

BiReality: Mutually-Immersive Telepresence

Norman P. Jouppi, Subu Iyer, Stan Thomas, and April Slayden

HP Labs

1501 Page Mill Rd.

Palo Alto, CA 94306

norm.jouppi@hp.com

ABSTRACT

BiReality (a.k.a. Mutually-Immersive Telepresence) uses a teleoperated robotic surrogate to provide an immersive telepresence system for face-to-face interactions. Our goal is to recreate to the greatest extent practical, both for the user and the people at the remote location, the sensory experience relevant for face-to-face interactions of the user actually being in the remote location. Our system provides a 360-degree surround immersive audio and visual experience for both the user and remote participants, and streams eight 704x480 MPEG-2 coded videos totaling 20Mb/s. The system preserves gaze and eye contact, presents local and remote participants to each other at life size, and preserves the head height of the user at the remote location. Initial user experiences are presented.

Categories and Subject Descriptors

H.5.1 Multimedia Information Systems; I.3.2 Computer Graphics - *Graphics Systems*; H.5.5 Sound and Music Computing

General Terms

Human Factors, Design, Experimentation

Author Keywords

Computer-supported collaborative work, video mediated communication, video conferencing, audio conferencing, spatial audio, robotics.

1. INTRODUCTION

We have developed a mutually-immersive telepresence system we call BiReality. Our goal is to recreate to the greatest extent practical, both for a user and people at a remote location, the sensory experience relevant for face-to-face interactions of the user actually being in the remote location. We call the system BiReality since its goal is to create two compelling copies of the real world, one at the remote location minus the user and one at the user's location minus the rest of the remote environment.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'04, October 10-16, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-893-8/04/0010 ...\$5.00.

In BiReality, immersive face-to-face interactions are facilitated by providing the following features:

- Wide visual field
- High resolution visual field
- Both remote participants and users appear life size
- Gaze preservation (people know where people are looking)
- Remote colors are perceived accurately
- High-dynamic range audio
- Directional sound field
- Head height is preserved

In contrast, traditional video conferencing typically provides only a single video stream, and this stream is incapable of providing both the full field of view of normal human vision and the resolution of the human visual fovea at the same time. Only one audio channel is provided, with only a limited dynamic range and no directional information. Current commercial video conferencing does not preserve gaze or present users' faces at life size. However experiments have shown that presenting users at life size increases immersion for people interacting with them[16].

1.1 Relationship to Previous Work

There are a number of previous and ongoing projects that are related to our work. Buxton[5] describes a number of early telepresence efforts. These include Hydra[18], in which each person was represented as a small display and a camera at a meeting. Paulos and Canny's Personal Roving Presence[15] allows a webcam user to maneuver through a remote environment using either a helium-filled blimp or a platform based on a radio-controlled toy car. Jouppi [9] presented a mobile telepresence surrogate with a 150 degree field of view of a remote location. BiReality differs from this previous work in that the user of our system interacts with people and objects in arbitrary real settings in a way that is much more immersive for both the user and the people they are visiting.

Recently much work has been done in addressing gaze preservation in virtual or desktop environments[7, 2]. Video-tunnels[4], FreeWalk[12], MAJIC[14], GAZE[19], GAZE-2[20], as well as other systems are examples of *symmetric* conferencing systems. They are symmetric in that each participant must use a copy of the same system to communicate



Figure 1: A BiReality surrogate standing. Note the far two sides of the user’s head are visible in the mirror.

with other people using the system. In contrast, BiReality is an asymmetric physical system. Any number of interaction participants may use a BiReality surrogate instead of being actually present at the physical interaction site.

2. INTRODUCTION TO BIREALITY

A BiReality system consists of two parts: a display cube at the user’s location and a surrogate in the place of the user at the remote location. The BiReality surrogate is shown at a standing height in Figure 1 and at a sitting height in Figure 7. The surrogate’s head displays live video of the user’s head from four sides. The display cube provides a complete 360-degree surround view of the remote location. Figure 2 shows a user in the display cube.

In order to present the user’s head in a natural appearance at the remote location, we are precluded from requiring the user to wear any technological device at shoulder level or above. This includes the use of headphones or head-mounted displays, and has implications for the system design as will be described later. The only technological device worn by the user is a wireless lapel microphone.

2.1 360-degree Surround Video

A key advantage of a 360-degree surround projection environment is that it allows the user to rotate locally in the display cube, and can eliminate all teleoperated mechanical rotations of the surrogate at the remote location. An overhead schematic view of the display cube is shown in Figure 3. In each corner of the display cube we have multiple cameras for capturing the view of the user for display on the surrogate’s head. In keeping with the 360-degree surround of the user, we also maintain a 360-degree surround view of the user on the head of the surrogate. Thus, when the user turns in place in the display cube, the video of their head from the four corners of the display cube will show them rotating at the remote location, without the need for any

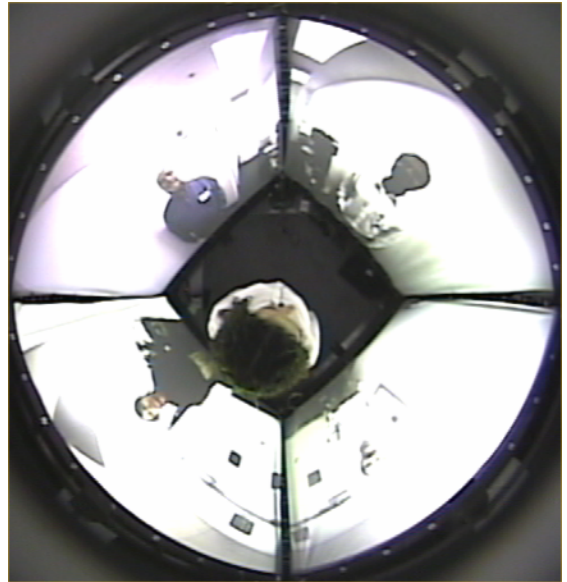


Figure 2: An overhead fisheye view of a user in the display cube.

teleoperated mechanical motion. To enable the user to repeatedly turn in place while using the system, we avoid any wired tether between the user and the rest of the system.

Figure 4 shows a cross section of the display cube. Unlike most display cubes which have floor to ceiling screens, the screens in our display cube are from 0.95 to 2.12 meters above the floor. The bottom of the display screen corresponds roughly to a desk or table height. We do not reproduce the remote location below this height since relatively little business is transacted under tables or desks in business settings. Similarly, the screens do not extend significantly above the standing height of the user since office ceilings are not typically used in business interactions either.

The display cube is intentionally designed to be relatively small by virtual reality standards, and provides a horizontal space of about 1.3 meters on a side for the user to move within. (In contrast, virtual reality CAVEs are often about three meters on a side.) When the user is in the BiReality display cube, the distance to the display cube walls are roughly on the order of the personal space of 0.5 to 1.3 meters that users expect in interpersonal interactions[8]. Besides being at a scale conducive to interpersonal interactions, this display cube size has two additional advantages. First, the display cube and projectors are easier to site since they have a smaller footprint. Second, if a larger cube was used, people sitting or standing close to the surrogate would need to be displayed at many times larger than real life on the display cube screens to subtend the correct angle for the user. We have found that many people experience an uncomfortable sensory dissonance when people are displayed at much larger than life near them, and find it somewhat menacing. Having a display cube size on the order of twice the average interpersonal interaction distance solves these problems.

In traditional display cubes a user wears headphones, but in our application we cannot obstruct the view of the user’s

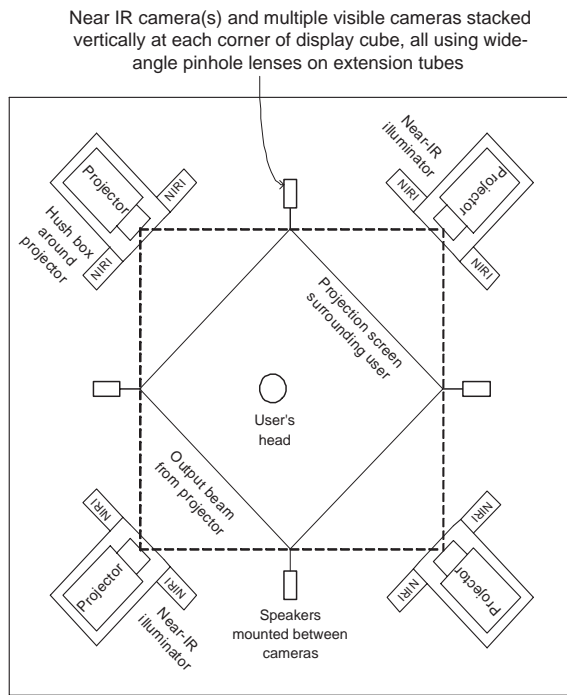


Figure 3: An overhead schematic view of a user in the display cube.

head and shoulders with any technological device. Unfortunately any sounds in a small 360-degree surround display cube with parallel walls reverberate for a very long time, making it hard to understand speech and nearly impossible to perceive directionality of sounds. Instead we use trapezoidal screens and angle them outward at the top. At the top of the display cube we have a layer of sound-absorbing foams and an anechoic ceiling. The angle of the display screens reflects sound up to the foams and ceilings where it is absorbed. This enables us to recreate the sound field of the remote location in great detail, even providing the ambiance of the remote location.

The user's face is lit on all four sides only by a light field from the remote location, via the cameras and projectors. Thus if a user is facing a bright window with a dark room behind them, the front of their face will appear to be brightly lit on the the surrogate display facing the window, and the other sides of their head will be darker on the other sides of the surrogate. In order to make the user look as natural as possible, we would also like to recreate the light fields from above and below the head of the surrogate for the user. Due to the need for sound absorption above the display cube, placing another screen above the display cube is not an option. (It is also a relatively expensive way to light the top of a person's head.) Instead we use a fill light, as is common in professional photography. The sound absorbing foams are designed with a square hole in the center to act as a mask for the fill light, preventing direct illumination from the fill light from impinging on the screens and washing out the projected imagery. The mask in the acoustic foam also provides an outlet for the convection of display cube air, silently preventing the display cube from getting hot and stuffy. Fi-

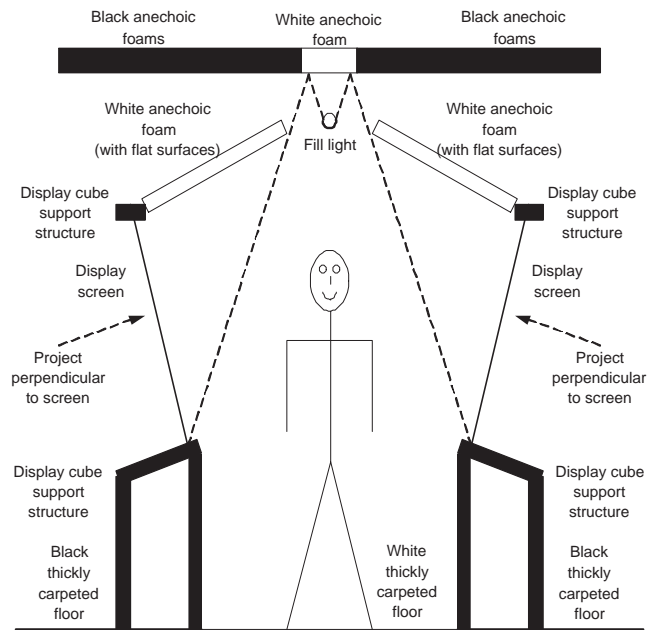


Figure 4: A cross-section schematic view of a user in the display cube.

nally, to provide a soft diffused fill light, we reflect the light off of the white ceiling in the room.

The user is free to move their head within the display cube over a range of about one meter. This prevents the user from feeling cramped and makes the system more natural for the user. However, this requires us to track the position of the user's head so we can provide a cropped life-size view of their head on the head of the surrogate. Moreover, since we project video on all four sides of the user's head, tracking by chromakeying or difference keying in the visible spectrum is not possible. However our projectors only project in the visible spectrum, and do not output in the near infrared. Thus we have developed a system using near-infrared difference keying to track the user's head. The display cube and user are evenly lit by near infrared illuminators (NIRI) mounted on the projector hush boxes, and this illumination is diffused by the screens. Since the NIR illumination is relatively constant, difference keying between NIR video of the user in the cube with a baseline image taken without a user lets us reliably track the user with relatively little computation. Next we translate the head position from NIR tracking into the image space of the color cameras. This is used for cropping and scaling of the color video of the user's head for display on the head of the surrogate. The system allows the user to place their head almost anywhere in the display cube and still keep a life-sized color view of their head on the head of the surrogate. Furthermore, the head tracking system automatically reduces the scale of the user's video to include gestures made by a user near his or her head.

The display screens showing the remote environment are in the background of all color views of the user's head. Thus when the user's head is cropped into a classic portrait style for display on the head of the surrogate, objects behind the head of the surrogate at the remote location are displayed behind the head of the user on the surrogate. We call this



Figure 5: User head tracking is complicated by the projected video of people at the remote location. Note the top “E” of a Snellen visual acuity chart at the remote location is visible to the left of the user.

feature remote backdrops, and it adds significantly to the immersion of the remote participants. Remote backdrops provide a visual effect similar to what would exist if the user was really part of the remote environment. In our system we do not attempt to segment the user’s head from the background (since this is a source of artifacts), so we obtain a clean presentation of the user’s head on the displays of the surrogate.

2.2 Preservation of User Head Height and Posture

People often stand in informal conversations or when presenting to an audience, but typically sit for extended discussions or when viewing presentations. In our system we automatically preserve the head height of the user at the remote location. When tracking the position of the user’s head in the display cube, we triangulate their head position to compute the height of the top of their head above floor level in the display cube. The surrogate is built with an extensible torso, and its extension is servoed so that the top of the user’s head is displayed on the surrogate at approximately the same height as the top of the user’s head at the user’s location. Thus when the user stands up, the surrogate will also stand up, changing the perspective seen by the user as well as the people viewing the user via the surrogate. Similarly, the user can sit down at the remote location merely by sitting down in the display cube. The preservation of the user’s height and sitting/standing posture is accurate to within plus or minus 2.5cm over the range of surrogate motion. Due to limitations of the linear actuator currently used in the surrogate, the range of motion is presently limited to 0.5 meters.

2.3 Simultaneous Preservation of Height and Vertical Gaze

In order to preserve gaze in the vertical direction independent of the user’s head height several additional techniques are required.

First, to capture a view of the user’s head from the proper angle for display on the head of the surrogate, we capture multiple views of the user’s head at different vertical heights from each corner of the display cube. We then select between them to use the view that is closest to the user’s eye level, based on near-infrared head tracking. We currently use four cameras in each corner of the display cube, roughly corre-

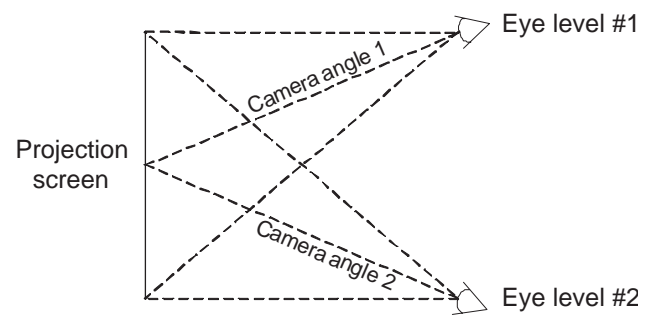


Figure 6: Cameras in the head of the surrogate tilt with the height of the user’s head to maintain the metaphor of a window onto the remote location.

sponding to the eye level of a tall person standing, a short person standing, a tall person sitting, and a short person sitting. By selecting among the multiple views we can present an eye-level view of the user’s head on the head of the surrogate. This is the view people expect to be presented with in other media such as television.

Second, as the user stands up and sits down their expected perspective of the remote location changes. The appropriate metaphor in this case is that the screens of the display cube form a set of windows into the remote location. If a user looks out of a window with their eyes level with the lower window sill, they will have a different perspective than if they look out of the window with their eyes level with the top window sill. Specifically, if they look out at the exact height of the bottom of the glass as in Figure 6, they will be able to see things on the other side of the window above this level, but objects below this level will be occluded (unless they stand up). To mirror this functionality and better preserve gaze in the vertical direction, the cameras in the surrogate’s head tilt up and down based on the user’s head height. Thus when the user’s head is level with the bottom of the projected area of the screen, the camera will be tilted up such that the bottom of its field of view is parallel to the ground. The video received from the cameras is also warped as a function of the camera angles to reduce overlaps and prevent gaps from occurring where the video streams abut in the corners of the display cube.

The details of preservation of vertical gaze fill more than a paper in themselves, but we give a brief overview here. Imagine as above that the user’s eyes are level with the bottom of the projected area in the display cube. In this case if a remote participant’s eyes are at the same level as the user’s, they will appear to look at each other when they look at each other’s images on the display cube and surrogate displays. Similarly, if the remote participant is standing while the user is sitting in the display cube, the tilt of the cameras results in the standing remote participant’s head being projected at a higher vertical level than that of the user, so that the user will be seen looking up while the remote participant will be seen looking down. Equivalently, if both local and remote participants are standing at the same height, their eyes will appear at roughly the same level on the screens of the surrogate, display cube, and in real life. Although the vertical preservation of gaze is not perfect, it is usually close in practice.



Figure 7: A surrogate in use in a noisy cafeteria. Conversations in this environment would be difficult or impossible without multichannel audio.

2.4 Surrogate Anthropomorphism

We strove to design the surrogate to roughly an anthropomorphic form factor so that interactions with remote participants could occur at common interpersonal distances. The surrogate only requires 2 PCs, and this allows us to fit it in a roughly personal form factor. The surrogate also has a skin with a pleasing industrial design.

To make the surrogate more anthropomorphic, we shaped it in a very simplified form of a human torso. The upper portion, representing the chest, shoulders, and arms, is wider than the lower portion which represents a person's legs. The PCs are contained entirely in the lower portion of the torso, along with an electrically-operated linear actuator. This enables the upper wider portion to be lowered over the lower portion when the user is sitting. The circular base on the bottom of the surrogate is intended to evoke the analogy of a personal hovercraft or flying saucer. The circular plastic base ring hides the wheels of the surrogate. The skin of the surrogate is fabricated in plastic without any sharp corners, just as people don't have sharp corners.

The colors of the surrogate were carefully chosen. A blue color suitable for business clothing was chosen for the skin color. (The color is brighter than would typically be worn by a man in business suit, but is more in keeping with what a woman might wear.) The circular base was painted in silver metallic paint in keeping with the personal hovercraft theme. The non-LCD portions of the head were painted in black to make it appear thinner and to not capture the attention of viewers, helping to focus attention on the user's head on the LCD panels. This is similar to the use of black clothing by A/V support people in professional stage venues.

2.5 Surrogate Head

Because the user can rotate their head in the display cube, large rotations of the surrogate head do not need to be provided, simplifying its design. Also, tilting the head up or down does not make much sense when there is no front or distinction among any sides of the head. Also, given the 360 degree surround projection environment, tilting the view seen by the user could also be disconcerting. Thus, the surrogate head does not tilt. In order to avoid attracting attention to the cameras, microphones, and speakers in

the surrogate head, we have spaced out the LCD panels by about 6cm and placed the cameras and speakers in the gap between the panels. The cameras are placed at the level of the user's eyes when they are displayed on the head of the surrogate. This improves the preservation of gaze in the vertical direction. The speakers output through a grille at the bottom of the gap. The LCD panels used in the surrogate have a 38cm diagonal. Thus we can present the user's head close to their actual head size on the surrogate. Microphones are placed at the top of the LCD panels.

2.6 Audio Issues

A number of audio issues are important in the design of the surrogate. To minimize PC case fan noise in the surrogate we have designed it so that all airflow exhausts down through the interior of the base ring. This prevents any direct path to remote participants for fan noise. The gap between the upper and lower portions of the surrogate torso serves both as an air intake and isolation to reduce direct transmission of fan noise. Carpeting in typical modern office environments also tends to attenuate the noise level of sound reflected off the floor. We have also taken care to minimize the fan noise from various sources in the PCs themselves, such as the CPU fan, chipset fan, and graphics accelerator fan.

We have implemented electronically-controlled directional audio output on the surrogate. Directional audio output has many uses. For example, directional output enables the user to whisper in a remote participant's ear. There is no difference in hardware between the four sides of the surrogate's head, and the front is merely the side currently displaying the front of the user's head. In order to implement directional audio output, we track the orientation of the user's head in the display cube (also via near-infrared difference keying). We then vary the volume of each of the speakers in the four corners of the surrogate's head as would be the case if the user were physically present and speaking while rotating their head. More details on the audio system can be found in [10].

2.7 ROI Editing of Compressed Streams

The captured video of the user's head is cropped to a pleasing head-and-shoulders portrait view for display on the head of the surrogate. The freedom of motion provided to the user in the display cube implies that large portions of the view of the user inside the display cube will need to be cropped out before display. If this cropping is performed on the surrogate, the bandwidth required to transmit the cropped portions will be wasted. Instead we crop the video in the compressed domain based on coordinates from the head tracking system before transmitting it to the surrogate. This can reduce the bandwidth required for the user's head by approximately 50%[1].

2.8 Head-Track Bandwidth

Although video is projected all around the user in the display cube, at any point in time they can only view about a 180 degree field of view without moving their head. Since the bandwidth required for 360 degree surround projection is significant, and it takes time for users to turn their heads around, we save bandwidth from the surrogate to the display cube by devoting more bandwidth to areas in front of the user's face. The orientation of the user's head is determined

via the near-infrared head tracking system, so that the user is not required to wear any tracking-related device on their head.

We do not attempt to use eye-tracking techniques to restrict the high-resolution area of the screen to only the portion of the screen currently in the user’s fovea. This is because the peak saccadic speed of a person’s eye movements can approach 600 degrees per second[11], while the latency of the eye tracking, communication to the surrogate, synchronization with compression, transmission back to the surrogate, decoding, and display can take more than a second. Thus system bandwidth reduction by eye tracking would significantly lag the speed of eye movements. In contrast, natural and comfortable head motions are an order of magnitude slower. So instead we display in high resolution the entire visual field comfortably viewable with the user’s fovea given a particular head orientation.

There are a number of means of reducing the transmitted bandwidth. Given the MPEG-2 compressed nature of the video, the easiest method and the one we have implemented is to reduce the bandwidth of the video by skipping frames. We use a IPP GOP length of 6. Skipping all the P frames and transmitting just the I frames results in a bandwidth reduction of around 50%, while deleting all P frames plus 4 out of 5 I frames results in a reduction of 90%. Deletion of frames is easy and requires very few CPU cycles on the surrogate.

If the user is facing within 22.5 degrees of a display cube corner, we transmit both streams adjacent to the corner at full bandwidth and reduce the streams in back to 10% of the original bandwidth by deleting all P frames plus 4 out of 5 I frames. If the user is facing within 22.5 degrees of the center of a display cube screen, we only transmit the single screen they are facing at full bandwidth, and reduce the bandwidth to their sides and back. In this case we reduce the bandwidth of the two sides by 50% by deleting all P frames, and reduce the bandwidth behind their head by 90% by deleting all P frames plus 4 out of 5 I frames. In both of these scenarios, the overall bandwidth from the surrogate to the display cube is roughly halved.

This halving of display bandwidth has relatively little effect on the immersion of the user, since the area of their attention is still usually at full bandwidth. It is possible to jerk your head quickly around and notice that the image behind you is frozen for a moment before the latency of the head tracking system and the video transmission and decoding adjusts to the new head orientation. With 30fps source video, the frame rate behind a user’s head can be reduced to 1fps. Ending frame deletion and resumption of video at higher frame rates has to be synchronized to the presence of an I frame in the source stream, so this can result in some additional delay. In the case of 30fps video with a GOP length of 6, this can introduce an additional delay of up to 5 frames, or 167 milliseconds.

2.9 Parallax Issues

In BiReality, we need to acquire a 360 degree horizontal and 73 degree vertical view at the user’s eye level while simultaneously displaying their head at camera level. The displays of the user’s head are not simultaneously emissive and transparent (or time multiplexed). This coupled with the need to tilt the field of view of the cameras to maintain vertical gaze means that we have to use cameras without a

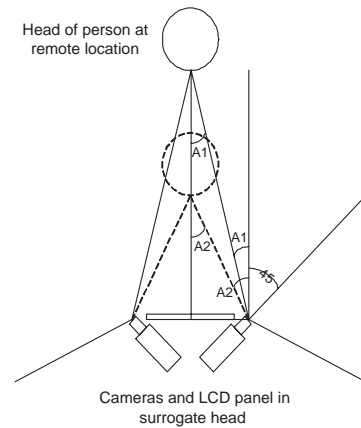


Figure 8: Reduction of parallax artifacts by variable cropping and warping dependent on distance to objects in overlap area.

common optical axis, and hence there is parallax between the four views acquired from the corners of the surrogate’s head. Each camera in a corner of the surrogate’s head captures a horizontal field of view of about 120 degrees, so that image information from people or objects standing up to 65cm from an LCD panel of the surrogate’s head are not lost. However, without correction this can result in more distant objects being duplicated. This is illustrated in Figure 8.

Since the view projected on each side of the display cube comes from a single camera and compression card, this relegates all parallax artifacts to the corners of the display cube. Because there are already artifacts in the corner due to the seam in the screen and the pinhole cameras, this minimizes the visibility of parallax artifacts to the user.

Optionally, distance sensors can be mounted between the cameras below the displays of the user’s head (see Figure 8). These distance sensors can be used to measure the distance to the nearest objects in the area of field of view overlap between the cameras. Then vertical strips of the overlap area can be cropped during projection to minimize the duplication of relatively distant objects without losing image information from people or objects near the surrogate. This can result in a more immersive experience for the user.

2.10 Image Quality

Each of the cameras in the surrogate and display cube outputs S-video. This produces sharper images and eliminates a class of color artifacts present in composite video. The eight video streams in the system are compressed into 704x480 MPEG-2 15fps streams each using 2.5Mb/s. In practice this yields a visual quality roughly equal to 20/150 vision.

3. USER EXPERIENCES

We have not yet performed extensive formal user studies of our telepresence system. However, invariably when a new user first enters the display cube, they have a big smile and a look of wonder in their eyes, since they have never experienced anything quite like it before, and the experience is very enjoyable. People in the remote environment tend to

react with surprise when they first see someone using a surrogate, but after a short period of joking about (which serves as a useful “icebreaker”), people at the remote location interact with the remote person via the surrogate largely as if they were actually physically present.

3.1 Display Cube

Besides providing a 360-degree field of view horizontally, the display cube provides a vertical field of view of more than 73 degrees for a user sitting in the center of the display cube. Together these provide a very strong feeling of immersion for the user. The freedom of motion provided to the user and the lack of any tethering are also valued as significant benefits.

3.2 Form Factor and Height Preservation

The anthropomorphic surrogate form factor is a big help in enabling conversations at conventional interpersonal distances. The ability to automatically adjust the surrogate height based on the user’s height and sitting/standing position is also a key feature. Since the current design is still too short for some standing people and too tall for some sitting people, we have designed a new surrogate with a height adjustment range of 0.67 meters (0.17 meters more than the current design).

3.3 Audio

The multichannel high-dynamic range directional sound system is universally impressive, and adds significantly to the level of immersion for the user. The direction of speaking participants during meetings is clearly discernible without visual stimuli due to the multi-channel capabilities of the system. It is also possible to focus on one speaking person in a room full of speaking people[3], enabling selective attention to and participation in parallel conversations at the remote location. We provided a button on the audio control joystick to switch between a spatial audio presentation and monoaural sound. Users reported dramatic increases in intelligibility of conversations via spatial audio versus telephone-equivalent sound during tests in the noisy cafeteria of Figure 7.

3.4 Video

Probably the biggest negative comment from users concerns the latency of the current system. One-way latency of the video is almost 700ms, so it is very noticeable. We are actively working on reducing this latency, but some amount of video latency with compression is inevitable. At least the audio and video are currently well synchronized, and users have reported that this aids in understanding speech in noisy environments with multiple simultaneous remote conversations. We hope that the next generation of video compression cards will have reduced latency. Eventually CPU-based codecs should be able to handle multiple bidirectional video streams at full resolution and frame rates with low latency[6].

3.5 Gaze Preservation

Gaze preservation in the BiReality system is a very challenging problem since the user has the freedom move around the display cube, including sitting down and standing up. The combination of switching between multiple cameras at different vertical heights in each corner of the display cube

coupled with tilting cameras in the head of the surrogate all controlled by tracking the user’s eye level works well to preserve gaze in the vertical direction. Horizontal gaze is best preserved for the user as seen by remote participants when the user is looking into the cameras in the corner of the display cube, and is sloppier when the user is looking at the center of a screen. In order to encourage the user to spend more time facing the corners we are working at reducing the gap between the screens in the corners and reducing parallax artifacts in the corners.

4. CONCLUSIONS

BiReality can bring immersive telepresence to ordinary public places. It leverages technologies such as computer graphics hardware and the internet which are rapidly increasing in capability and decreasing in cost. Initial user feedback on our prototype system has been very favorable.

Our BiReality system simultaneously affords both freedom of motion for the user in the display cube and a highly immersive experience. By providing a 360-degree surround environment at both the user’s and remote locations, the user can perform all rotations locally merely by rotating their own body. This eliminates a need for teleoperated remote rotation, and makes the experience much more immersive. The surrogate automatically adjusts its height to match the user’s displayed height with their actual height in the display cube. This allows users to stand or sit at the remote location merely by standing or sitting in the display cube. We preserve gaze even as the user is standing up or sitting down in the display cube through the use of vertical camera diversity in the display cube, tilting cameras in the head of the surrogate, and image warping. We vary the bandwidth of the video being presented around the user based on near-infrared video tracking of the orientation of the user’s head. This provides the highest quality video for foveal portions of the user’s vision, while requiring less bandwidth for peripheral areas and areas behind their head. By providing near CD-quality audio and four directional channels, the audio subsystem enables users to whisper back and forth with individual remote participants, accurately auralize the location of remote sound sources, and utilize the cocktail party effect to improve the intelligibility of remote speakers.

In summary, we believe we have created a compelling immersive face-to-face telepresence system. However much work remains to refine the system and formally evaluate its strengths and weaknesses. The system also provides an interesting vehicle for future research into many human factor issues, such as the role of head height in human interactions.

Acknowledgements

The authors would like to thank Jacob Augustine, Shivarama Rao K., and Deepa Kuttippambal for their help with the BiReality software, and Wayne Mack for his help with the construction of the surrogate and display cube.

5. REFERENCES

- [1] J. Augustine, S. R. Kokrady., N. P. Jouppi, and S. Iyer. Region of Interest Editing of MPEG-2 Video Streams in the Compressed Domain. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2004.

- [2] H. H. Baker and et. al. Computational and Performance Issues in Coliseum Videoconferencing. In *Proc. ACM Multimedia*, 2003.
- [3] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 2nd edition, 1997.
- [4] W. A. Buxton and T. P. Moran. EuroPARC's Integrated Interactive Intermedia Facility (iiif): Early Experience. In *IFIP WG 8.4 Conference on Multi-User Interfaces and Applications*, pages 11–34, 1990.
- [5] W. A. S. Buxton. Telepresence: Integrating Shared Task and Person Spaces. In *Graphic Interface 1992*, 123–129.
- [6] M. Chen. A Low-Latency Lip-Synchronized Videoconferencing System. In *Proc. ACM CHI 2003*, 465–471.
- [7] M. Chen. Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconference. In *Proc. ACM CHI 2002*, 49–56.
- [8] E. T. Hall. *The Hidden Dimension*. Anchor Press, 1990.
- [9] N. P. Jouppe. First Steps Towards Mutually-Immersive Mobile Telepresence. In *Proc. of the ACM Conference on Computer Supported Cooperative Work*, pages 354–363, 2002.
- [10] N. P. Jouppe and S. Iyer. A Headphone-free Head-tracked Audio System for Advanced Audio Telepresence. In *Proc. the 117th Audio Engineering Society Convention*, October 2004.
- [11] M. W. Matlin and H. J. Foley. *Sensation and Perception*. Allyn and Bacon, 4th edition, 1997.
- [12] H. Nakanishi, C. Yoshida, T. Nishimura, and T. Ishida. FreeWalk: A 3D Virtual Space for Casual Meetings. *IEEE Multimedia*, 6(2):20–28, April 1999.
- [13] B. O'Conaill, S. Whittaker, and S. Wilbur. Conversations Over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication. *Human-Computer Interaction*, 8:389–428, 1993.
- [14] K. I. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In *Proc. ACM CSCW 1994*, 385–393.
- [15] E. Paulos and J. Canny. PRoP: Personal Roving Presence. In *ACM CHI 1998*, 296–303.
- [16] A. Prussog, L. Mühlbach, and M. Böcker. Telepresence in Videocommunications. In *Proc. the Human Factors and Ergonomics Society 38th Annual Meeting*, pages 180–184, 1994.
- [17] A. Sellen. Remote conversations: The effect of mediating talk with technology. *Human-Computer Interaction*, 10:401–444, 1995.
- [18] A. Sellen, B. Buxton, and J. Arnott. Using Spatial Cues to Improve Videoconferencing. In *Proc. ACM CHI 1992*, 651–652.
- [19] R. Vertegaal. The GAZE Groupware System: Mediating Joint Attention in Multiparty Communication and Collaboration. In *Proc. ACM CHI 1999*, 294–301.
- [20] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung. GAZE-2: Conveying Eye Contact in Group Video Videoconferencing Using Eye-Controlled Camera Direction. In *Proc. ACM CHI 2003*, 521–528.