected light traces over the audience more slowly compared to the flat mirror, since the speed of the angle formed between the display surface and the incident projector light is effectively retarded relative to the absolute rotation of the surface. (If the display surface were a single complete cylinder, the specular reflection would not move at all.) As a result, the convex mirror yields higher angular resolution of the three-dimensional imagery, producing higher-quality 3D stereopsis. However, a convex mirror has several disadvantages. Due to the large angular divergence of the mirror, many projector rays reflect to the far side of the display where they are unseen, while many forward-facing rays can not be reflected into by any mirror angle. The result is a smaller usable volume relative to the mirror size. The missing sample squares in Row 3 of Fig. 6 indicate the points that fell outside the concave mirror's smaller visible volume.

The concave display surface is built into an elliptical shape designed to focus the light of the projector (which, in its unfolded optical path, is 56cm from the mirror surface) to a line 1m away from the center of the display (which is a typical average viewing distance), as seen in Fig. 7(b). The utility of this shape is that at any instant at most one audience member will see the light reflected by the projector. We leverage this mirror shape in the next section so that in the case of tracked viewers, the display can render the proper vertical perspective for each viewer in a straightforward manner using a single $(V_h, V_d)$ per rendered frame.

Different mirror shapes also affect the shape of the display's focal surface. The focal surface for a given viewer is composed of the multiple mirror slices that are illuminated as the mirror spins. For a flat mirror, the focal surface is a cone centered around the mirror's axis. Convex and concave mirrors have asymmetrical focal surfaces that change based on viewing angle. Convex mirrors produce a set of concave focal planes; concave mirrors produce a set of convex focal planes. This represents another advantage of the concave mirror, as the human face is shaped more like a convex cylinder than a concavity. When the shape of the focal surface approximates the object being displayed, accommodation cues are more accurate and aliasing [Zwicker et al. 2006] is minimized.

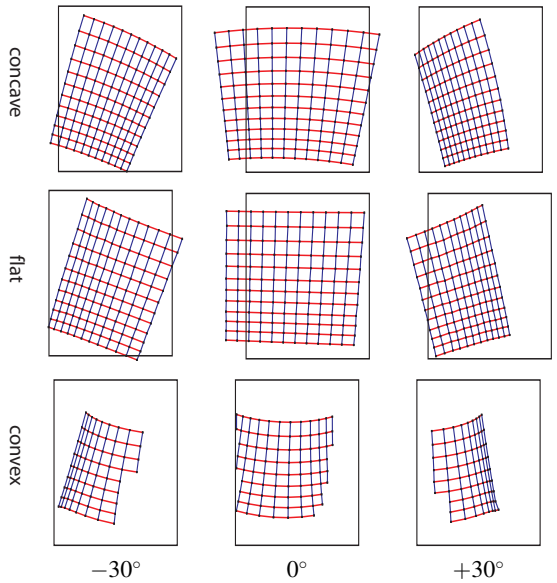



**Figure 6:** *A grid of points on a frontoparallel plane is processed through the 6D lookup table to produce warped geometry displayed on the projector. Three LUTs for three mirror shapes are demonstrated, evaluated for three different mirror angles; the black rectangles show the extent of the actual projector frame.*

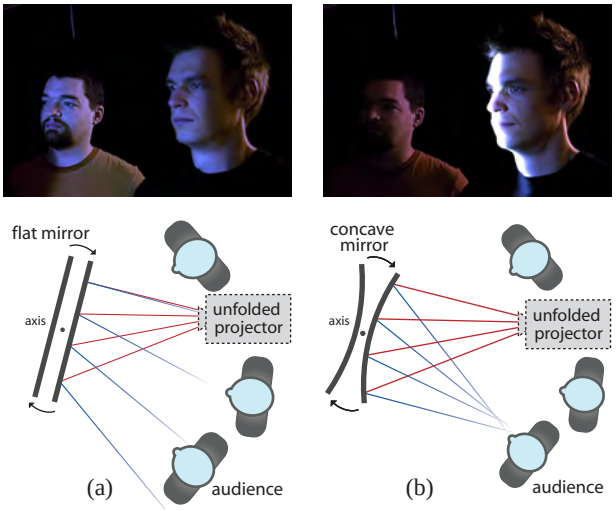



**Figure 7:** *(a) Light diverging from a flat anisotropic display surface can illuminate multiple viewers simultaneously, requiring a single projector frame to accommodate multiple viewer heights. (b) Light reflected by a concave display surface typically projects imagery to at most one viewer at a time, simplifying the process of rendering correct vertical parallax to each tracked viewer.*

## 6 Face Tracking for Vertical Parallax

To provide accurate gaze and eye contact, the rendered face of the remote participant must appear to be rendered correctly into world

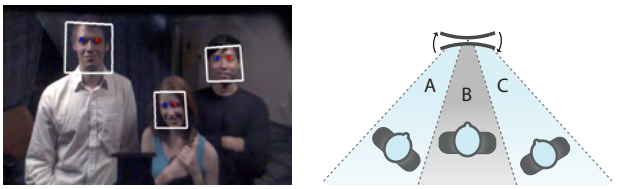
space coordinates as seen by all audience members. Rendering the face for the same viewing height $V_h$ and distance $V_d$ for all audience members can make the face appear to be gazing at an inaccurately high or low angle to some viewers, even though the natural horizontal-parallax of the display will provide generally accurate horizontal perspective to all viewers. Although vertical gaze direction is detected with less sensitivity than horizontal gaze direction [Chen 2002], a true sense of eye contact requires both to be within a few degrees of accuracy. To render vertical perspective accurately to multiple viewers, we track viewer positions in the 2D video feed showing the RP's view of the audience (Fig. 8(a)). We note that [Jones et al. 2007] demonstrated tracked vertical parallax for a single viewer with an active tracking system, but did not employ the correct projection mathematics for our system or handle multiple viewers with a passive tracking system. For our tracking system we use the face detection algorithms in the OpenCV library based on [Viola and Jones 2004] and filter the tracking data using a Kalman filter to reduce noise. The filtered detected face data provides a good estimate of the azimuth and inclination of each audience face relative to the eyes of the RP, though the single camera does not provide a distance measurement. We instead approximate depth based on the size of the detected face. While variation in face size across audience members biases such distance measurements, the visual error is essentially undetectable as it results only in subtle changes to perspective foreshortening; future systems could include a stereo camera pair to more accurately triangulate facial depth. To photograph simulated audience viewpoints in several figures in this paper, we used the Augmented Reality Toolkit [Woods et al. 2003] to track square markers attached to each camera.

When rendering tracked vertical perspectives for multiple viewers, we use the focusing concave display surface so that any one video projector frame can be assumed to address just one of the audience members. For each display surface rotation angle we determine the tracked audience member who is closest to the central reflected ray of the mirror. We then render the face using the lookup table entries corresponding to the height and depth $(V_h, V_d)$ of this closest viewer (Fig. 8(b)). In this way, the display's horizontal parallax provides binocular stereo with no lag as viewers move horizontally, while vertical parallax is achieved through tracking. We believe this is a good approach since it respects the finding of [Chen 2002] that we are more sensitive to horizontal gaze direction than vertical gaze direction, and also since people's body motion is more likely to produce horizontal head translations than vertical ones.

## 7 Results

Figure 10 evaluates the accuracy of the projection technique for three different mirror shapes. We projected a test cube, a real face scan, and a mannequin head scan onto the concave, flat, and convex display surface geometries, each with its appropriately computed lookup table. In the concave case, it was necessary to shrink objects to fit within the smaller usable display volume. The surfaces are seen from two simultaneously tracked camera perspectives at different heights. Despite the changing mirror geometry, the results show the perspective of the displayed objects to be consistent with the camera viewpoints as well as with each other. For a ground truth comparison, we removed the spinning mirror from the display and replaced it with the actual mannequin head. Though the lighting differs, the size and shape of the virtual head closely matches that of the real head. In general, the concave mirror yields the best combination of display volume, user addressability, and focal cues for a head-sized display. The concave mirror is used for all subsequent results in this paper and video unless stated otherwise.

We have not conducted a formal user study to judge the effectiveness of the display as a system for improved teleconference communication but were pleased that subjects using the display consistently reported a sense that the remote participant was able to make eye contact with them. This effect was felt most strongly when the remote participant was looking away, then suddenly turned or glanced towards a particular audience member.



(a) (b)

**Figure 8:** *(a) Faces are tracked in the 2D video feed (b) As the mirror rotates, the head is rendered for the appropriate height and distance of the nearest viewer for correct vertical perspective.*

To provide a quantitative measurement of eye contact accuracy, we captured and transmitted a test object featuring five registration targets that enable angular orientation to be measured (Fig. 9). The testing procedure is as follows: we placed a camera at one end of the teleconferencing system, and placed the test object at the other end of the system. After sighting the transmitted image of the camera lens through the sighting hole of the test object, we photographed the transmitted image of the test object. Then we measured on the photograph how far the apex registration target was from the center of the four edge registration targets. From this deviation we calculated the gaze error in each direction of the system. The measured errors ranged between 3 to 5 degrees on the 3D display, most of which is attributable to geometric noise and the 2.5 degree separation between independent views. For the remote 2D display, the error ranged between 1 to 2 degrees.

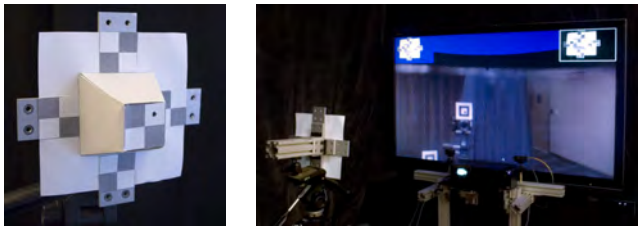


**Figure 9:** *A test object (left) is aimed at a camera shown on the 2D display (right). The camera photographs the transmitted image of the test object to measure gaze accuracy for the 3D display. By switching the locations of the camera and test object, we also measured the accuracy of the 2D display.*

## 8 Future Work

Our experience with our system suggests several avenues for future work to improve the various system components. Improving the quality of the rendered face could be done in several ways. Color could be achieved by placing multiple synchronized projectors in the same optical path, or by building a three-chip high-speed video projector. Grey level reproduction could be improved by applying more advanced halftoning algorithms such as [Ostromoukhov 2001]; however, such algorithms would have to be optimized to run at thousands of frames per second. Also, given that we can render grey levels to the device, it would improve the stereopsis effect to apply the antialiasing technique of [Zwicker et al. 2006] to the rendered imagery. Additionally, it is a drawback that the remote participant does not receive a three-dimensional view of the people in the audience, even though the screen is positioned and calibrated to optimally match the actual audience position. Replacing the 2D video screen with an autostereoscopic binocular display as in [Per-


**Figure 10:** *Comparison of the different mirror shapes for simultaneously tracked upper and lower views. (Row 1) Concave mirror (Row 2) Flat mirror (Row 3) Convex mirror. For the convex mirror, the geometry was scaled by 0.75 to fit within the smaller display volume. In the 4th and 8th columns we replaced the mirror with the actual mannequin head to provide a ground truth reference.*

lin et al. 2000] or [Sandin et al. 2005] could remove this limitation. Furthermore, extending our one-to-many system into one that accommodates any number of remote participates is also of interest. Currently, the 3D display volume can show only one subject, so a meeting involving $N$ subjects in $L$ different locations would require $N \times (L-1)$ head-sized displays. Finally, it would be of interest to conduct a user study to further evaluate the extent to which eye contact has been achieved, and determine whether the system improves the sense of telepresence and effective communication as a result.

## 9 Conclusion

In this work, we have presented a 3D Teleconferencing system able to transmit the face of a remote participant in 3D to an audience gathered around a 3D display, maintaining accurate cues of gaze, attention, and eye contact. To develop this system, we generalized the projection mathematics for projecting three-dimensional imagery onto arbitrarily shaped display surfaces, facilitating the use of a novel concave display surface able to focus the video projector patterns to individual users. We also track the positions of the faces of the audience so that the display can render the correct vertical perspective of the face to each viewer. The result is a teleconferencing system takes a significant step towards maintaining the many nonverbal cues used in face-to-face human communication.

## Acknowledgements

The authors wish to thank David Krum, Monica Nichelson, Richard DiNinni, Scott Fisher, Bill Swartout, Randy Hill, Randolph Hall, and especially John Parmentola for their support, assistance, and inspiration for this work. This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM) and the University of Southern California Office of the Provost. The high-speed projector was originally developed by a grant from the Office of Naval Research under the guidance of Ralph Wachter and Larry Rosenblum. The content of the information does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

## References

ARGYLE, M., AND COOK, M. 1976. *Gaze and Mutual Gaze*. Cambridge University Press, London.

CHEN, M. 2002. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, 49–56.

GEMMELL, J., TOYAMA, K., ZITNICK, C., KANG, T., AND SEITZ, S. 2000. Gaze awareness for video-conferencing: a software approach. *Multimedia, IEEE 7*, 4 (Oct-Dec), 26–35.

GRAYSON, D. M., AND MONK, A. F. 2003. Are you looking at me? Eye contact and desktop video conferencing. *ACM Trans. Comput.-Hum. Interact. 10*, 3, 221–243.

GROSS, M., WÜRMLIN, S., NAEF, M., LAMBORAY, E., SPAGNO, C., KUNZ, A., KOLLER-MEIER, E., SVOBODA, T., GOOL, L. V., LANG, S., STREHLKE, K., DE MOERE, A. V., AND STAADT, O. 2003. blue-c: A spatially immersive display and 3d video portal for telepresence. *ACM Transactions on Graphics 22*, 3 (July), 819–827.

ISHII, H., KOBAYASHI, M., AND GRUDIN, J. 1993. Integration of interpersonal space and shared workspace: Clearboard design and experiments. *ACM Trans. Inf. Syst. 11*, 4, 349–375.

JERALD, J., AND DAILY, M. 2002. Eye gaze correction for videoconferencing. In *ETRA '02: Proceedings of the 2002 symposium on Eye tracking research & applications*, ACM, New York, NY, USA, 77–81.

JONES, A., MCDOWALL, I., YAMADA, H., BOLAS, M., AND DEBEVEC, P. 2007. Rendering for an interactive 360ř light field display. *ACM Transactions on Graphics 26*, 3 (July), 40:1–40:10.

KAJIYA, J. T., AND KAY, T. L. 1989. Rendering fur with three dimensional textures. In *Computer Graphics (Proceedings of SIGGRAPH 89)*, 271–280.

LIU, J., BELDIE, I. P., AND WÖPKING, M. 1995. A computational approach to establish eye-contact in videocommunication. In *in Videocommunication, Int. Workshop on Stereoscopic and Three Dimentional Imaging*, 229–234.

MATUSIK, W., AND PFISTER, H. 2004. 3D tv: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics 23*, 3 (Aug.), 814–824.

MORRIS, E. 2004. The Fog of War: 13 questions and answers on the filmmaking of Errol Morris. *FLM Magazine*. http://www.errolmorris.com/content/eyecontact/interrotron.html.

MUHLBACH, L., BOCKER, M., AND PRUSSOG, A. 1995. Telepresence in videocommunications: a study on stereoscopy and individual eye contact. *Human Factors 37*, 2, 290–305.

MUKAWA, N., OKA, T., ARAI, K., AND YUASA, M. 2005. What is connected by mutual gaze?: user's behavior in video-mediated communication. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, ACM, New York, NY, USA, 1677–1680.

NAYAR, S. K., KRISHNAN, G., GROSSBERG, M. D., AND RASKAR, R. 2006. Fast separation of direct and global components of a scene using high frequency illumination. *ACM Transactions on Graphics 25*, 3 (July), 935–944.

NGUYEN, D., AND CANNY, J. 2005. Multiview: spatially faithful group video conferencing. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, 799–808.

NGUYEN, D. T., AND CANNY, J. 2007. Multiview: improving trust in group video conferencing through spatial faithfulness. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, 1465–1474.

OSTROMOUKHOV, V. 2001. A simple and efficient error-diffusion algorithm. In *Proceedings of ACM SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, 567–572.

OTT, M., LEWIS, J. P., AND COX, I. 1993. Teleconferencing eye contract using a virtual camera. In *CHI '93: INTERACT '93 and CHI '93 conference companion on Human factors in computing systems*, ACM, New York, NY, USA, 109–110.

PERLIN, K., PAXIA, S., AND KOLLIN, J. S. 2000. An autostereoscopic display. In *Proceedings of ACM SIGGRAPH 2000*, Computer Graphics Proceedings, Annual Conference Series, 319–326.

POULIN, P., AND FOURNIER, A. 1990. A model for anisotropic reflection. In *Computer Graphics (Proceedings of SIGGRAPH 90)*, 273–282.

PRUSSOG, A., MUHLBACH, L., AND BOCKER, M. 1994. Telepresence in videocommunications. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, USA, vol. 1, 180–184.

QUANTE, B., AND MUHLBACH, L. 1999. Eye-contact in multipoint videoconferencing. In *Proceedings of the 17th International Symposium on Human Factors in Telecommunication*.

RASKAR, R., WELCH, G., CUTTS, M., LAKE, A., STESIN, L., AND FUCHS, H. 1998. The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *Proceedings of SIGGRAPH 98*, Computer Graphics Proceedings, Annual Conference Series, 179–188.

REES, J. 2008. Critics pan CNN's fake election holograms. *New Zealand Herald* (Nov 7).

ROSE, D. A. D., AND CLARKE, P. M. 1995. A review of eye-to-eye videoconferencing techniques. *BT technology journal 13*, 4, 127–131.

SANDIN, D. J., MARGOLIS, T., GE, J., GIRADO, J., PETERKA, T., AND DEFANTI, T. A. 2005. The varrier autostereoscopic virtual reality display. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Papers*, ACM, New York, NY, USA, 894–903.

SELLEN, A. J. 1995. Remote conversations: The effects of mediating talk with technology. *Human Computer Interaction 10*, 401–444.

STEINMEYER, J. 1999. *The Science behind the Ghost: A Brief History of Pepper's Ghost*. Hahne.

TAGUCHI, Y., KOIKE, T., TAKAHASHI, K., AND NAEMURA, T. 2009. Transcaip: A live 3d tv system using a camera array and an integral photography display with interactive control of viewing parameters. *Accepted to IEEE Transactions on Visualization and Computer Graphics*.

VERTEGAAL, R. 1999. The gaze groupware system: mediating joint attention in multiparty communication and collaboration. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, 294–301.

VIOLA, P., AND JONES, M. J. 2004. Robust real-time face detection. *International Journal of Computer Vision 57*, 2, 137–154.

WILBURN, B., JOSHI, N., VAISH, V., TALVALA, E.-V., ANTUNEZ, E., BARTH, A., ADAMS, A., HOROWITZ, M., AND LEVOY, M. 2005. High performance imaging using large camera arrays. *ACM Transactions on Graphics 24*, 3 (Aug), 765–776.

WOODS, E., MASON, P., AND BILLINGHURST, M. 2003. Magicmouse: an inexpensive 6-degree-of-freedom mouse. In *GRAPHITE 2003*, 285–286.

YANG, R., AND ZHANG, Z. 2004. Eye gaze correction with stereovision for video-teleconferencing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 26*, 7 (July), 956–960.

YANG, J. C., EVERETT, M., BUEHLER, C., AND MCMILLAN, L. 2002. A real-time distributed light field camera. In *Rendering Techniques 2002: 13th Eurographics Workshop on Rendering*, 77–86.

ZHANG, S., AND HUANG, P. 2006. High-resolution, real-time three-dimensional shape measurement. *Optical Engineering 45*, 12.

ZHANG, Z. 2000. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 11, 1330–1334.

ZWICKER, M., MATUSIK, W., DURAND, F., AND PFISTER, H. 2006. Antialiasing for automultiscopic 3D displays. In *Rendering Techniques 2006: 17th Eurographics Workshop on Rendering*, 73–82.