First Steps Towards
Mutually-Immersive Mobile Telepresence

Norman P. Jouppi
HP Labs
Palo Alto, CA 94304
+1-650-857-2268
norm.jouppi@hp.com

ABSTRACT

Mutually-Immersive Mobile Telepresence uses a teleoperated robotic surrogate to visit remote locations as a substitute for physical travel. Our goal is to recreate to the greatest extent possible, both for the user and the people at the remote location, the sensory experience relevant for business interactions of the user actually being in the remote location. The system includes multi-channel bidirectional video and audio on a mobile platform as well as haptic feedback. This paper describes our first system prototypes and initial experiences using them.

KEYWORDS: Video Conferencing, Audio Conferencing, Human Visual Perception, Multimedia, Multi-User Networked Applications, Robotics, User Interface Hardware, Haptics, Multi-Channel Audio.

INTRODUCTION

We are developing Mutually-Immersive Mobile Telepresence as an alternative to business travel. The business travel we would like to replace ranges from commuting through crowded urban areas (e.g., Los Angeles freeways) to airline travel across the continent or around the world.

Several technologies have been proposed as alternatives to business travel. These include technologies such as audio conference calls and video conferencing. While these techniques are being used, the amount of actual physical travel for business purposes has continued to increase[1]. Why have alternatives such as current video conferencing technologies not replaced more business travel? One hypothesis is that it is because such technologies are *not immersive*. What are some of the aspects of physical business travel that make experiencing remote locations immersive?

- Wide visual field
- High resolution visual field
- Gaze is preserved (people can make eye contact)
- Both remote participants and users appear life size[1]
- Remote colors are perceived accurately
- High-dynamic range audio
- Directional sound field
- Mobility at the remote location
- Ability to manipulate remote objects

In contrast, users of traditional video conferencing are limited to a single camera position. Typically a single video stream is provided, 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 user's faces at life size. However experiments have shown that presenting users at life size increases immersion for people interacting with them[16].

Mobility at the remote location is a key enabler of casual meetings. Casual meetings have been documented as important for effective collaboration, communication, and innovation[10, 19]. Office floor plans of research labs are often designed with this in mind, providing casual open seating areas and informal meeting areas such as kitchenettes. Discussions in hallways and during outings are often a key factor in the success of conferences and offsites. None of these are supported by commercial video conferencing. In recognition of the importance of unplanned interactions, numerous previous experimetal systems have attempted to facilitate unplanned interactions in cycberspace. Examples of these include Cruiser[17], FreeWalk[14], and Piazza[8]. In contrast, we are interested in facilitating such interactions in traditional physical spaces but at a distance.

In Mutually-Immersive Mobile Telepresence our goal is to provide all of the above benefits of physical travel in an immersive way without physical travel and at lower cost. This goal has a number of implications. First, as stated previously, the interface must be mutually immersive. This means that we cannot use technologies such as head-mounted displays,

[1] We define people or objects to be "life size" if they subtend the same horizontal and vertical visual angles when seen remotely as if they were physically present in the position of the surrogate.

which obscure the face of the user. Furthermore, careful lighting of the user's face is important, so that their appearance at the remote location is natural. Second, the system must be unencumbering for widespread adoption and natural operation. This means motion capture harnesses or confining devices should not be worn by the user. Third, due to limitations of current 3-D display technologies (i.e., such as requiring the user to wear glasses or providing a limited field of view), we do not provide stereoscopic video. However, we feel that this is not a significant limitation in our application. Humans rely on more than ten perceptual methods for depth discrimination [11], of which binocular vision is one. Binocular vision is most important for hand-eye coordination and interaction with objects within grasping distances. However, the majority of the content in a business visit is beyond arms length, such as viewing slide presentations or engaging in discussions with colleagues across a conference table. Unlike other applications, such as teleoperated space station assembly, we do not plan to support teleoperated dextrous manipulation; rather the manipulators we provide are primarily for pressing buttons at the remote location. Fourth, we desire the system to be widely usable by people in all walks of life, not just by researchers or “geeks”. Fifth, the operation of the system should be under the conscious control of the user, just as a user's body is under their conscious control. Thus, as an example, we are not interested in systems that move the remote head based on the gaze of the user, since people sometimes move their eyes unconsciously while in thought, and they move their eyes independently of their head.

There are a number of previous and ongoing projects that are related to our work. Buxton[3] 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. Yamaashi, et. al.[20] describe a system called “Extra Eyes”, using both a wide angle camera and a foveal camera. The foveal image is displayed on a screen below the wide angle image, and a box is drawn in the wide angle image showing the position of the foveal image. Paulos and Canny's Personal Roving Presence[15] allows the user to maneuver through a remote environment using either a helium-filled blimp or a platform based on a radio-controlled toy car. Both devices carry small video monitors for the display of the user's face and a single video camera for acquiring images to be sent to the user. Both devices are controlled through an interface built into a web browser. Our work 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.

FIRST GENERATION SYSTEM IMPLEMENTATION

The system consists of a teleoperated robotic surrogate at a remote location and the user at either a compact low-cost user station or an immersion room. To complete construction of a first-generation prototype system quickly with a small number of people, we have used off-the-shelf components wherever possible. The following sections describe the current status of our prototype system.

Model 1 Surrogate

Figure 1 shows an early version of the model 1 surrogate in use. We have designed the surrogate to be approximately the form factor of a sitting person. This overall size is dictated by the size of the four PC systems it contains, each one with five or more PCI/AGP/ISA cards. The surrogate consists of a base, mezzanine level, and a head. The head height shown in the figure works well in meetings (higher head heights cause too much of an obstruction for other remote attendees). It also works well when moving, since the higher the head the more unstable the platform becomes.

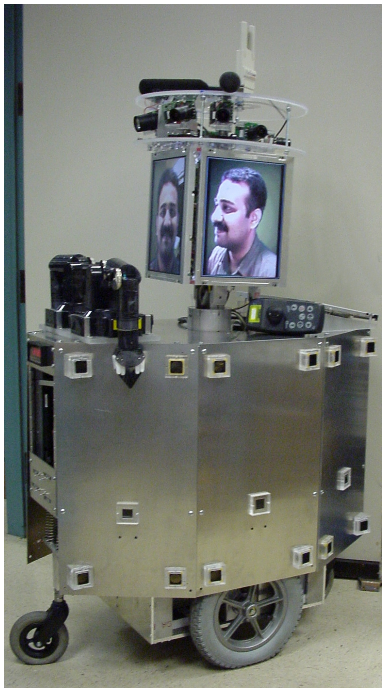

Figure 1: An early version of the surrogate in use.

Base

The base of the unit contains 2.4 Kilo Watt-Hours of batteries. This battery capacity is currently enough to power the surrogate for three hours. More than ninety percent of the power is consumed by the electronics, so as technology scales with Moore's Law [2] the battery life will increase.

The base also contains electric wheelchair motors, motor controllers, actuator controllers, and wheels. We have chosen a configuration with symmetric drive wheels in the center and caster wheels symmetrically placed in the center of the front

[2] As long as Moore's Law [13] holds, the density of semiconductor components quadruples every three years, with attendant speed improvements and power reductions.

and back. This configuration allows the surrogate to turn in place, something that is natural for people to do. In contrast other wheel configurations, such as those standard in an automobile, require either a large turning radius or the execution of multipoint turns. These types of turns require conscious thought and are not consistent with our goal of providing immersion. The wheelchair motors are connected to the drive wheels by clutches, so that surrogates can be manually pushed around during service or deployment. Linear actuators are used to control the tilt and pan of the surrogate's head.

The surrogate suspension system is designed to be able to traverse terrain compliant with the American with Disabilities Act (ADA) [4]. This act dramatically simplifies and constrains the environment that the surrogate must operate within. Because of this, the surrogate does not have to be able to climb stairs or to have full mobility of its manipulators. Some examples of helpful constraints from the ADA specification include that curbs must have ramps at crosswalks, buttons can only be mounted from 35 to 54 inches off the floor, and hallways must be at least 3 feet wide.

Mezzanine

The mezzanine level of the surrogate contains four ATX standard PC systems. Each system drives the LCD panel on one side of the head. These systems also have multiple Viewcast Osprey-1000 H.261 video capture and compression cards running in CIF resolution (352 by 240) as well as a Viewcast Osprey-100 video capture card for remote backdrops (discussed later). Each of them uses a graphics accelerator based on a NVidia Riva TNT-2. This accelerator is capable of texture mapping multiple video streams over an entire screen with alpha blending at video frame rates, including downloading new textures for each frame. Two of the PC systems also control the manipulators (arms), while the third system acts as a router and connects to a wireless local-area network (WLAN). The fourth PC system drives a service display on the back of the head and also controls the wheelchair electronics and sonar. All PC systems in the surrogate can be accessed by pointing an infrared wireless keyboard at their corresponding display. The windows near the center of each body panel are for infrared keyboard sensors.

Forty Polaroid ultrasonic sensors ring the surrogate. They are contained in square projections on the surrogate skins (as shown in Figure 1). The sensors are used as an ultrasonic sonar ranging system as part of the surrogate's collision avoidance and navigation assistance system.

Two teleoperated robotic arms are mounted directly on top of the front of the mezzanine. By having two arms instead of one the surrogate appears more anthropomorphic. Two arms also provide useful redundancy and enhanced flexibility of motion. The arms are lightweight and pliant, to avoid any possibility of injury to people or objects at the remote location. We have experimented with haptic feedback from various robotic arms using a SensAble Technologies Phantom Desktop. Haptic feedback is especially important for the user to determine whether they have actually touched objects in the remote location, since only monocular video is provided. The arms may be used for simple tasks such as pushing buttons.

Head

The head of the surrogate has three levels: LCD panels, video capture, and audio capture. The cameras and microphones are fixed to the head so that when the user moves their surrogate head their visual and auditory views move as well. This produces the same effect for both the user and the remote people they are visiting as if the user was physically present and moving their head.

The head of the surrogate is built with four LCD panels at right angles to each other. Panels on the front and both sides of the head display live video of the user's head. The LCD panels have special wide view coatings which allow them to be viewed over a range of almost 180 degrees both horizontally and vertically. This is in contrast to laptop LCD panels, which can only be viewed over small angles without losing contrast or appearing to have inverse video artifacts. The surrogate's head gives the appearance of an orthographic projection of the user's head. In the long run, as flexible display technologies become available, it would be better to use a display with a shape closer to that of the user's head. The surrogate has two small speakers mounted under the LCD panels of the head, and a subwoofer inside its body.

Eight compact CCD board cameras are mounted directly on top of the displays. To reduce gaze errors, we would like the cameras to be behind the display of the user's eyes on the LCD panel, but this is impractical since the LCD panels are lit. The next best place to mount the cameras is directly above the LCD panels, since this introduces the least error between the surrogate and what would be seen if the user is actually present at the remote location. Five of the board cameras are used to acquire video to send back to the user. Four of these cameras are mounted portrait style, covering a 95 vertical by 72 horizontal degree field of view. The middle (foveal) camera is mounted in landscape orientation, covering a 26 vertical and 35 horizontal degree field of view. (Foveal video will be described in more detail later.) The optical center of the four peripheral cameras is pointed down 20 degrees from horizontal. In comparison, the center of human peripheral vision is pointed down about 15 degrees. We have found a slightly larger angle useful, since there are few hazards overhead in a typical office environment. Furthermore, bright unshielded ceiling lights can negatively affect the exposure settings of the cameras if the lights are too close to the center of their field. In the human eye, the fovea is above the center of the total field of view. Similarly, the foveal camera is centered horizontally, but positioned vertically at 2/3 of the vertical range covered by the peripheral cameras.

This proves useful during slide presentations, since projection screens are typically centered above the horizontal plane of the viewers for maximum visibility. The three remaining cameras acquire remote backdrops (described later). All eight cameras use automatic white balance.

Above the camera level the head contains four short shotgun microphones at right angles to each other. The shotgun microphones are directional, and capture sounds coming from the left, front, right, and back of the surrogate. These are processed by our audio circuitry, sampled, compressed, and transmitted to the user's location. There they are decompressed and used to drive speakers positioned around the user. The top level of the head also contains a two-pronged antenna for connection to the megabit WLAN.

User Station

The user station (shown in Figure 2) was our first system interface at the user site. Its advantages are that it is compact and low cost, but it is not very immersive. The user station consists of an HDTV monitor encased in a light-tight enclosure. The HDTV monitor runs at a resolution of 1920 by 1080, and is controlled by a graphics accelerator based on a NVidia Riva TNT-2 in a dual Pentium II PC. In front of the monitor is a half-silvered mirror at a 45 degree angle. This reflects an image of the user off a second full mirror on to a video camera mounted above the monitor. This arrangement allows the user to make eye contact with people that they see on the monitor. This arrangement is similar to that of Buxton and Moran's Reciprocal Video Tunnel[2].

Two other cameras on the sides of the monitor acquire the side views of the user's face. These views are not radially perpendicular to the view seen by the camera above the monitor, although they are displayed at right angles on the head of the surrogate. This arrangement is needed so that the side cameras do not look into each other, but rather capture a view of the user's face surrounded by chromakey background. The user station does not preserve gaze on the profile views.

Diffused light domes are used to light the user's face. Without the diffused lights, the user's face appears dark and menacing. However, the bulbs in the light diffusers are only 25 Watts, so that reflections off the half-silvered mirror do not overwhelm light coming from the HDTV monitor.

Care has been taken to manage the color of acquired and displayed images. The cameras acquiring the user's face have manual color temperature controls set to the color temperature of the diffused lights. With automatic white balance, the chroma-key blue screen around the user can cause the user's face to appear bright red or green. The displays are set to the same color temperature of the diffused lights, but their color saturation is increased to compensate for reduced contrast of the display/mirror combination[5].

On the desk in front of the display is an infrared wireless keyboard, joystick for audio control, Immersion Corporation force-feedback joystick, SensAble Technologies Phantom Desktop, and a LCD panel. The LCD panel displays the status of the system and surrogate.

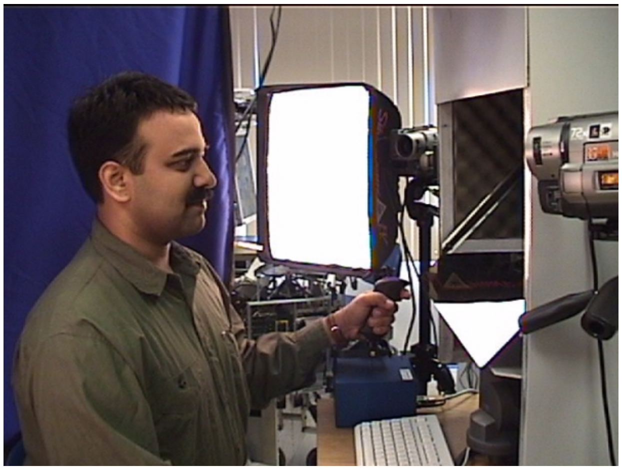

Figure 2: A user at the user station.

Audio Telepresence

In the audio telepresence component[9], we attempt to recreate as accurately as possible both the remote sound field for the user, as well as the user's contributions to the remote sound field as if they were actually present. We transmit four channels of near CD-quality audio from the remote location to the user. Each audio channel is currently sampled at 23KHz[3] with 16-bit resolution. A single channel of 23KHz 16-bit audio is transmitted from the user's lapel microphone to the surrogate. The high dynamic range of the system allows users to whisper back and forth with remote participants. This is a key enabler for private conversations in public spaces, overcoming a limitation in existing technologies[10].

We ring the user with four Bose Acoustimass speakers and a subwoofer for accurately recreating the remote sound field. The user can control the relative gain of the four speakers by steering with a joystick. This is useful for amplifying the voice of a single person speaking in a noisy room and enhancing the “cocktail party” effect available with multichannel audio. Remotely steerable audio provides more capabilities than what the user would have if they were physically present - it is a case of “Beyond Being There[6].”

We have developed our own software to acquire, compress, decompress, and output the audio data. This software also reduces the gain of the microphones when the speakers are active in order to prevent echoes in the audio system. True echo cancellation would be difficult given the changing acoustics experienced as the system moves about, combined with variable network delays through the internet.

[3] This sample rate can match that of CDs when faster WLANs are available.

Haptic Collision Avoidance and Navigation Assistance

A research-quality force-feedback joystick is used for driving the surrogate. The collision avoidance and navigation assistance system outputs forces on the joystick based on input from the ultrasonic sensors. If the speed requested by the user pushing the joystick will cause the surrogate to impact an obstacle within 4 seconds, a “wall” sensation is presented haptically to the user in the direction of the obstacle and the surrogate slows down so that any impact will be 4 seconds away. If the user continues pushing on the joystick, as the surrogate gets closer to the obstacle it will continue to slow down so that the time to impact remains 4 seconds. If this continues further, the surrogate will eventually stop just in front of the obstacle when the time to impact in 4 seconds results in a speed below the minimum speed of the surrogate.

If obstacles are present to the sides of the surrogate, haptic walls will be generated for sideways motions of the joystick. Users can use haptic walls as aids in navigating through tight corridors or around obstacles.

Head Tracking and Remote Backdrops

A chromakey blue curtain surrounds the user at the user station. We have developed head-tracking techniques based on chromakeying. In order to present the user's face life-sized at the remote location, it needs to fill the LCD display. But users naturally shift around in their chairs, so to prevent the user's head from falling off the surrogate's displays we capture a wider field of view at the user station and then translate and zoom with texture mapping to fill the surrogate's display with the user's head. Given the size of the head's LCD panels, this presents the user's head at roughly life size. The use of the blue screen allows us identify the bounding box of the user's head at video speeds with low overhead on a PC.

After we compute the bounding box of the user's head, we compute an exponentially-weighted moving average of the bounding box with a factor of 0.9X when running at ten frames per second. This has several benefits. First, if there is a group of rogue pixels in one frame that causes the bounding box to momentarily expand, by exponentially weighting it with previous samples the disruption to the display is minimized. Second, sometimes users gesture with their heads - for example they may nod their head in assent or shake their head no. By exponentially averaging the bounding box, gestural movements that take on the order of a second or less can be transmitted through the head tracking system without being attenuated by head tracking that is “too fast”.

Instead of leaving the blue-screen behind the user on the displays of the surrogate, we capture a wide field view out the opposite side of the surrogate's head. We first texture map this onto the whole screen, then alpha blend the user's face over it with the alpha being set to a function of the blueness of the pixel. This generates a final image of the user's head in front of scenery from the remote location. The remote backdrop adds to the level of immersion experienced by people at the remote location.

We use a wide-angle lenses for the acquisition of remote backdrop images, even though the angle subtended by the LCD display in the field of view of a person at the remote location is actually much smaller (unless the person has their face a few inches from the display.) To cover such a small angle one would normally use a short telephoto lens. However, because the viewers of the surrogate at the remote location will be scattered over many different angles, and the remote backdrop camera is mounted above the LCD panel, the parallax errors will usually be quite large. This can generate confusion for the people at the remote location as they try to figure out what small patch of the background is being displayed on the screen. In our experience, a wide-angle lens works much better. This covers much more background area than would be present if there was no LCD screen, however the background is easily identifyable even if the view of the background is largely obscured by the user's head. Note that for some backgrounds (e.g., flat painted walls), the field of view of the remote backdrop camera doesn't matter.

Hand gestures are an important part of human communication. When the user gestures with their hands, the head tracking software automatically expands the bounding box to cover their hands as well, transmitting the hand gestures to the remote location (see Figure 3). This may squeeze the aspect ratio of the display of the user's head, but it is preferable to losing the gesture.

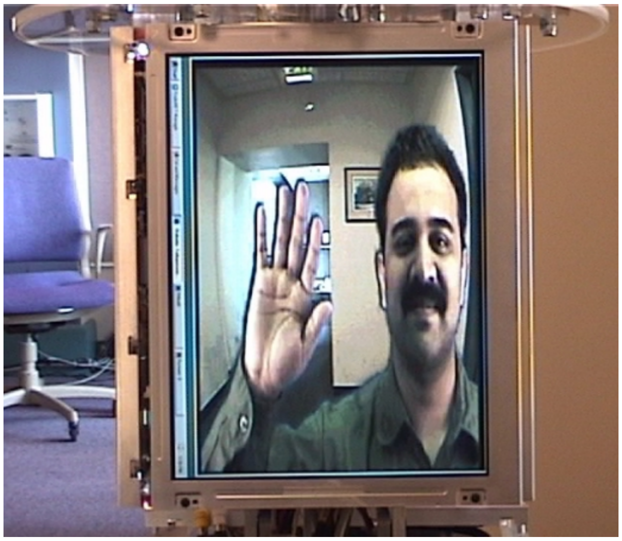

Figure 3: Head tracking automatically backs off to include gestures.

Foveal Video

A screen of roughly 17,000 by 4,800 pixels would be required to create a screen with a 140 degree field of view in a 32:9 aspect ratio with the angular resolution discernible by a young person [12]. Given the current pace of improvements in display technology, it will be some time until this is available. Even worse, the bandwidth required to transmit all of these pixels at video frame rates would be enormous. However, the human eye is capable of this resolution only in its fovea, and not over the whole field of view[11]. Away from the fovea, the resolution of the retina falls off in a Gaussian fashion. We take advantage of this characteristic to reduce both the bandwidth requirements as well as the display requirements in our system.

In our system we acquire a set of video images of varying angular resolution. The centermost video has the highest resolution and the video streams further out from the center are of progressively lower resolution. We composite the separate streams into a single presentation for the user. We call this *foveal video*. Foveal video is attractive only for images coving a very wide field of view. Although the human fovea only covers a small angle of view at any instant, it can rapidly move around. Thus, the center high resolution video must cover an angle of view sufficient to contain the majority of the foveal movement for a significant amount of time. One example of this is the presentation screen during a business meeting. For much of the meeting the viewer's attention will probably be focused on the screen. It is still useful to relay images from the rest of the room, but they do not usually have to be of sufficient resolution to allow reading of remote text. Another example is from moving through a remote location. Views to the sides are used mostly for context and to ensure proper navigation; central views benefit from higher resolution to identify offices and people at a distance.

We acquire images at different angular resolutions by using lenses with different focal lengths. While in theory it would be possible to build a camera sensor with varying resolution over its field, the resolution of current cameras is so low that we need to use multiple cameras anyway. Furthermore in the future, if cameras do achieve resolutions of 17,000 by 4,800 pixels, a camera that had a multiresolution sensor would need to use a lens with a huge field of view approaching 180 degrees. It is difficult to get high-quality lenses that cover such a wide field of view and yet are inexpensive. Thus the use of multiple cameras and video composition seems like it will be necessary for the foreseeable future.

Unfortunately using multiple cameras does create a number of challenges. First, the exposure can vary widely from one camera to another, depending on whether or not there is a bright light source in its field of view. This can cause the abutting video areas to differ in exposure, creating noticeable discontinuities. Second, the cameras can be misaligned, making it difficult to join the videos in a seamless manner. We address this problem by mounting the cameras on precise miniature rotational stages. Third, parallax errors between the cameras mandate that either distant objects on the border between different views will appear twice in the composited image, or there will be a blind spot for nearby objects between the different views, or both. We have chosen to bias the views in favor of duplication over blind spots for objects further than 3 feet from the surrogate.

Another problem with using multiple cameras and compression cards is a lack of synchronization. A lack of synchronization will cause the views to update at different times. If the frame rate is low or the delay between views is large, temporal tears in the image will be objectionable. If the surrogate is stopped, moving objects may momentarily disappear in the cracks between video streams. Even worse, if the surrogate is moving the image will breakup and become confusing to the user. We did not have the capability to synchronize our video compression hardware, so we tried to ameliorate this problem by using as high a frame rate as we could. This reduces the temporal length of the tears. We also use the same video card for all video streams and we set identical compression parameters for each card. This equalizes the compression and decompression delay of the cards, further reducing the temporal length of the tears.

Anamorphic Video

People are used to seeing many objects in perspective. This shortens one dimension more than the other. People can usually read text if it is foreshortened by a factor of 2.5 or less. In order to provide the user with a sense of place, we would like to display video from a wide field of view (e.g., 140 degrees). However the user station only fills about 60 degrees of the user's field of view. In order to display a wide field of view in a small angular space we display the video anamorphically, that is with different scales in the vertical and horizontal dimension. In the user station we compress the foveal view and the centermost peripheral views by 33 percent in the horizontal dimension relative to the vertical dimension. We compress the outermost peripheral views by another factor of two. Thus the central portions are displayed at a scale of 0.75:1 horizontal to vertical, and the outermost peripheral regions are displayed at a scale of 0.375:1.

We abruptly change the scale from the mildly anamorphic central regions to the more steeply anamorphic outer areas. Other options include making the transition gradual, either with a linear ramp in scale or a geometric ramp in scale. Ramping more smoothly creates less of an artifact for objects that straddle the boundary. Figure 4 shows the display of the user station with a foveal region in the center and abruptly changing anamorphic scaling.

Immersion Room

The Immersion Room (see Figure 5) was our second method for controlling the surrogate. Instead of using a single monitor, we use two BarcoGraphics 6300 projectors with the widest possible lenses in a rear projection configuration. The projectors are housed in custom-designed “hush boxes” to reduce their already low noise levels. The screen is bent at an angle of 127 degrees at the center where images from the two projectors meet. Unlike the 16:9 aspect ratio of the user station, the immersion room screen has an aspect ratio of

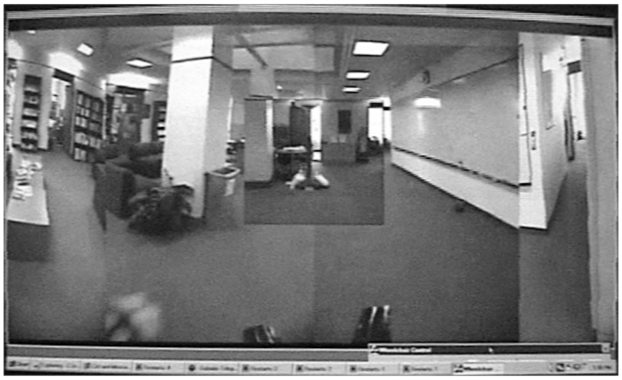

Figure 4: The view seen on the user station, showing a combination of foveal and anamorphic video.

16:6. Moreover, it fills close to 120 degrees of the user's field of view. The same image as presented on the user station is presented on the immersion room screen, but without the anamorphism. In order to capture a direct view of the user, we make a small slit in the projection screen and mount a camera with a pinhole lens on an extension tube behind the screen. Because the projectors have wide angle lenses and the screen is bent, there is plenty of room behind the bend in the screen to mount a camera. Since the pinhole lens is mounted on a narrow tapered tube, only a small dark spot is present on the screen. This camera allows people to make eye contact with the people they are visiting remotely, as with the user station.

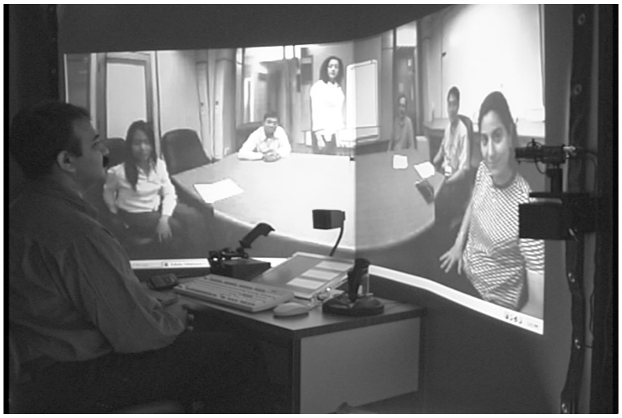

Figure 5: The immersion room.

Except for the projectors and cameras, the rest of the equipment at the immersion room remains the same as the user station. However, instead of using a chromakey blue curtain behind the user, we painted the walls of the immersion room chromakey blue.

Network Requirements

Network bandwidth and latency are both important issues for mutually-immersive mobile telepresence.

Network Bandwidth

The bandwidth of all eight video streams (five going from the surrogate to the user and three from the user to the surrogate) at the NTSC output of their cameras is about 1 gigabit per second. The actual bandwidths in the system are a function of the image quality desired and the network bandwidths available. We are currently compressing all video streams down to H.261 CIF resolution (352 by 240). The views going from the surrogate to the user are compressed at 15 frames per second and at a rate of 320 Kbits/sec. The video streams from the user to the surrogate are compressed at 10 frames per second and a rate of 200 Kbits/sec. We find less bandwidth is required for the user's face than for the view presented to the user. The bandwidths total 1.6 Mbits/sec from the surrogate to the user and 0.6 Mbits/sec from the user to the surrogate.

The audio bandwidths total 0.4Mbit/sec from the surrogate to the user and 0.1Mbit/sec from the user to the surrogate. The bandwidth for controlling the motion of the surrogate is negligible.

The surrogate is connected to the internet through a wireless LAN (WLAN). We are currently using a radio network from RadioLAN in the 5.8GHz ISM band. It was the first WLAN that supported data rates of up to 10 Mbits/sec, although it is now becoming obsolete. Data rates of up to 54 Mbits/sec are now available using IEEE 802.11a.

Network Latency

Latency is a key issue in telepresence. For some applications, such as remotely viewing presentations, large latencies are tolerable. However for interactive conversations or for remote manipulation latency is a bigger problem. Currently in our system, the minimum one-way latency in the video channel is about 400 milliseconds. Almost all of this latency is due to delays in video compression and decompression. Our audio channels are currently unsynchronized with the video, and are transmitted as quickly as possible. We do this since previous research has shown that it is better for audio to be faster than the video in interactive conversations rather to delay the audio for synchronization with slower video [7]. Depending on the distance between the surrogate and the user, long-distance network delays may add an additional 200 milliseconds one-way (if the surrogate and the user are on opposite sides of the Earth and satellites are not used).

For best immersion, the total round trip latency should be a couple of hundred milliseconds or less. Lower latency compression and display technology will help, but speed-of-light communication costs form a lower bound on latency. Although longer latencies may occasionally be irritating, we believe tolerating several hundred millisecond latencies to go around the Earth for a meeting using mutually-immersive mobile telepresence is still much preferable to tolerating latencies of several days when actually physically traveling around the Earth using commercial subsonic aviation.

Economics

The electronic components in the surrogate currently cost about $20K. The most expensive of these are the wide-view LCDs, which cost about $2K each. The cost of the mechanical and packaging components are about $10K. These costs can come down if the surrogate is produced is larger quantities. Since most of the cost of the surrogate is its electronics, its cost should scale down over time due to Moore's Law.

USER EXPERIENCES

Due to the limited space available, we discuss the results of our experiences with the system in general terms only. Over a dozen people have used the system, and more than a hundred have seen the system in operation. The following observations are a synopsis of their comments.

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

Users unanimously prefer the immersion room over the user station interface for several reasons. First, there is more of a feeling of openness in the immersion room while the user station is more cramped. The large projection wall of the immersion room allows video images to be presented at close to life size and fill a large portion of the user's field of view, whereas the user station images are smaller than life size and only fill a small portion of the user's field of view. Second, there is better visual contrast in the immersion room, while the user station suffers from poor contrast due to the half-silvered mirror. Third, in the immersion room the audio directionality matches the video presentation, while sounds coming from the left at the user station correspond to objects that are presented on the left side of the monitor (which is in front of the user).

Gaze and Life Size Issues

The remote physical presense provided by the surrogate is a significant improvement over conventional videoconferencing systems. It allows surrogate users to sit around a meeting table just as if they were physically present. The surrogate maintains a 1-to-1 correspondence between people and people-sized spaces, helping remote users to maintain a unique individual identity in the minds of the other participants. In contrast, users of traditional group video conferencing systems often end up pasted on a monitor shared between many remote users. Being large, unwieldy, and encumbered with wiring connections to the wall, such systems usually end up stuck off to the side of a meeting room, isolating users of the systems. Furthermore, when many users share a single monitor, they tend to be thought of with a shared group identity specific to their remote location.

Audio Experiences

The multichannel high-dynamic range sound system has been 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 pick out one speaking person in a room full of speaking people, enabling selective attending to and participation in parallel conversations at the remote location.

To give an example of the dynamic range achieved, after everyone left a meeting room but the surrogate user, it was possible to auralize the position of a clock ticking on a wall of the conference room without using any visual stimuli. Some users have described the audio realism achieved as “spooky,” since it sounds like remote participants are really present but they aren't.

Video Experiences

In general foveal video has been favorably received. However, the resolution of the current video compression cards is not great enough to read slide presentations remotely unless they consistently use large text, even using the enhanced resolution of the fovea. Future, higher resolution video codecs should improve this significantly.

Anamorphic video (only on the user station) is useful during meetings as it allows more of the meeting participants to be seen by the user. However, users find it difficult to judge distances to objects when driving the surrogate with anamorphic video. The effect is a bit like that of many automobile passenger-side rear view mirrors that say “objects are closer than they appear,” except that in the case of anamorphic video, the apparent distance to an object varies depending on where it is on the display.

Color Management

Diligent color management is key for enabling immersion in both directions. Presenting the remote location in vivid colors that appear lifelike significantly aids the immersion of the user in the remote environment, especially if they have ever been in the remote location in person. Similarly, flesh tones can be difficult to recreate, but are important for presenting the user as a equal participant at the remote location. Although inattention to color issues can result in complaints, we have not yet performed formal studies of color accuracy versus perceived immersion.

Experience with Remote Backdrops

The remote backdrops received almost universally favorable responses. In some cases, since the user's head fills most of the screen, people at the remote location were not be consciously aware of its operation when the surrogate was stationary. However when the surrogate is in motion, remote backdrops give a strong impression of motion on the part of the user, since the backdrop is moving with the motion of the surrogate while the user's head is fixed on the screen.

Mobility Experiences

Initial user reaction to the system mobility has been positive. The system is clearly a large improvement for the user over traditional video conferencing. Users especially liked cruising down the halls with the surrogate. Most people enjoy operating remote-controlled toys, and controlling the surrogate shares some of the same qualities.

We have found that the automatic collision avoidance and navigation assistance system is very useful, particularly given the delay of the communication channel and as an aid when learning how to drive remotely. When we first tried driving the surrogate remotely without it we ran into a number of objects. Since the ultrasonic sensors make an clicking sound that is not loud but is audible, it is important to turn them off when the surrogate is parked. Otherwise they can distract the remote participants in a meeting.

We have found the ability to pitch the head up and down is important when learning how to drive the surrogate. Pointing the head down while driving is similar to looking at your feet when walking. However, we have found that the current head actuators are too noisy for use in meetings without distracting other participants.

Form Factor Issues

Although people in the United States with disabilities have made great strides in recent years in terms of equal opportunities, access, and image, some people still have negative stereotypes of people in wheelchairs. Because the hardware dictated the same form factor of a person in a wheelchair for the model 1 surrogate, some people have reported negative stereotypes of people using the surrogate for similar reasons.

Although the model 1 surrogate is narrow enough to fit through standard office doorways, it is too long to turn in place in the middle of a doorway. Also, because there is no remote video coverage out the rear of the surrogate, it is hard to back up out of a crowded office.

Based on these results, we are working on a new version of the surrogate that has a footprint that would allow it to turn in place in a doorway. This surrogate would also have a form factor that would be closer to that of a walking person, and transmit video from all directions back to the user.

Head Height Issues

Although use of a single surrogate head height reduces mechanical complexity, we found compromising on a single head height resulted in negative user experiences. For example, if after a meeting everyone is standing around talking with each other, the low head height of the model 1 surrogate requires remote participants to “look down” on the user of the surrogate. Unfortunately height can be an indicator of more than one type of stature in human interactions, so some users of the surrogate were reluctant to put themselves at a relative height disadvantage. Our next surrogate will have an adjustable head height, covering a range of heights appropriate for sitting or standing users.

Experience with Manipulators

Initially users (all without prior teleoperated manipulation experience) thought using the arms would be easy. However after many months of work we found that the surrogate's arms were still very difficult to use. Another observation was that a few remote participants were ill-at-ease when sitting by the surrogate's arms (perhaps they had seen too many science-fiction horror films with marauding robots).

Based on the training required, the limited capability they afforded, and the contrast between reality and user expectations, we decided to omit manipulators from surrogates for the near future.

Haptic Experience

A final observation is that there is still room for lots of research in haptics. The force-feedback joystick we use, although worlds better than consumer force-feedback joysticks, still has a very limited dynamic range and a limited maximum force. Yet it is one of the more expensive parts of the system. We would like to output forces in many different gradations, but have been limited to providing a spring force to center the joystick combined with a fixed force wall.

Networking Issues

Our current WLAN network does not provide any quality-of-service (QOS) features. This causes problems for transmission of real time data. For the video, this can result in the video stream getting stuck occasionally for up to a second. Unless one is driving this is not too serious of a problem. However, for the audio intermittent delays can cause terrible disruption to the remote sound field - many small gaps in speech can cause it to be unintelligible and are highly irritating. To avoid small gaps in the reproduction of the remote sound field, our audio telepresence system uses adaptive buffering. This trades off increased latency for improved continuity.

WLAN networks with QOS guarantees are still a topic of research, although some products with some capabilities are becoming available. We are looking forward to being able to use these capabilities to reduce our audio latency.

CONCLUSIONS

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

As part of the project, we are investigating a number of different technologies. Our work in head tracking and remote backdrops has application to more traditional forms of video conferencing as well. Foveal video is an important step towards providing very wide fields of view and enough detail to be useful, without waiting years for technological progress to achieve similar resolution over the entire field of view. The audio telepresence component enables users to whisper back and forth with individual remote participants, auralize the location of remote sound sources, and utilize the cocktail party effect to improve the intelligibility of remote speakers. The 1-to-1 correspondence between participants and "people-sized spaces" provided by the surrogate helps make remote meeting attendees full-fledged participants. The mobility of the surrogate is an important enabler for casual meetings. Finally in the future, if deployed in volume and after continued scaling of the electronic components, the system should cost less to rent than an automobile.

In summary, we believe future mutually-immersive mobile telepresence systems have the potential to be an economically compelling substitute for many types of business travel.

Acknowledgements

The author is grateful for those who helped with various aspects of the project. Wayne Mack built the Model 1 surrogate. Dan Scales provided the interface to the wheelchair electronics. Subu Iyer helped with the acquisition and display of video. Summer interns Mark Newman, Keith Emnett, and Mike Pan were a great help in the initial system implementation, control of the manipulators, and the audio telepresence system, respectively.

REFERENCES

1. Bureau of Transportation Statistics. Transportation Statistics Annual Report. Technical report, U.S. Department of Transportation, 1999. Available at www.bts.gov.

2. 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.

3. W. A. S. Buxton. Telepresence: Integrating Shared Task and Person Spaces. In *Graphic Interface '92*, pages 123–129, 1992.

4. Evan Terry Associates. *Pocket Guide to the ADA: Americans with Disabilities Act Accessibility Guidelines for Buildings and Facilities*. John Wiley & Sons, revised edition, 1997.

5. E. J. Giorgianni and T. E. Madden. *Digital Color Management*. Addison-Wesley, 1998.

6. J. Hollan and S. Stornetta. Beyond Being There. In *ACM CHI '92*, pages 119–125, 1992.

7. E. A. Issacs and J. C. Tang. Studying Video-Based Collaboration in Context: From Small Workgroups to Large Organizations. In K. E. Finn, A. J. Sellen, and S. B. Wilbur, editors, *Video-Mediated Communication*, pages 173–197. Lawrence Earlbaum Associates, 1997.

8. E. A. Issacs, J. C. Tang, and T. Morris. Piazza: A Desktop Environment Supporting Impromptu and Planned Interactions. In *Proceedings of CSCW '96*, pages 315–324, 1996.

9. N. P. Jouppi and M. J. Pan. Mutually-Immersive Audio Telepresence. In *Proceedings of the 113th Audio Engineering Society Convention*, October 2002.

10. R. Kraut, R. Fish, R. Root, and B. Chalfonte. Informal Communication in Organizations: Form, Function, and Technology. In R. Baecker, editor, *Groupware and Computer-Supported Cooperative Work*, pages 287–314, 1993.

11. M. W. Matlin and H. J. Foley. *Sensation and Perception*. Allyn and Bacon, 4th edition, 1997.

12. M. Milne. Armchair Resolution. *ACM SIGGRAPH Computer Graphics*, 33(3):6–8, August 1999.

13. G. E. Moore. Cramming More Components onto Integrated Circuits. *Electronics*, pages 114–117, April 1965.

14. H. Nakanishi, C. Yoshida, T. Nishimura, and T. u Ishida. FreeWalk: A 3D Virtual Space for Casual Meetings. *IEEE Multimedia*, 6(2):20–28, April 1999.

15. E. Paulos and J. Canny. PRoP: Personal Roving Presence. In *ACM CHI '98*, pages 296–303, 1998.

16. A. Prussog, L. Mühlbach, and M. Böcker. Telepresence in Videocommunications. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting*, pages 180–184, 1994.

17. R. W. Root. Design of a Multi-Media Vehicle for Social Browsing. In *Proceedings of CSCW '88*, pages 25–38, 1998.

18. A. Sellen and B. Buxton. Using Spatial Cues to Improve Videoconferencing. In *Proceedings of CHI '92*, pages 651–652, 1992.

19. S. Whittaker, D. Frohlich, and O. Daly-Jones. Informal Workplace Communication: What Is It Like and How Might We Support It? In *ACM CHI '94*, pages 131–137, 1994.

20. K. Yamaashi, J. R. Cooperstock, T. Narine, and W. Buxton. Beating the Limitations of Camera-Monitor Mediated Telepresence with Extra Eyes. In *ACM CHI '96*, pages 50–57, 1996.