Virtual Space Teleconferencing using a Sea of Cameras

Henry Fuchs, Gary Bishop, Kevin Arthur, Leonard McMillan
University of North Carolina at Chapel Hill*

Ruzena Bajcsy, Sang Wook Lee, Hany Farid
University of Pennsylvania†

Takeo Kanade
Carnegie Mellon University‡

Abstract

A new approach to telepresence is presented in which a multitude of stationary cameras are used to acquire both photometric and depth information. A virtual environment is constructed by displaying the acquired data from the remote site in accordance with the head position and orientation of a local participant. Shown are preliminary results of a depth image of a human subject calculated from 11 closely spaced video camera positions. A user wearing a head-mounted display walks around this 3D data that has been inserted into a 3D model of a simple room. Future systems based on this approach may exhibit more natural and intuitive interaction among participants than current 2D teleconferencing systems.

1 Introduction

In the near future, immersive stereo displays, three-dimensional sound, and tactile feedback will be increasingly capable of providing a sensation of *presence* in a virtual environment [Sutherland, 1968; Bishop *et al.*, 1992]. When this technology is applied for use in long-range communication, the goal is to provide a sense of *telepresence* to the participant.

The true promise of telepresence lies in its potential for enhanced interactivity and increased flexibility when compared to alternative technologies, such as conventional teleconferencing. Telepresence should not merely allow the viewing of remote participants, but it should also allow us to participate in the same space with them. This experience includes the ability to glance in the direction that a speaker points, or look at the other participants, or even stare at the ceiling.

Applications for telepresence are far-reaching. They include such tasks as remote medical diagnosis, instruction, and entertainment.

In this paper we present an approach to telepresence that uses stereo correlation techniques to extract a dense 3D description of a remote scene for presentation to the user. We discuss alternative approaches that other researchers have taken and compare both their approaches and design goals with our own. We present preliminary results of our efforts and conclude with a discussion of open issues and directions for future work.

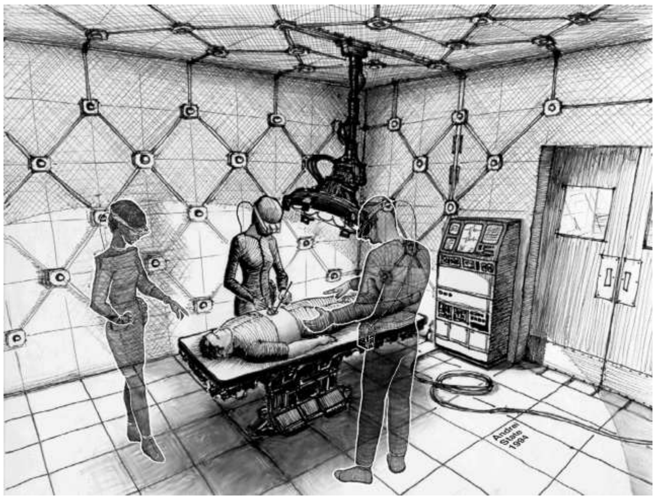

Figure 1: Concept sketch of a medical tele-consultation scenario.

2 Our Approach

We want to achieve a telepresence capability with

*Department of Computer Science, Chapel Hill, NC 27599-3175; e-mail {fuchs, gb, arthur, mcmillan}@cs.unc.edu.

†GRASP Laboratory, Department of Computer and Information Science, Philadelphia, PA 19104-6228; e-mail {bajcsy, swlee, farid}@grip.cis.upenn.edu.

‡The Robotics Institute, 5000 Forbes Avenue, Pittsburgh, PA 15213-3891; e-mail tk@cs.cmu.edu.

- natural, intuitive interaction for each of the participants with little or no training requirements,
- non-intrusive system sensors and displays,
- independent control of gaze for each participant,
- support for multiple independent participants at each site and multiple (2 or more) sites.

These considerations lead us to systems in which each participant wears a head-mounted display to look around a remote environment whose surface geometries are continuously sensed by a multitude of video cameras mounted along the walls and ceiling, from which depth maps are extracted through cross-correlation stereo techniques [Fuchs and Neumann, 1993].

Views acquired from several cameras can then be processed and displayed on a head-mounted display with an integrated tracking system to provide images of the remote environment. In this model the cameras perform two functions: they individually provide photometric images, or textures, from different viewpoints within the scene, and, in combination, they are used to extract depth information.

The extraction of depth information from a set of two or more images is a well known problem in robotic vision [Barnard and Fischler, 1982; Dhond and Aggarwal, 1989]. In the traditional application one or two cameras mounted to a mobile platform are used to acquire depth information at each pixel within the overlapping region of the sensor arrays as the robot moves through the environment. This same basic approach has also been applied to computer graphics as a technique for static model generation [Koch, 1993]. Our approach differs in that a multitude of permanently mounted cameras are used to acquire dynamic, or more accurately throw-away, models of a scene.

In contrast to other applications of computer vision, for telepresence we desire precise and dense depth information in order to provide high visual fidelity in the resulting display, and are not concerned with recognizing or constructing models of the objects in the scene or with gathering information for navigation and collision-avoidance for autonomous vehicles.

We feel that the passive "sea-of-cameras" approach to telepresence has a number of advantages over other methods, and is very nearly feasible with current or soon-to-be-available technology. One of the advantages of the approach is that we need not generate a model for objects that are not seen at a given time, because we can generate the scene description from the point of view of the remote camera whose line of sight most nearly matches that of the participant.

3 Previous Work

Previous approaches to telepresence tend to fall into one of the following categories: (1) a remote system provides incremental updates to a locally maintained model [Caudell *et al.*, 1993; Ohya *et al.*, 1993; Terzopoulos and Waters, 1993], (2) dynamic textures are mapped onto an essentially static model [Hirose *et al.*, 1993], (3) images from a multicamera conference room are projected onto a large field-of-view display, and (4) a boom mounted stereo camera pair is controlled by the movement of a remote observer wearing a head-mounted display.

The first two approaches assume that the remote geometry is largely static or constrained to move along predefined paths. Both approaches require a high-level understanding of the scene's composition, and they require that the scene be broken down into its constituent parts. For instance, each individual person and object in the scene must be modeled. The use of these techniques has been practically limited to human models and simple objects and environments that have been previously digitized.

The third and fourth approaches allow for both dynamic and unconstrained remote environments. A multi-camera wide-angle teleconference allows for limited gaze redirection, but it does not provide strong depth cues nor does it allow for freedom of movement within the environment. We believe that the types and ranges of interaction are greatly reduced by these limitations. Boom-mounted cameras, on the other hand, are perhaps the closest approximation to an ideal telepresence set-up. They provide for both stereoscopic and motion parallax depth cues and require little auxiliary processing of the acquired images. The disadvantages of using boom-mounted cameras are related to mechanical limitations of the system; the boom should have nearly the same range of motion as an actual observer, and the motion of the boom-mounted cameras is intrusive and perhaps even dangerous for the applications we envision. In our proposed approach, there is no remote camera positioning system that attempts to mimic the local participant's head movements. In effect there are only "virtual cameras" that correspond to the participant's eye positions in the environment.

4 Depth Acquisition

4.1 Overview

The major steps in recovering depth information from a pair or sequence of images are: (1) preprocessing, (2) matching, and (3) recovering depth (see [Dhond and Aggarwal, 1989] for a review of stereo algorithms). The preprocessing stage generally consists of a rectification step that accounts for lens distortion and non-parallel axis camera geometry [Tsai, 1987; Weng *et al.*, 1992b]. The process of matching is the most important and difficult stage in most stereo algorithms. The matching process determines correspondence between "features" that are projections of the same physical entity in each view. Matching strategies may be categorized by the primitives used for matching (e.g. features or intensity) and the imaging geometry (e.g. parallel or non-parallel optical axis). Once the correspondence between "features" has been established, calculating the depth is usually a straightforward computation dependent on the camera configuration and optics.

One of the most common stereo reconstruction paradigms is matching image features from two parallel axis views (see [Weng *et al.*, 1992a] for a review). This method provides a disparity value d for matched pairs of points for each point in either the left or right image. The depth z can then be recovered by the well known equation: $z = \frac{fb}{d}$, where f is the focal length of the pin-hole camera model and the baseline b is the distance between the focal points of the two cameras. This approach to recovering stereo is attractive because of its simplicity; however, recovering an accurate, dense 3D depth map with this procedure has proven to be a formidable task.

Two standard parameters that most stereo algorithms vary are the *baseline*, and the *mask size* over which correlation is performed. Varying these parameters affects different properties of the recovered depth map. In particular, a large mask size will result in a high density depth map (i.e. good recovery in the absence of "features") but poor localization of features in the 'x' and 'y' dimension. A large baseline allows for high resolution in the 'z' dimension, but increases the likelihood of errors due to occluding boundaries and repetitive patterns in the scene.

A stereo algorithm is presented that attempts to exploit, maximally, the benefits of small and large baselines and mask sizes. In particular, a *multi-baseline, coarse-to-fine* approach to stereo is adopted [Okutomi and Kanade, 1993; Kanade, 1993; Farid *et al.*, 1994], where several closely spaced views are taken (multi-baseline) and matching across these views is done for several different mask sizes (coarse-to-fine). The use of several views and mask sizes introduces a need for more sophisticated *matching* and *combination* strategies. Several such control strategies are introduced for matching across the multi-baseline which greatly reduce errors due to repetitive patterns and false matches that arise from specularity and occluding boundaries. Control strategies are also introduced for combining information across varying mask sizes which lead to dense, high resolution depth maps.

The following sections describe the details of the method we have used to achieve our preliminary results presented in Section 6.2.

4.2 Intensity Matching

In order to recover dense depth maps, intensity matching, as opposed to feature matching, is used in the stereo algorithm presented in this section. In particular, matching correlation error is given as the sum of absolute value of differences of intensities over a sampling window:

$$\sum_{x=1}^{n}\sum_{y=1}^{n} \frac{\mid I(x,y) - \hat{I}(x,y) \mid}{n^2} \qquad (1)$$

where I and $\hat{I}$ are the intensity values in the images being matched and n is the dimension of the square mask size over which correlation is performed.

4.3 Wide-Baseline Stereo

In order to take advantage of the benefits of using a small and large baseline, matching may be performed over a sequence of images. There are several strategies that may be adopted for matching across such a sequence of images; below, we present one such approach.

Whereas the original multi-baseline stereo algorithms [Okutomi and Kanade, 1993; Kanade, 1993] perform correlation to the left- or right-most image in a sequence of images, the algorithm described here correlates to the center image in the sequence. Correlating to the center view, in effect, reduces the baseline by a factor of two thus making errors due to occlusion, etc. less likely. The benefit of the full baseline is partially recovered later, as will be described below.

Consider for the moment the right half of a seven image sequence $(L_1, L_2, L_3, C, R_1, R_2, R_3)$, that is, images C through R_3. The matching point of a point P_0 in image C can be determined in image R_1 by searching along an epi-polar line. Let the point P_1 be the matching point in image R_1. The matching point for P_1 in image R_2 can then be determined by searching about an epi-polar line centered at the projection of P_1 in image R_2. Finally, the matching point for P_2 in image R_3 can be determined by searching about a epi-polar line centered at the projection of P_2 in image R_3. The disparity for P_0 is then simply $P_3^x - P_0^x$, where P_i^x is the x component of the point P_i. [1] In order to avoid errors due to occlusion, if the correlation error of a point in image P_i is above a pre-defined threshold, then the previously matched point P_{i-1} is directly projected into the last image in the sequence.

The projection of points is trivial given a known distance between neighboring images in the sequence. Given an image sequence with n images, a point P_i in image i is projected into image $i+1$ as follows:

$$P_{i+1} = P_i \times ((i+1) \times \frac{n}{2})/(i-n) \qquad (2)$$

Errors in the projection can be compensated for by increasing the search neighborhood about the projection point.

The process of computing disparity for a single point is repeated for each point in image C, resulting in a disparity map relating points in image C to those in image R_3. The process is then repeated to compute a

[1] This assumes parallel axis camera geometry.

disparity map relating points in image C to those in image L_3.

In order to take advantage of the full baseline (image L_3 to R_3), it is necessary to "combine" the left and right disparity maps. In an ideal world these maps would be identical and simply adding them would suffice. However, due to occlusions, noise, intensity variations, false matches, etc. this approach is unrealistic and results in a large number of errors. Hence a simple "combination rule" to combine the left and right disparity map is adopted on a per-pixel basis:

$$\begin{gathered}\text{if } (|D_L - D_R| < \varepsilon_D \text{ and } |C_L - C_R| < \varepsilon_C) \text{ then}\\ D_F = (D_L + D_R)\\ \text{else if } (C_L < C_R\ ) \text{ then } D_F = 2 \times D_L\\ \text{else } D_F = 2 \times D_R\end{gathered}$$

where, D_L and D_R corresponds to the left and right disparity maps, respectively, C_L and C_R correspond to the left and right correlation errors, respectively and D_F corresponds to the final disparity value. ε_D and ε_C are pre-defined thresholds set to a value of 1 in the results presented in Section 6.2. These two thresholds dictate the error tolerance between the left and right disparity maps.

To this point correlation has been performed only over a single mask size. In order to benefit from the properties of correlating over a large and small mask size, disparity maps are computed for a number of mask sizes. In particular, using the process described in the previous section, disparity maps are computed for mask sizes ranging from 3×3 to 15×15. Associated with each of these disparity maps is a correlation map, which associates a correlation value with each point in the image. The final disparity map is computed by initially setting all disparity values to be that of the coarsest map (15×15 mask). Each point in the final disparity map is then updated through the smaller mask sizes as long as the correlation error of a smaller mask is less than or equal to the correlation error of a larger mask.

5 Display

After computing a depth map, we desire to use this information along with the captured images to provide an effective 3D presentation to the user.

Individual depth maps may be displayed in the conventional manner as 3D height fields with the corresponding image texture mapped onto the geometry. For a given camera view, we create a polygonal model that will look correct when viewed from that camera position or nearby, and the view will degrade as the user moves farther away from the correct position. The program is capable of switching to a better camera view (and corresponding depth map) as the user walks about the room. Future work will investigate combining multiple depth maps to create a larger model that is more consistent for a wider range of views.

6 Preliminary Results

We have conducted two different experiments in an attempt to test the sea-of-cameras approach.

6.1 Geometry from Synthetic Cameras

In the first experiment a ray tracing program was used as a synthetic camera. The ray tracing program was modified to output the depth information as well as the reflected color at each pixel. Using these synthetic camera models, we were able to construct the geometry of an environment using only the depth information available from a single camera's point of view.

Figures 3 to 5 show images of an operating room scene containing high resolution scanned human models obtained from Cyberware, and a user walking around this scene and wearing a head-mounted display. The dimensions of the ray-traced images are 240×320. Ideal depth information from the ray tracer was used along with four synthetic camera views to display the environment. At any particular time, the geometry corresponding to a single camera's view is displayed and the camera view chosen depends on the user's position and orientation as he walks about the room.

6.2 Depth from Acquired Images

In the second experiment we employed the multi-baseline stereo method described earlier on a series of images of a human subject. In order to obtain a sequence of images while insuring parallel axis motion, a CCD camera (Sony XC-77RR, 25mm lens) is mounted on the end effector of a PUMA 560. An image sequence is then obtained by repeatedly moving the PUMA a fixed distance horizontally in front of an object, and digitizing an image at each step.

An 11-image sequence of the upper torso of a human subject was taken. The camera was translated 3 cm between successive views, giving a full baseline of 30 cm. The subject was approximately 1 m from the camera. The stereo reconstruction algorithm described in Section 4 was run on the subsampled 256×256 images (images were originally 512×512). The resulting depth map was postprocessed using a 15×15 Gaussian filter for smoothing. Figure 2 shows three images from the sequence and the resulting depth map.

For display, a 3D polygonal mesh was created by texture-mapping the image from the center camera's view onto a height map. Figures 6 to 8 show scenes of a user walking around this reconstructed model within a virtual room of predetermined geometry.

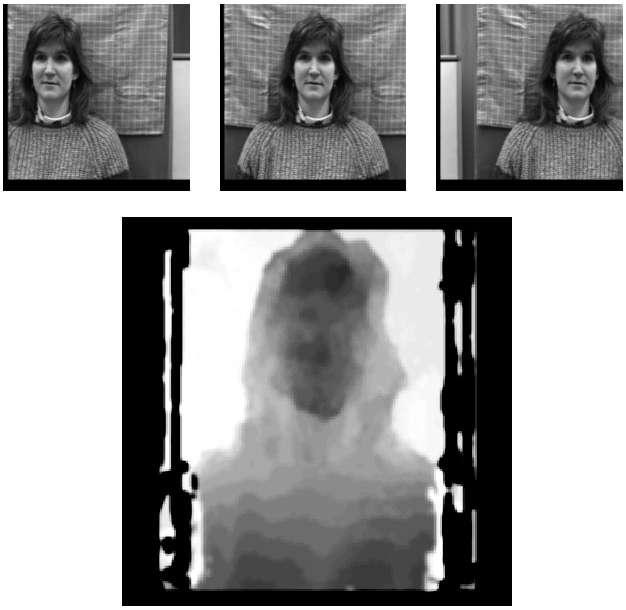

Figure 2: Subset of image sequence (top row) and depth map.

7 Discussion

Although we believe our preliminary results show the promise in our approach to telepresence, there are a number of problems and uncertainties that remain to be addressed.

One concern is whether it will be possible with multi-baseline stereo (or other) methods to construct depth maps of high enough accuracy and resolution for applications requiring high visual fidelity. Furthermore, having fixed cameras in the environment will make the method susceptible to obscuration of important parts of the scene by objects or people. These problems may be alleviated somewhat by the use of a movable array of cameras attached to a work light that can be positioned to provide accurate and unoccluded data for a particular region of interest.

The extra time required to acquire and process the captured images imposes significant demands that exceed those of conventional video teleconferencing or conventional computer graphics and virtual reality applications. To address this, one of the authors is directing an effort at Carnegie Mellon University to develop a video-rate (30 frames per second) stereo machine that can take up to 6 camera inputs [Kanade, 1993].

We believe that these issues can be resolved over time, and that the cost of the technology will fall, eventually making the proposed system practical for real use.

8 Conclusions

Our initial experiments show promise for the sea-of-cameras approach to virtual space teleconferencing. We are hopeful that sufficient computational resources may be acquired in the next few years to achieve an interactive-rate system. Such a system would qualitatively alter the presentation and mode of interaction in telepresence applications and would stand in stark contrast to today's conventional 2D teleconferencing. Among the variety of applications are also ones that don't require real-time depth extraction – movies and training "films" in which 3D dynamic environments can be extracted off-line for later "walk-around" visualization. We are excited at the prospect that once someone will have experienced this new approach, he or she will not willingly go back to a conventional 2D system.

9 Acknowledgements

Funding at the University of North Carolina was provided by ARPA contract DABT63-93-C-C048 and the NSF/ARPA Science and Technology Center for Computer Graphics and Scientific Visualization, NSF Prime Contract Number 8920219. Funding at the University of Pennsylvania was provided by ARPA contract N00014-92-J-1647, ARO contract DAAL03-89-C-0031PRI, and NSF contracts CISE/CDA-88-22719, CDA91-21973, and GER93-55018.

References

[Barnard and Fischler, 1982] S. T. Barnard and M. A. Fischler. Computational stereo. *Computing Surveys*, 14(4):553–572, 1982.

[Bishop *et al.*, 1992] G. Bishop, H. Fuchs, et al. Research directions in virtual environments. *Computer Graphics*, 26(3):153–177, August 1992.

[Caudell *et al.*, 1993] T. P. Caudell, A. L. Janin, and S. K. Johnson. Neural modelling of face animation for telecommuting in virtual reality. In *Proceedings of the IEEE Virtual Reality Annual International Symposium*, pages 478–485, September 18-22, 1993.

[Dhond and Aggarwal, 1989] U. R. Dhond and J. K. Aggarwal. Structure from stereo - a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19(6):1489–1510, 1989.

[Farid *et al.*, 1994] H. Farid, S. W. Lee, and R. Bajcsy. View selection strategies for multi-view, wide-baseline stereo. Technical Report MS-CIS-94-18, University of Pennsylvania, Department of Computer and Information Science, 1994.

[Fuchs and Neumann, 1993] H. Fuchs and U. Neumann. A vision of telepresence for medical consultation and other applications. In *Proceedings of the 6th International Symposium on Robotics Research*, October 2-5, 1993. To appear.

[Hirose *et al.*, 1993] M. Hirose, K. Yokoyama, and S. Sato. Transmission of realistic sensation: Development of a virtual dome. In *Proceedings of the IEEE Virtual Reality Annual International Symposium*, pages 125–131, September 18-22, 1993.

[Kanade, 1993] T. Kanade. Very fast 3-d sensing hardware. In *Proceedings of the 6th International Symposium on Robotics Research*, October 2-5, 1993. To appear.

[Koch, 1993] R. Koch. Automatic reconstruction of buildings from stereoscopic image sequences. *Proceedings of Eurographics '93*, 12(3):339–350, 1993.

[Ohya *et al.*, 1993] J. Ohya, Y. Kitamura, H. Takemura, F. Kishino, and N. Terashima. Real-time reproduction of 3d human images in virtual space teleconferencing. In *Proceedings of the IEEE Virtual Reality Annual International Symposium*, pages 408–414, September 18-22, 1993.

[Okutomi and Kanade, 1993] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, 1993.

[Sutherland, 1968] I. Sutherland. A head-mounted three dimensional display. In *Proceedings of the Fall Joint Computer Conference*, pages 757–764. Thompson Books, Washington, D.C., 1968.

[Terzopoulos and Waters, 1993] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[Tsai, 1987] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, 1987.

[Weng *et al.*, 1992a] J. Weng, N. Ahuja, and T. Huang. Matching two perspective views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):806–825, 1992.

[Weng *et al.*, 1992b] J. Weng, P. Cohen, and M. Herniou. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(10):965–980, 1992.

Figure 3: Simulated operating room scene, with head data courtesy of Cyberware and modelled body data.

Figure 4: Split-screen still frame from video footage of a user walking around the above scene. This shows the user (lower portion) and his view for one eye.

Figure 5: View of the same synthetic camera scene, but the user has moved forward in the room.

Figure 6: Scene containing depth map extracted from 11 camera views.

Figure 7: An off-axis view of the scene.

Figure 8: A closer view showing the resolution of the extracted geometry.