3D Tele-Collaboration Over Internet2

Herman Towles, Wei-Chao Chen, Ruigang Yang, Sang-Uok Kum and Henry Fuchs - University of North Carolina at Chapel Hill
Nikhil Kelshikar, Jane Mulligan [*] and Kostas Daniilidis - University of Pennsylvania
Loring Holden and Bob Zeleznik - Brown University
Amela Sadagic and Jaron Lanier - Advanced Network and Services,Inc.

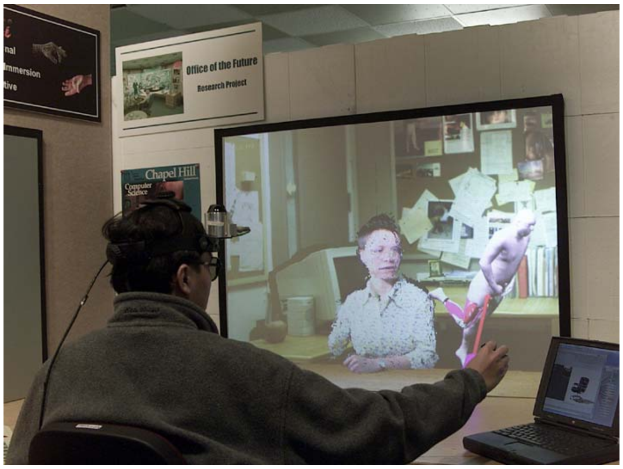

Figure 1: Our 3D tele-immersion realization with the local user manipulating the *visible human* data set using a virtual laser pointer. The local user sees the remote collaborator in perspectively correct 3D stereo, and both users are able to manipulate in 3D the shared visible human object.

Abstract

Our long-term vision is to provide a better every-day working environment, with high-fidelity scene reconstruction for life-sized 3D tele-collaboration. In particular, we want to provide the user with a true sense of presence with our remote collaborator and their real surroundings, and the ability to share and interact with 3D documents. The challenges related to this vision are enormous and involve many technical tradeoffs, particularly in scene reconstruction.

In this paper we present a significant step toward our ultimate goal. By assembling the best of available hardware and software technologies in scene reconstruction, rendering, and distributed scene graph software, members of the National Tele-Immersion Initiative (NTII) are able to demonstrate 3D collaborative, tele-presence over Internet2 between colleagues in remote offices.

Categories and Subject Descriptors: H.5.1[Multimedia Information System]: Video teleconferencing; **Keywords**: telepresence, tele-collaboration.

1 Introduction

We foresee a future[19] when we will be able to interact with our colleagues in any locale just as if they were across the table in our office. However, we are still years away from being able to realize a tele-immersion environment that will provide a true sense of social presence - a realistic feeling that we share and inhabit the same space with remote friends and colleagues. Today's conferencing systems are deficient in many ways. Cameras and displays lack the necessary resolution and field-of-view to create a true sense of immersion. It is not uncommon to experience an off-screen voice while using today's group conferencing systems. Participants are not shown life-size, and we find it difficult to make eye-contact with our collaborators because cameras and displays are not co-located. In group situations, audio is not spatialized. While we may be able to share whiteboards and 2D documents today, the human-computer interfaces (mouse and keyboard) we use to modify or markup these documents are not necessarily the way we would work with our colleagues if all were gathered at the same table. As a result, we have multiple human senses defining breaks in presense, making the experience less than natural.

With the formation of the National Tele-Immmersion Initiative (NTII) [17] researchers in computer vision, computer graphics and human-computer interfaces have joined forces to develop a new 3D tele-collaboration testbed. This system, shown in Figure 1, is capable of live, 3D scene reconstruction and view-dependent, stereo display while operating between remote sites over Internet2. In addition, participants can interact with shared, 3D objects with new human-computer interfaces such as a virtual laser pointer.

Although NTII research is a long-term effort, we believe the capabilities that we can demonstrate mark a significant step in 3D tele-collaboration systems. This paper documents our system and results. Our specific contributions are:

- **Integrated, Networked Collaborative Environment:** We combined state-of-the-art research components from computer vision, computer graphics and human-computer interaction to create one of the first systems designed to enhance the sense of social presence for tele-collaborative tasks.
- **True 3D Scene Representation:** We make no assumptions about the scene content by extracting true depth points with sufficient resolution and generality. This also allows us to realistically composite different types of 3D objects including static room background and 3D shared objects.
- **Accurate Immersive Display:** By using life-size, head-tracked stereo display, we create a portal to a remote place with a compelling continuum between the local and remote sites.

2 Related Work

To provide a better sense of presence today, conferencing products [23] and research systems such as the Access Grid (AG) [4] are available which provide a larger field-of-view and higher display resolutions. Most of these systems are multi-channel extensions of standard 'one-camera to one-display' products. Such system are also typically not calibrated to display the capture with exact scale.

To solve the gaze-awareness problem, some researchers have mounted the capture camera behind a display screen. The MONJUnoCHIE [1] system does this with a rear-projected image in a configuration similar to a tele-prompter. A new CAVE-based system [10] uses an electrically switchable glass screen allowing cameras to be positioned behind the screen. The IST VIRTUE project is an impressive solution [5, 24] that uses multiple cameras placed around the screen to generate a new, behind-the-screen virtual camera view using image-based rendering techniques for 2D-only display.

Several systems have been developed that can create 3D view-dependent imagery. CMU's Virtualized Reality [16] project has a 3D-room with 49 cameras used to capture and reconstruct persons in action. Their current system works off-line and is not a networked application. CMU also developed a hardware for real-time stereo reconstruction [8]. Matusik *et al.* [13] produce novel views based on silhouettes at interactive rates of 10-15Hz. Snow *et al.* [22] use a voxel occupancy algorithm for non-realtime reconstruction. The Keck laboratory at the University of Maryland has a 64 camera system for studying human action and off-line reconstruction [7]. The Multiple Perspective Interactive Video at UC-San Diego is used for off-line queries of video segments [9].

The University of Illinois at Chicago has made many significant contributions to tele-immersion research. Much of their work has focused on building tele-immersive data exploration (TIDE) [11] environments for collaborative exploration of complex 3D datasets via a networked array of CAVEs and ImmersaDesk systems. Participants are often represented by avatars rather than real 3D representations.

[*] now at the University of Colorado at Boulder.

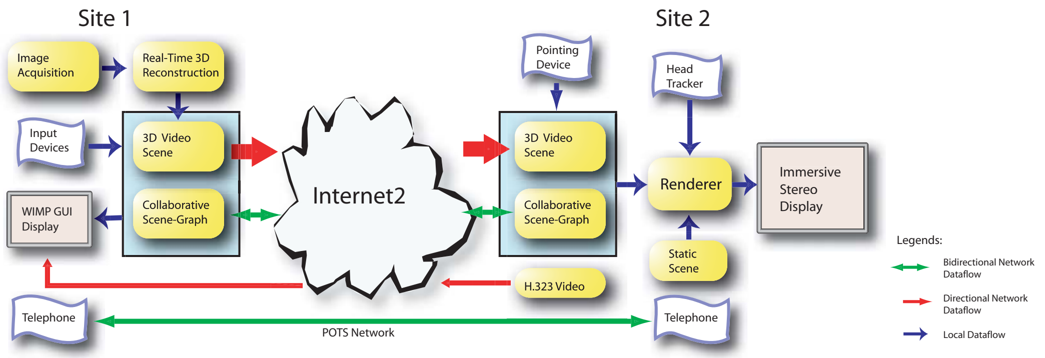

Figure 2: Block diagram of the NTII tele-immersion system.

At UNC we began our view-dependent, tele-immersion research focused on recreating the most compelling visual experience possible by removing the requirement for live, scene reconstruction[3]. We created a very high-fidelity, 3D static model of a real office scene using image-based techniques. Presented with a head-tracked stereo projection system, we were able to demonstrate a compelling portal to a static office , and more importantly to convince ourselves and many others that such a sense of presence, if interactive, would be very exciting.

3 System Overview

In developing our tele-immersion testbed, we were faced with a few simple and many complex design considerations involving hardware component selection, software architecture, and physical layout of the acquisition and rendering environment. Figure 2 illustrates the configuration of our current system. It allows one user on either side of the network. Primarily for economic reasons, our current system is asymmetric. The 3D video scene is reconstructed on Site 1 and displayed on Site 2. A conventional H.323 video feed is provided from Site 2 to Site 1. On the display side, Site 1 uses traditional monitor display, whereas Site 2 incorporates our head-tracked passive stereo display system proven to be effective in[3]. Because of the lack of a good solution for multi-user stereo display and the difficulty in scaling up 3D reconstruction system, we currently allow one user on Site 2.

In order to reconstruct real-time 3D video, we implemented trinocular stereo algorithm which provides higher quality reconstruction than binocular stereo. To obtain a more complete coverage and to reduce occlusion, we run multiple stereo algorithms with images acquired from around the user. Our reconstruction system does a static background subtraction and the reconstruction volume is approximately one cubic meter. This reconstruction volume is large enough to surround a person in our ‘across the desk’ one-on-one tele-collaboration scenario. To avoid rendering the 3D video into a purely virtual set, the renderer composites the foreground reconstruction of the remote participant into image-based 3D office model [3]. All scenes are transformed into a unified coordinate system before composition and rendering.

To achieve realistic test conditions across Internet2, the image acquisition and 3D reconstruction configuration was duplicated at 3 sites - Advanced Network and Services in Armonk, NY, UPenn in Philadelphia, PA, and UNC in Chapel Hill, NC. The immersive display system is only present in UNC. With an initial goal of 5 FPS at 320×240 camera image resolution, the system generates up to 75 Mbps one-way network traffic. We paid much attention to ‘last-mile’ network issues at the test locations to ensure sufficient bandwidth in our testbed.

The collaborative tool in our current system is a variant of [26], which can be thought of as a gesture-based 3D whiteboard, wherein both users share the same scene graph that can be very quickly sketched on and edited using pointing devices. A user at Site 2 uses a 3D pointing device imitating a virtual laser pointer. While we eventually plan to add multi-channel, spatialized audio, today’s system uses full-duplex, echo-cancellation speakerphones via the H.323 link or plain-old-telephone-service (POTS).

In summary, our tele-immersion testbed provides:

- ‘One-on-One’ 3D tele-immersion experience.
- Life-size, view-dependent, passive-stereo display.
- Shared 3D data objects manipulated with virtual laser pointer.
- Half-duplex 3D operation today.
- Operation between three sites over Internet2.
- Audio over H.323 or POTS.

4 Real-Time Acquisition and Reconstruction

This section describes the conversion of 2D images into a 3D video stream. Our image acquisition setup captures multiple camera images synchronously, which serve as input for the real-time 3D reconstruction system. The resulting 3D video is then transmitted over Internet2.

4.1 Image Acquisition

Our tele-immersion system currently operates in half-duplex mode, which means 3D acquisition and 3D display are not currently co-located. As such, camera layout is unconstrained by display placement. We have explored several camera arrangements, including two shown in Figure 3. In both cases, the cameras are arranged on a horizontal arc surrounding the reconstruction volume. In the seven camera array, five triples (or views) are formed by sharing two cameras in adjacent triples.

We use Sony 1394 digital color cameras with progressive scan. The cameras are connected to five quad-processor (550MHz Pentium III) servers running Windows 2000. Each machine captures an image triple and produces a depth map - one *reconstruction view* per server. For accurate 3D reconstruction, one of these machines also acts as the external (hardware) trigger server to synchronize image exposure of all cameras.

4.2 3D Reconstruction

Correspondence matching is the primary computational bottleneck in stereo algorithms. We have investigated a number of techniques for our application. A review of these techniques can be found in [20]. We conclude that trinocular Modified Normalized Cross Correlation (MNCC) is best suited for our tele-immersion system [14].

The cameras are first calibrated and external parameters all registered to the same coordinate system. Our camera calibration technique is built upon the calibration toolbox in [2]. The three input images are then rectified pairwise so that epipolar lines lie along the same horizontal image lines to facilitate the search for correspondences.

Non-parallel Trinocular Configurations – The trinocular epipolar constraint is a well known technique to refine or verify correspondences, and to improve the quality of stereo range data. It is based on the fact that for a hypothesized match in a pair of images, there is a unique location we can predict in the third camera image where we expect to find evidence of the same world point [6].

Correlation Matching – The modified normalized cross-correlation (MNCC) is defined as

$$corr_{MNCC}(I_L, I_R) = 2\ \mathrm{cov}(I_L, I_R)/(\sigma^2(I_L)\ +\ \sigma^2(I_R)). \quad (1)$$

where I_L and I_R are the left and right rectified images over the selected correlation windows. For each pixel (u, v) in the left image, the metrics above produce a correlation profile $c(u, v, d)$ where disparity d ranges over acceptable integer values. The maximum value in this profile is treated as a match. Foreground-background segmentation [12, 15] allows us to consider less than half of the pixels in the reference image. A detailed exposition of the trinocular algorithm can be found in [15].

Reconstruction Results – The correlation and selection procedure produces a disparity map. We further apply median filtering to remove a few outliers in the disparity map. With knowledge of the camera calibration, the median filtered disparity map combined with a registered

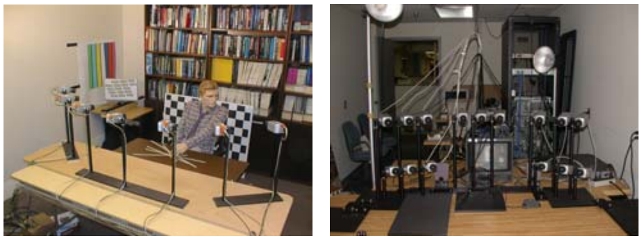

Figure 3: Camera acquisition arrays for 5 trinocular camera triples. (Left) Seven camera setup. (Right) 15 camera setup

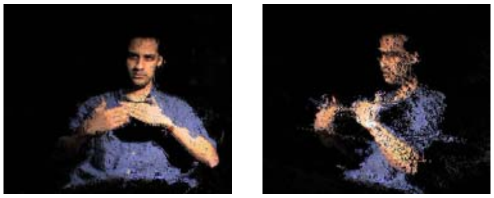

Figure 4: 3D Point Clouds from two viewpoints.

color texture are used to reconstruct a 3D point cloud. Figure 4(c) and Figure 4(d) show the point cloud produced from a typical image triple from two different view perspectives.

4.3 3D Video Transmission

Each of the several acquisition views generate a color point cloud using a reference image C. To reduce the raw data size, instead of explicitly storing the 3D coordinate of these points, we may instead represent them from the camera pose of C as depth images, which consists of a 2D color image(24bpp) and a disparity image(16bpp). The depth images and camera poses from all acquisition views are then transmitted over the network to the remote renderer. The 3D video data are tagged with frame IDs for synchronous playback on the remote site.

Our current assumption is one that foresees a future containing abundant network resources, and currently we are transmitting the 3D video data in an uncompressed format using TCP/IP. Clearly this can and will be improved in the future.

5 Collaborative Graphics

In our system, since the remote user is displayed using 3D graphics, we have the ability to augment the real-world with synthetic objects. This promises interactions that in many ways may be more powerful than what people in the same room can accomplish traditionally. For example, two doctors can bring up a virtual representation of a patient in order to discuss possible ways to perform some operation, make annotations, and save them later for educational purposes.

Scene Graph Sharing – Our current implementation takes the first steps in this direction. We modify the implementation of [26] to incorporate the virtual graphics into the renderer described in the next section. The virtual graphics are added to the renderer in a loosely coupled fashion by simply getting a callback when a frame needs to be rendered. We only use a subset of the original functionalities in [26] because we do not require sharing of scene graphs from external applications. The collaborative graphics subsystem shares its scene graph with that in the remote site by replicating object changes in real-time to remote copies. A review of networked virtual environment architectures, and a tutorial for standard methods of information sharing, can be found in [21].

3D Pointing Device – In our tele-immersion system, using a conventional 2D pointer such as a mouse to interact with the scene graph would not exploit the full potential of the system. For this purpose, we custom-built a virtual laser pointer that instead of emitting a laser beam, draws a virtual laser ray in the virtual world. This pointer, which is tracked via an embedded magnetic tracker receiver, can be used not only to point to something in the environment, but pressing the button can be used to “skewer” an object in order to reposition or examine it.

6 Rendering and Display

One of the primary goals in the system is to immerse the user in an environment consisting of all the scene components mentioned so far. In this section we will discuss implementation issues of the renderer, and of our head-tracked stereo display system.

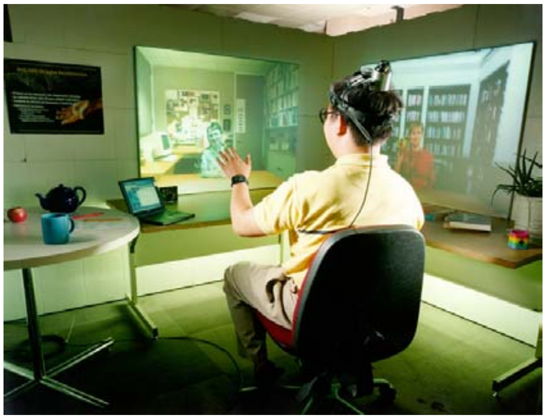

Figure 5: Demonstration at UNC of dual 3D tele-immersion sessions with Armonk and Philadelphia.

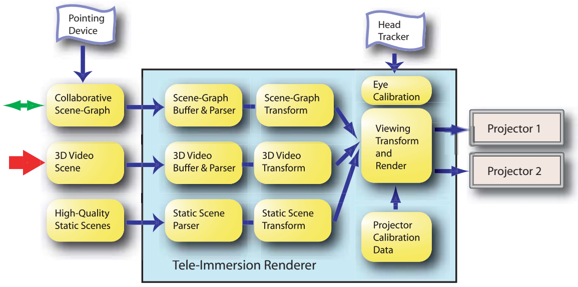

Figure 6: Rendering system overview. The renderer updates the display independently of the scene graph to improve user experience.

6.1 Display

Our tele-immersion display environment is built with a corner desk with two front-projective displays as shown in Figure 5. Each stereo display window is $1.2m \times 0.9m$ in size, large enough for life-size projection of our remote colleague at his/her desk.

The display surfaces are covered with a polarization-preserving fabric for passive stereo operation. To avoid the flicker issues of time-division multiplexed passive stereo, we use two projectors for stereo on each display surface, one for each eye. Our two-projector stereo solution delivers the full projector resolution to the display surface, and its 100% duty cycle produces a brighter display than other solutions. Our stereo solution uses circular polarization[3]. In practice we find crosstalk less of an issue with circular than linear polarization solutions.

6.2 Rendering

Figure 6 shows the block diagram of our implementation of the tele-immersion renderer. Each input scene is parsed and transformed into a global 3D coordinate system before being presented to user. The renderer takes the display system calibration and the tracker information as input to render stereo imagery from the user’s point of view.

Input scenes come from various sources, and their update rates are different. For example, the low-bandwidth collaborative scene graph updates much faster than the 3D video in our current implementation. Therefore, the renderer buffers the scenes asynchronously and presents the user with the latest scenes available. Rendering view-dependent imagery requires a constant update of the user’s eye-positions. For this purpose, we employ a HiBall^{TM}tracking system [25], which requires the viewer to wear a head-mounted optical sensor as seen in Figure 5.

We implemented the renderer using OpenGL API on commodity PC graphics hardware. There are two PC configurations that we currently use - a one-PC architecture and a three-PC architecture. The one-PC system consists of a 933MHz Pentium III PC with a nVidia GeForce2 MX graphics card. One multi-threaded renderer implements all functionalities in Figure 6. The three-PC system consists of three 933MHz Pentium III PCs, each equipped with nVidia GeForce2 QuadroPro graphics card. Two of the machines act as rendering machines, each implementing a renderer connected to only one projector. The third machine serves as the network aggregation point that multicasts the incoming scene data to the rendering machines. The aggregation machine also communicates with the rendering machines to synchronize scene redraw.

To create life-size imagery and to compensate for projector keystone distortion, we calibrate the projectors using an efficient method [18]. A 2D homography matrix is computed for each projector and incorporated into the rendering pipeline without extra rendering cost.

7 Results and Discussions

Because of the system complexity, we implemented and tested each subcomponent (reconstruction, rendering, scene graph) individually. Each sub-component can be incorporated or removed, depending on the application. The 3D reconstruction system alone runs at 2 to 3 fps producing 15K to 20K 3D points per view. The frame capture synchronization and TCP/IP network transport both have adverse impact on the system latency. Today, we see end-to-end latencies from 1.5 to 7 seconds.

Not all the delay is attributable to the front-end processes. On the receive/rendering side, all view frames must be received and parsed. These steps contribute less than 500 msec to the total latency in the one PC configuration. The delay is larger for the three-PC configuration because of the buffering between the aggregation and rendering machines. Clearly, overall latency could be reduced significantly with a pipelined design and the use of more efficient networking protocols.

Actual rendering loop performance obviously varies depending on the size of the 3D video scene being received, which may vary from only one to as many as five reconstruction views. Rendering performance ranges from 20-30 fps on the one-PC system, to 50-100 fps on the three-PC system. These rendering rates translate into a very responsive, view-dependent stereo presentation for the user. Collaborative scene graph operation does not require much processing and network bandwidth, thus operates smoothly at interactive rates up to 20Hz.

During the past year we have had many users experience our system, and we were able to observe their behavior and responses with the system; we are encouraged by their reactions. For example:

- **Ease of Use:** The system has a very intuitive user interface that includes only one virtual laser pointer. We observed that it is often self-explanatory and our users quickly learn to use the system.
- **Spontaneous Interactions:** Almost all users react to the system as if they are working with a person sitting across the table. They use gestures and body movements in the course of collaboration. Although many users do not change viewpoint very much, the view-dependent display create the illusion of immersion for the user into the virtualized world.
- **Extended Usage:** Almost no users complained about fatigue and dizziness due to extended usage of our system. In particular, one user with no prior experience of virtual reality systems used the system to conduct a remote interview for over an hour.

8 Conclusions and Future Work

With the development of our 3D tele-immersion testbed, we have seen a glimpse of what we foresee naturally-immersive environments are going to be like in the future. Buoyed by these successes, we are continuing to work on such issues as:

- Improved 3D reconstruction quality
- Reduce latency and susceptibility to network conditions
- Improved display quality and rendering performance
- Full-duplex operation
- Elimination of head tracker and eyeglasses
- Spatialized audio

The first three areas are our primary focus. Improved reconstruction quality means improving spatial resolution, frame rate, and a larger overall reconstruction volume, all of which are directly related to processing power and numbers of cameras. In this regard, we are planning to build a 60-camera acquisition theater at UNC and UPenn, and to apply the massive computing resources of the Pittsburgh Supercomputing Center to the real-time reconstruction task. Another important aspect is improving the quality of stereo correspondence. To this end, we are continuing to explore active lighting techniques that are imperceptible to the occupants. As the number of cameras increases, it is also important to develop rapid camera calibration and registration techniques.

To reduce operational latency, we must re-architect the end-to-end tele-immersion system and pipeline at several stages. More importantly, we are actively developing methods for online network characteristics monitoring and efficient high-level protocols that will allow the tele-immersion application to adapt to changing network conditions.

Thirdly we are actively working on new distributed rendering architectures capable of providing faster and higher resolution display walls built with commodity hardware.

Finally, the major advantage of a 3D tele-immersive system is the natural interaction to the remote world. With the current display setup, we have difficulty making a full-duplex system because users cannot easily establish eye-contact wearing the passive stereo glasses, and wearing a head-mounted tracker makes the system unfeasible for everyday use. We are investigating other options including auto-stereoscopic displays that will make our current system more practical.

References

[1] T. Aoki, K. Widoyo, and H. Yasuda. MONJUnoCHIE System: A Next-generation Videoconference System Supporting High Sense of Reality and User-friendly Information Sharing. In *IPSJ SIGNotes AV and Multimedia Information Processing*, 1998.

[2] J.-Y. Bouguet. Camera Calibration Toolbox for Matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc`.

[3] W.-C. Chen, H. Towles, L. Nyland, G. Welch, and H. Fuchs. Toward a Compelling Sensation of Telepresence: Demonstrating a Portal to a Distant (Static) Office. In *IEEE Visualization 2000*, pages 327–333, October 2000.

[4] L. Childers, T. Disz, R. Olson, M. E. Papka, R. Stevens, and T. Udeshi. Access Grid: Immersive Group-to-Group Collaborative Visualization, 2000.

[5] E. Cooke, P. Kauff, and O. Schreer. Imaged-Based Rendering for Tele-Conference Systems. In *Proceedings of WSCG 2002*, February 2002.

[6] U. Dhond and J. K. Aggarwal. Structure from Stereo — A Review. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1489–1510, November 1989.

[7] D. M. Gavrila and L. S. Davis. 3-d model-based tracking of humans in action: a multiview approach. In *Proceedings of IEEE Computer Vision and Pattern Recognition1996*, pages 73–80, 1996.

[8] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka. A stereo machine for video-rate dense depth mapping and its new applications. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 196–202, June 1996.

[9] P.H. Kelly, A. Katkere, D.Y. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain. An architecture for multiple perspective interactive video. In *ACM Multimedia*, pages 201–212, 1995.

[10] A. M. Kunz and C. P. Spagno. Technical System for Collaborative Work. In *Proceedings of Workshop on Virtual Environments 2002*, May 2002.

[11] J. Leigh, A. Johnson, T. DeFanti, S. Bailey, and R. Grossman. A Tele-Immersive Environment for Collaborative Exploratory Analysis of Massive Data Sets. In *Proceedings of ASCI 99*, pages 3–9, June 1999.

[12] F. C. M. Martins, B. R. Nickerson, V. Bostrom, and R. Hazra. Implementation of a Real-time Foreground/Background Segmentation System on the Intel Architecture. In *IEEE ICCV1999 Frame-Rate Workshop*, September 1999.

[13] W. Matusik, C. Buheler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. In *Proceedings of ACM SIGGRAPH*, 2000.

[14] J. Mulligan and K. Daniilidis. Trinocular Stereo for Non-Parallel Configurations. In *Proceedings of ICPR2000*, September 2000.

[15] J. Mulligan and K. Daniilidis. Trinocular Stereo: A New Algorithm and its Evaluation. *International Journal for Computer Vision*, 47(1):51–61, April 2002.

[16] P. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *Proc. Int. Conf. on Computer Vision*, pages 3–10, 1998.

[17] Advanced Network and Services, Inc. `http://www.advanced.org`.

[18] R. Raskar. Immersive Planar Displays using Roughly Aligned Projectors. In *IEEE VR 2000*, pages 109–116, March 2000.

[19] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In *ACM SIGGRAPH 1998*, pages 179–188, July 1998.

[20] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal for Computer Vision*, 47(1):7–42, May 2002.

[21] S. Singhal and M. Zyda. *Networked Virtual Environments - Design and Implementation*. ACM Press Books, SIGGRAPH Series, 1999.

[22] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *IEEE Conf. Computer Vision and Pattern Recognition*, June 2000.

[23] TeleSuite Corporation. `http://www.telesuite.com/`.

[24] VIRTUE Project. `http://bs.hhi.de/projects/VIRTUE.htm`.

[25] G. Welch and G. Bishop. SCAAT: Incremental Tracking with Incomplete Information. In *ACM SIGGRAPH 1997*, pages 333–344, August 1997.

[26] B. Zeleznik, L. Holden, M. Capps, H. Abrams, and T. Miller. Scene-Graph-As-Bus: Collaboration between Heterogeneous Stand-alone 3-D Graphical Applications. In *Eurographics 2000*, volume 19, 3, pages 91–98, August 2000.