Telepresence with no Strings Attached: An Architecture for a Shared Reality Environment

Christian Côté, Shawn Arseneau, and Jeremy R. Cooperstock*
Centre for Intelligent Machines, McGill University
3480 University Street
Montreal QC H3A 2A7
Canada
Tel: +1-514-398-5992

Abstract

The Shared Reality Environment is an ongoing project that explores the use of virtual reality technologies to achieve realistic computer-mediated human-human interaction. The project integrates immersive displays, spatialized audio, haptics, and gesture recognition, through a minimal latency network architecture. As our primary goal is to provide distributed participants with a convincing sense of co-presence without inhibiting natural, spontaneous interaction, the environment must employ unobtrusive technology wherever possible.

We present an architecture that uses image processing to track the user in an immersive enclosure, monitor the user's gestures and incorporate other (remote) participants into the shared space. Such a video-based approach permits untethered interaction without the constraints of a complex user interface. Moreover, because the various image processing algorithms involved are all based on an effective background removal step, we are able to leverage our use of both hardware and software.

Keywords: augmented reality, telepresence, person tracking, gesture recognition, image processing

1 Introduction

Long-distance communication has improved steadily since the advent of telephony 125 years ago, with important advances in signal quality and latency reduction. However, despite tremendous hype and early enthusiasm[1] for videoconference systems, users often find the quality of interaction impeded, rather than improved by, the addition of video. The benefit of seeing the person at the other end of the line is generally insufficient to compensate for increased latency and reduced audio quality associated with the technology. Videoconferencing systems are still far from a satisfactory substitute for physical presence.

One of the most significant limitations with videoconferencing is that the displays tend to be quite small, for example, a small television monitor or a window on a computer desktop. Such displays do not allow for peripheral vision and furthermore, fail to convey important social cues such as gaze awareness between conversants. Worse, the additional latency imposed by the audiovisual encoding and decoding process results in an unnatural turn-taking conversation, reminiscent of long-distance telephone calls in the days before fiber optics.

A second important factor is audio. The limited fidelity of a monaural sound system may be adequate for a two-party, handheld telephone conversation, in which the separation of microphone and speaker permits simultaneous speech from both sides. However, conversations become stifled when half-duplex communication is imposed (e.g. by a speakerphone) or when multiple participants join the discussion. In the latter case, the lack of spatial separation between the audio signals prevents our brains from attending to one speaker at a time (the "Cocktail Party Effect" [1]).

The Shared Reality Environment (SRE) project is an attempt to bridge the gap between current videoconferencing technologies and physical presence. Employing immersive, rear-projection displays, spatialized audio systems, haptic feedback devices and gesture recognition tools, we hope to overcome the limitations enumerated above. Although the SRE development is in its early stages and remains very much a work in progress, we envision that it will lead to a greater sense of co-presence and enable unimpeded interaction between distributed participants.

To achieve such a level of interaction, we must successfully accomplish a number of steps. First, for each media source, meaningful information must be segmented from the background. This not only means filtering out noise, but also eliminating information that

[1] By way of example, AT&T invested over $500 million in the development of the Picturephone, the first videoconferencing system commercially available, and predicted sales of 1 million units by 1980. In reality, only a few hundred units were sold.

is of no interest to the participants. Next positions of objects of interest must be localized, in order to carry out accurate spatial rendering of the sources in the remote sites. At this stage, analysis could also be performed to recognize users' gestures or actions, as required. The data must then be transmitted, with minimal latency, to each remote location, where the audiovisual information is reproduced in an accurate manner. This requires a perspective projection based on each viewer's position and multi-channel audio mixing for effective sound spatialization.

Each of the above steps poses formidable challenges. Moreover, solutions that apply to one sensory modality do not necessarily work for others. As our efforts have so far concentrated on the visual domain, the remainder of this paper reports on our progress in this direction.

2 Background

The Shared Reality Environment is one of many telepresence research projects conducted around the world [8, 10, 12, 13, 14, 16, 17]. Among these, the "Office of the Future" [17] at the University of North Carolina is particularly interesting, as it imaginatively addresses many of the same issues as our own: seamlessness of the technology, realism of the experience, and quality of the interaction. However, the constraints imposed in adapting the technology to a real office environment are somewhat restrictive in terms of the scope of applications we would like to accommodate. Indeed, distributed musical rehearsals or distance education are typically ill-suited to an office environment.

Our objective with the SRE is to overcome the limitations of conventional telepresence tools using novel technologies and practices. For example, high-end virtual reality immersive displays, such as the CAVE [5], are more compelling and visually engaging than desktop computer monitors for human-computer interaction. Similarly, we believe that large screen displays, in which participants are projected at "life-size" allow more effective human-human interaction [10, 17, 18]. Likewise, high-resolution spatialized audio can support such demanding applications as musical rehearsals or performances [8] as well as multiple simultaneous conversations. Haptic feedback can be introduced to help bridge the physical separation of remote individuals. Such feedback could range from reproducing the floor vibrations in response to a user walking around to simulating the tactile response of a surgeon's instrument as it moves through different tissues [6]. Finally, gesture recognition could allow virtual shared objects, such as CAD models, to be manipulated by distributed design teams, enabling a new range of computer-supported collaborative tasks. By incorporating all these technologies into a single system, we believe it is possible to establish a *shared reality* in which distributed users are able to interact freely, unhindered by the constraints of conventional "state-of-the-art" videoconference systems.

The SRE is composed of multiple rooms, each of which contains an enclosure of three screens of rear-projected video, a multi-channel sound system for the generation of spatialized audio, various haptic transducers and gesture recognition mechanisms, all interconnected via a high-speed network. An important element of our research is the concept that interaction with the technology must be transparent to the users. This implies that the computer should recognize what the user is attempting to do, and not the other way around. Furthermore, we avoid any form of body-worn trackers or other special clothing restrictions that one might be tempted to use for gesture recognition or person tracking. The approach is similar in philosophy to that adopted in Simon Penny's "Traces" [11], which uses computer vision to perform wireless full body tracking. This is motivated by the desire to free the user from any technology-imposed constraints that could inhibit spontaneous expression. While satisfying this objective is a daunting task, the environment is designed to be "walk-in and use" with the same ease as picking up a ringing telephone to speak to the person at the other end of the line.

3 Architecture

As a first step toward these goals, we have developed an architecture that simultaneously performs user tracking, gesture recognition and representation of remote participants in the virtual space, using a series of image-processing algorithms. While not necessarily as computationally efficient nor as accurate as techniques employing bodily-worn sensors, this video-based approach helps make the technology transparent to the user, and further leverages the hardware and software already in place for basic videoconferencing. The following sections explore in more depth how the design for each task compares with other alternatives.

In order to prototype this architecture, we are using a two-screen testbed as pictured below in Figure 1. Experimental details employing this testbed are covered in section 4.

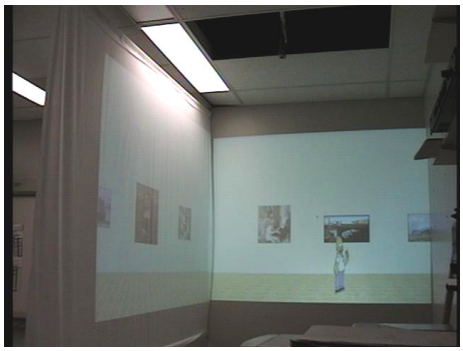


Figure 1 Picture of the SRE prototype

3.1 Participant Representation

One of the first challenges to establishing communication between several remote participants is representing each of them realistically in a single virtual space. These representations, or avatars, may not be perfect, but should be sufficiently convincing to allow seamless interaction.

A naive solution would be to simply have an unmodified video projection, where the camera output directly feeds the video wall display; in other words, conventional videoconferencing with several large screens. While this would certainly be easy to implement and could be accurate provided all participants remained stationary at predetermined positions, it only offers mediocre realism for the range of applications that we wish to support. Furthermore, this model would be impractical for anything more than three participants.

A second alternative is to isolate a rectangular bounding box that encompasses each user and paste these video fragments onto a generic video background. While this solves the mobility problem and allows interaction of more than three participants, it does not offer a very convincing sense of realism. Quite simply, the use of rectangular video fragments implies the inclusion of small areas of background video that do not belong to the users', thus detracting from the quality of display. Worse still, if two participants are displayed next to each other, unnatural occlusion of one by the background portions of video from a second participant's bounding box could occur.

To help avoid these problems, we perform background removal on the camera input to obtain an image of the user isolated from the scene. This image is then inserted into the virtual environment (see Figure 2). This approach has the advantage of offering an acceptable level of realism, especially pertaining to occlusion. Indeed, only those parts of the space that are behind the user are hidden from view, as would be the case in a physical setting. The main shortcoming of our approach is the absence of volume of our avatar. One possible solution would be to use stereo cameras to collect depth information and construct a relief model of the participant [10]. While this method holds promise, experimental results clearly indicate that the technology needs to mature before we can adopt it for practical applications.

Figure 2 Remote participant avatar with background removed

3.2 Latency Considerations

The use of a background removal step to represent the user also helps us attain our low latency objective. Satisfying this goal for the video stream presents a number of challenges. A first approach is to make use of M-JPEG or MPEG encoded video. Unfortunately, high-quality MPEG hardware tends to be expensive and the encoding algorithm introduces a delay of several frames. While the cost and latency imposed by M-JPEG encoding is less significant, compression and decompression time are still in the order of 50ms per frame, on top of the image acquisition and network transport time.

Avoiding compression presents the option of transmitting raw data. For high resolution, 30 fps video, this requires massive amounts of bandwidth. Even on a 100 Mbps Ethernet, transmission of a single frame of 640x480 resolution at 24 bits per pixel takes approximately 100 ms. Borrowing from compression techniques, we note that much of the data in a sequence of video frames is redundant, for example, a static background, which may constitute the majority of each frame. Since our earlier processing step has already removed the background in its entirety, we can obtain a significant decrease in transmission time by sending only the remaining image components, as raw data, provided they are reasonably small.

3.3 User tracking

One of the main strengths of immersive display systems is the fact that the rendering of the scene is accurate from any position the user may occupy in the enclosure. However, to compute the correct perspective,

the software needs to know the exact position of the viewer at all times.

Typical trackers for the CAVE include electromagnetic devices such as Ascension's Flock of Birds™, wearable transmitters, or optical markers [4]. While the performance of such devices is impressive, their reliability comes at the expense of reduced user freedom, requiring either that the user wears one or more sensors or is tethered by a cable. As neither of these limitations are consistent with our philosophy of freeing the user from the machine, we consider, instead, the use of purely optical tracking, allowing for interaction that more closely resembles that available for physical co-presence.

Our approach makes use of the processed video image with the background removed, a step that has already been performed for the purpose of representation in the virtual space. Since the background removal leaves us with only the user in each video frame and we know where each camera is located, it is possible, with the use of three cameras, to track any feature we choose. The coordinates of that feature, for example the head, are then relayed to the immersive display software to obtain the appropriate rendering.

As a first test of this approach, we have tracked the center of mass of the user on the floor plane using a ceiling mounted camera. Once the tracking is activated, one has only to step into the SRE before the rendered display matches the user's viewpoint. While the algorithm is fairly primitive and cannot compete with the accuracy of commercial trackers, the initial performance results are highly encouraging. Should the need for more precise measurement arise, a more sophisticated algorithm, e.g. an elliptical head tracker [3], could replace the current body tracker.

Other wireless optical tracker designs, such as the "Inexpensive tracker for the CAVE" [15] could also be used in our environment, and may offer higher precision, albeit at the loss of reuse of both equipment and software.

3.4 Gesture Recognition

The ability to recognize users' gestures for the manipulation of synthetic objects shared by all participants is one of the more interesting functions of the SRE. This allows a broad range of applications that are inaccessible in current videoconferencing technology, such as collaborative design work using virtual CAD models (see Figure 3). However, integrating such functionality without compromising our design objective of transparency presents a serious challenge.

Figure 3 Simulation of CAD model manipulation in the SRE

Gesture recognition and collaborative manipulation are typically achieved through the use of wearable data gloves or trackers, trackball or joystick-like devices, markers or optical algorithms [4]. Body-worn trackers, such as data gloves [9], offer attractive features including high precision and haptic feedback. However, they constrain users by requiring them to wear specific gear, which may be uncomfortable for extended periods of time. Regardless, the requirement of putting on a special device in order to access certain functionality of the environment risks reducing the spontaneity of expression. We would consider it absurd if, for example, in order to read our email, we first had to put on special "email-reading glasses." Similarly, we consider it unacceptable to demand that participants stop in the middle of a meeting in order to put on data gloves for CAD manipulation. While this is perhaps less of an issue for the use of trackballs and joysticks, such interfaces often require special expertise and training to operate effectively. This is mainly due to the non-isometric mapping between the device and the effect that actions induce on the object being manipulated.

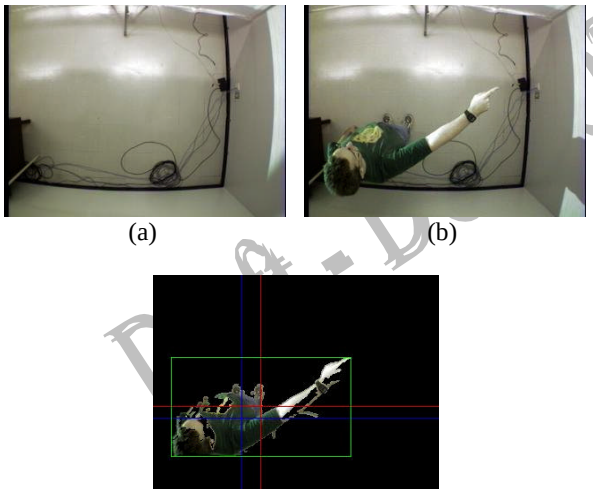
Video-based methods, on the other hand, offer a direct coupling between users' actions and the corresponding effect. With image processing algorithms, we are free to make our input mechanism as powerful as the recognition accuracy of the video processing permits. Furthermore, because physical contact is unnecessary, we free the user from encumbering constraints.

A generic video-based gesture recognition algorithm can be constructed as follows: First the region of interest, i.e. the user, is isolated from the rest of the scene. Next, the cut-out user image is segmented into its constituent components, such as head, arms and legs. Finally, analysis is performed to establish the relative motion of each body part. Since we have already obtained a background-subtracted image for participant representation and tracking, the first phase of gesture recognition is already complete. Various techniques exist for the segmentation and analysis phases, including the relatively low-cost blob approach employed by Penny's "Traces" tracker [11].

4 Image Processing Results

One of the key factors in designing an image processing algorithm for the SRE is its need to run in real-time. This is necessary to maximize the sense of realism, as latencies in excess of a single frame time might be noticeable, detracting from the user's experience.

. Some work has been done with standard background removal [7] where a single image, taken of the environment without the user, is subtracted from the live video to produce an image stream whose frames contain only those pixels that differ from the original background. This approach suffers from several limitations, most importantly, its reliance on constant lighting intensity of the background in order to function correctly. An improvement, known as the background primal sketch technique [19], addresses this particular problem without unreasonable computational overhead by taking the median value of the pixel color over a series of images. The threshold is also calculated on a pixel-by-pixel basis to help distinguish the user in regions that are more susceptible to lighting variations. Sample output of the background removal process employing the primal sketch is illustrated in Figure 4.

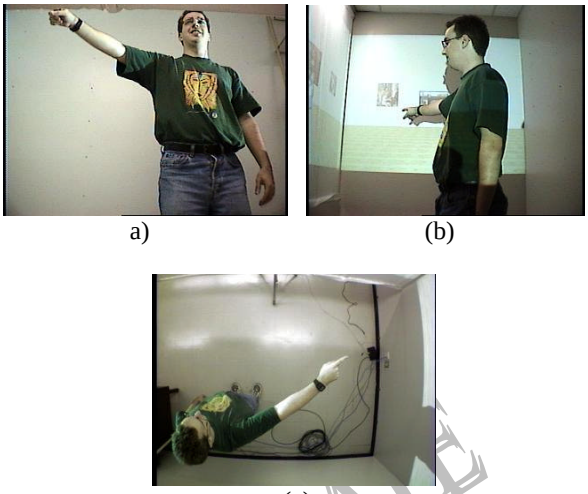


(a) (b)

(c)

Figure 4 (a) original scene, (top view) (b) user in scene (c) difference image using primal sketch technique with left/bottom crosshairs indicating the center of mass and the bounding box denoting the largest 4-connected region in the scene

For simple gesture recognition, the background removal process is performed on the video from three cameras in the SRE, providing front, right side and top views of the user (see Figure 5). The resulting images can then be segmented and analyzed, as described earlier.



a) (b)

(c)

Figure 5 User views from the SRE cameras. (a) front, (b) side and (c) top view

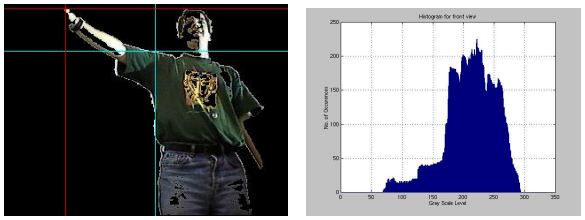
For the *image transmission algorithm*, the difference image from the front camera view is used as a mask on the original scene to create an image where only the user appears on a black background. At present, this is an idealization, as some interpolation between multiple camera views is generally more appropriate, depending on the viewer's angle with respect to each camera.

The next step is to track the user within the SRE's floor plan. This is done by extracting the center of mass of the largest 4-connected region in the difference image from the top camera view, which normally coincides with the user within the scene.

The *gesture recognition algorithm* uses information from all three camera views in order to ascertain the user's pointing direction. The first step is to retrieve the center of mass information used in the *tracking algorithm*. Next, a bounding box is formed around the largest 4-connected region in the scene (see Figure 4). Using a coordinate system with *x,y* denoting the user's position on the floor, and *z* positive toward the ceiling, the *x,y* direction of user pointing is formed by the vector from the center of mass to the center of the bounding box. The front and side view are used to obtain the corresponding *z* value and serve as an additional error correction for the *x,y* coordinate components.

These two views use a slightly different process to retrieve the pointing direction due to the perspective of the user. The first step is to retrieve the center of mass from the respective difference images. The horizontal histogram is then used to determine whether the left or right arm is being extended. Finally, the location of the extremity, normally the user's hand, as well as that of the largest increase in values, typically denoting the shoulder, are located and used in the calculation of the

arm elevation vector (see Figure 6). The combination of *x*, *y*, and *z* parameters fully specify the direction in which the user is pointing and can then be used to aid in object selection or manipulation tasks.



(a) (b)

Figure 6 (a) difference image (front view) (b) resulting horizontal histogram

For the algorithms employed here it is important to note that there remain some serious constraints that are now being addressed. Due to the nature of the pixel-by-pixel calculations, the cameras must be stationary and the background scene must remain static, meaning that the display screens cannot be within the field of view of any cameras. As the latter is an unrealistic constraint, we are presently investigating the use of both infrared illumination and structured light [17] to assist in the removal of dynamic backgrounds.

5 Conclusions and Future Work

We have presented an efficient, inexpensive method to reduce video bandwidth, avoid occlusion, track users and recognize gestures using video cameras and real-time software. We have demonstrated that although individual design components may be inferior to alternative solutions, the combined system offers efficiency that could not be attained by bringing together disparate technologies. Further, we have shown that it is possible to construct a telepresence system in which the technology is transparent. Users need only "walk-in and use" to start interacting with others participants.

While several enhancements of the technologies presented here are currently in progress, our initial results are highly encouraging. For the user tracker, we are planning to use a different algorithm to track the head in addition to the center of mass. This, along with the use of a larger number of cameras, should provide more stable results. For the gesture recognition engine more complex manipulations, such as translation and rotation, will need to be supported. Improved tracking resolution will likely become increasingly important as well.

On a larger scale, several aspects have yet to be addressed. The use of spatialized audio in conjunction with video displays remains largely unexplored, as is the integration of haptic feedback. We also seek to have the computer play a more active role in interpreting the user's actions in order to provide context-sensitive feedback. While these remain open research topics, we are heartened to see that significant progress is being made and believe that high levels of fidelity and realism, as well as unencumbered, natural interaction, will soon be taken for granted in immersive telepresence applications.

6 Acknowledgments

The authors would like to thank Matt Szymanski of VRCO and Tom DeFanti of the University of Chicago for their assistance with the CAVE library, as well as Leslie Sponder and Annie Bolduc for their comments on earlier drafts of the paper. Aoxiang Xu developed the low-latency video transport protocol described in Section 3.2. Support has come from the Natural Sciences and Engineering Research Council of Canada, Fonds pour la Formation de Chercheurs et l'Aide a la Recherche (FCAR), Petro-Canada, Canarie Inc., and the Canadian Foundation for Innovation. This support is gratefully acknowledged.

7 References

1. Arons, B., A Review of the Cocktail Party Effect. Journal of the American Voice I/O Society. 1992.
2. Arseneau, S. and Cooperstock, J. "Real-time Image Segmentation for Action Recognition." IEEE Pacific Rim Conference. Victoria, Canada. 1999.
3. Birchfield, S. "Elliptical Head Tracking Using Intensity Gradients and Color Histograms." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Santa Barbara, California. pp.232-237. 1998.
4. Buxton, B. A directory of sources for input technologies, online resource: http://www.dgp.toronto.edu/people/BillBuxton/InputSources.html
5. Cruz-Neira, C., Sandin, D. J. and DeFanti, T. A. "Surround-Screen Projection-Based Virtual reality: The Design and Implementation of the CAVE" Computer Graphics, SIGGRAPH Annual Conference Proceedings, pp.135-142, 1993
6. Howe, R., Peine, W., Kantarinis, D. and Son, J. "Remote Palpation Technology." IEEE Engineering in Medicine and Biology Magazine. Volume 14, Issue 3. pp. 318-323. May-June 1995.
7. Kahn, R. and Swain, M. "Understanding people pointing: the Perseus system." Proceedings International Symposium on Computer Vision IEEE Comput. Soc. Press. pp.569-74. Los Alamitos, CA, 1995.
8. Konstantas, D., Orlarey, Y., Carbonel, O. and Gibbs, S. "The Distributed Musical Rehearsal Environment." IEEE Multimedia, Volume 6, Issue 3. pp. 54-66. July-Sept. 1999.

9. Medl, A., Marsic, I., Andre, M., Kulikowski, C. and Flanagan, J. "Multimodal User Interface for Mission Planning." AAAI Symposium on Intelligent Environments, pp.102-109. Stanford, CA, 1998.

10. Ogi, T., Yamada, T., Tamagawa, K. and Hirose, M. "Video Avatar Communication in Networked Virtual Environment." The 10th Annual Internet Society Conference. Yokohama, Japan. 2000.

11. Penny, S., Smith, J. and Bernhardt, A. "Traces: Wireless Full Body Tracking in the CAVE." ICAT Virtual Reality Conference. Japan. 1999.

12. Reynolds S, Cammaert G. Building and using telepresence classrooms. Fid Review, vol.1, no.2-3, 1999, pp.93-6. Publisher: FID, Netherlands

13. Roussos M, Johnson A, Moher T, Leigh J, Vasilakis C, Barnes C. Learning and building together in an immersive virtual world. Presence-Teleoperators & Virtual Environments, vol.8, no.3, June 1999, pp.247-63. Publisher: MIT Press, USA

14. Salem B. Implementing a multimodal user interface for telepresence systems. SPIE-Int. Soc. Opt. Eng. Proceedings of Spie - the International Society for Optical Engineering, vol.3840, 1999, pp.46-53. USA

15. Sharlin, E.; Figueroa, P.; Green, M.; Watson, B. A wireless, inexpensive optical tracker for the CAVE, Virtual Reality, 2000. Proceedings. IEEE , 2000 , pp 271-278.

16. Stone, R.J., Virtual reality and telepresence: an initiative within an initiative, Advanced Robotic Initiatives in the UK, IEE Colloquium on , 1991 , pp. 7/1 -7/3

17. Welch, G., Fuchs, H., Raskar, R., Towles, H. and Brown, M. "Projected Imagery in your 'Office of the Future'." IEEE Computer Graphics and Applications. Volume 20, Issue 4. pp.62-67. July-Aug. 2000.

18. Yamaashi, K., Cooperstock, J.R., Narine, T. and Buxton, W. Beating the Limitations of Camera-Monitor Mediated Telepresence with Extra Eyes. Proc. of CHI'96, Conference on Human Factors in Computing Systems, Vancouver, May 1996.

19. Yang, Y. and Levine, M. "The Background Primal Sketch: An Approach for Tracking Moving Objects." Machine Vision and Applications, vol. 5, pp.17-34, 1992.