Chapter 1

REAL-TIME 3D-TELE-IMMERSION

K. Daniilidis*†, J. Mulligan, R. McKendall, D. Schmid
University of Pennsylvania
{kostas,janem,mcken,daschmid}@grip.cis.upenn.edu

G. Kamberova
Washington University
kamberov@cs.wustl.edu

R. Bajcsy
NSF CISE Directorate
rbajcsy@nsf.gov

Abstract

In this paper we present the first implementation of a new medium for tele-collaboration. The realized testbed consists of two tele-cubicles at two Internet nodes. At each tele-cubicle a stereo-rig is used to provide an accurate dense 3D-reconstruction of a person in action. The two real dynamic worlds are exchanged over the network and visualized stereoscopically. The remote communication and the dynamic nature of tele-collaboration raise the question of optimal representation for graphics and vision. We treat the issues of limited bandwidth, latency, and processing power with a tunable 3D-representation where the user can decide over the trade-off between delay and 3D-resolution by tuning the spatial resolution, the size of the working volume, and the uncertainty of reconstruction. Due to the limited number of cameras and displays our system can not provide the user with a surround-immersive feeling. However, it is the first system that uses *3D-real-data* that are reconstructed *online* at another site. The system has been implemented with low-cost off-the-shelf hardware and has been successfully demonstrated in local area networks.

1. INTRODUCTION

Advances in networking and processor performance open challenging new directions for remote collaboration via immersive environments. With the continuing progress in bandwidth and protocols for the information highway, new education and business structures become feasible. The incorporation of graphical models in remote training is already a reality: Two astronauts from two different continents can already train together in a virtual space-shuttle [Macedonia and Noll, 1998]. However, nothing that they see is real: they see each other as their graphical avatars and the space shuttle is a virtual model. The demand for

[*] To appear as a chapter in Ales Leonardis and Ruzena Bajcsy (Eds.) The Confluence of Computer Graphics and Vision, NATO Advanced Research Workshop Series, KLuwer Academic Publishers, 2000

[†] This work has been supported by Advanced Network and Services (National Tele-immersion Initiative), ARO/MURI DAAH04-96-1-0007, and NSF CDS-97-03220. The authors gratefully acknowledge the exciting collaboration with the leading scientist of the National Tele-immersion Initiative Jaron Lanier as well as with Henry Fuchs and the other colleagues at the University of North Carolina, Chapel Hill.

remote collaboration among expert physicians during complex operations or between engineers for virtual prototyping is increasing, too.

The purpose of this paper is to show, in the context of tele-immersion, the utility of an integrated approach coming from two fields: Computer Vision and Computer Graphics. The problem of tele-immersion, not only requires two different technologies: data acquisition/reconstruction (the typical domain of Computer Vision) and fast realistic and interactive data display (the typical domain of Computer Graphics) but also it requires rethinking some of the basic representations of the data in view of the constraints coming from the demands of real time, low latency, high spatiotemporal resolution, and low cost.

While the Computer Vision community is mainly concerned with scene reconstruction to be used in different tasks such as navigation/manipulation or recognition, here the goal is different. In tele-immersion applications, the goal is *communication* amongst people who are geographically distributed but are meeting in the space of each local user augmented by the real live avatar of the remote partner. This is quite different from the conventional virtual reality. What is most important is not the realism but the usefulness with respect to the task in hand, for example, collaboration or entertainment. It is also different from traditional off-line versions of image-based rendering which just replace virtual with real worlds. Therefore, the challenging issue for computer vision beside the representation is the real-time processing - which has long been a focus for the visualization and graphics community.

What will follow is a description of a fully integrated dynamic 3D telepresence system working over the network. The highlights of the system are:

1. Full reconstruction and transmission of dynamic *real* 3D-data in combination with any *virtual* object.
2. Real-time performance using off-the-shelf components.
3. Optimal balance between several quality factors (spatial resolution, depth resolution, work volume).

Why is 2D not enough ?. Nowadays, most advanced tele-conferencing and telepresence systems transmit 2D-images. In order to get additional views, the systems use either panoramic systems and/or interpolate between a set of views [Chen and Williams, 1993, Seitz and Dyer, 1996, Scharstein and Szeliski, 1996]. We argue here, that for collaboration purposes 3D-reconstruction can not be avoided. First, view morphing approaches are able to interpolate views over a very restricted range of weakly calibrated viewpoints. Second, even if a system is fully calibrated [Scharstein and Szeliski, 1996] we need a calibration between the observer tracker and the cameras. In a collaboration scenario, where multiple persons discuss real 3D properties of mechanical objects or even give instructions requiring 6DOF movements there is no camera placement constellation which can produce the required variability of viewpoints resulting from the head movements of a user and reflect the feeling of distances. Therefore, we pursue a 3D image based rendering which is viewpoint independent based on stereo reconstruction.

2. RELATED WORK

Here we are not going to review the huge number of existing papers (confer the annual bibliographies by Azriel Rosenfeld) on all aspects of stereo (the reader is referred to a standard review [Dhond and Aggrawal, 1989]). Application of stereo to image based rendering is very well discussed and reviewed in the recent paper by Narayanan and Kanade [Narayanan et al., 1998]. Although terms like virtualized reality and augmented reality are used in many reconstruction papers it should be emphasized that we address here a reactive telepresence problem whereas most image based rendering approaches try to replace a graphic model with a real one *off-line*.

Stereo approaches may be classified with respect to the matching as well as with respect to the reconstruction scheme. Regarding matching we differentiate between sparse feature based reconstructions (see treatise in [Faugeras, 1993]) and dense depth reconstructions [Okutomi and Kanade, 1993, Narayanan et al., 1998]. Approaches such as [Belhumeur, 1996, Tomasi and Manduchi, 1996] address the probabilistic nature of matching with particular emphasis on the occlusion problem. Area-based approaches [Matthies, 1992] are based on correlation and emphasize the real-time issue like our approach.

An approach with emphasis on virtualized reality is [Narayanan et al., 1998]. This system captures the action of a person from a dome of 51 cameras. The processing is off-line and in this sense there is no indication how it could be used in telepresence beside the off-line reconstruction of static structures.

With respect to reconstruction, recent approaches can be classified as strongly or weakly (or self-calibrated) approaches. Self-calibration approaches [Maybank and Faugeras, 1992] provide a metric reconstruction from multiple views with an accuracy which is suitable only for restricted augmented reality applications like video manipulation where the quality of depth is not relevant. Weakly calibrated approaches [Kutulakos and Vallino, 1998] provide real time performance and are suitable for augmenting scenes only with synthetic objects. Our approach is the first that provides an optimal balance between depth accuracy and speed and therefore can be applied in tele-immersion.

3. SYSTEM DESCRIPTION AND SCENARIO

The tele-immersion testbed we work with is a continuously evolving system. Before delving into the individual algorithms we will describe the first hardware lay-out realized in spring 1999. Each side consists of

1. a stereo rig of two CCD-cameras,
2. a PC with a frame grabber,
3. a PC with an accelerated graphics card capable of driving stereo-glasses synchronization.

The spring-99 version has an Intel Pentium-II 450 MHz and a Matrox-Genesis Frame Grabber at the local site (called A in Fig. 1.1). The latter includes the TI C80 processor as a component. The CD-cameras are the Sony XC-77. For visualization we use the Diamond FireGL-4000 board with CrystalEyes stereo-glasses. Both sites are connected to the network and send their data using the TCP/IP protocol. Implementation of networking is for the local area network and will be extended to include compression and to compensate for lossy transmission protocols in a wide area Internet2 connection.

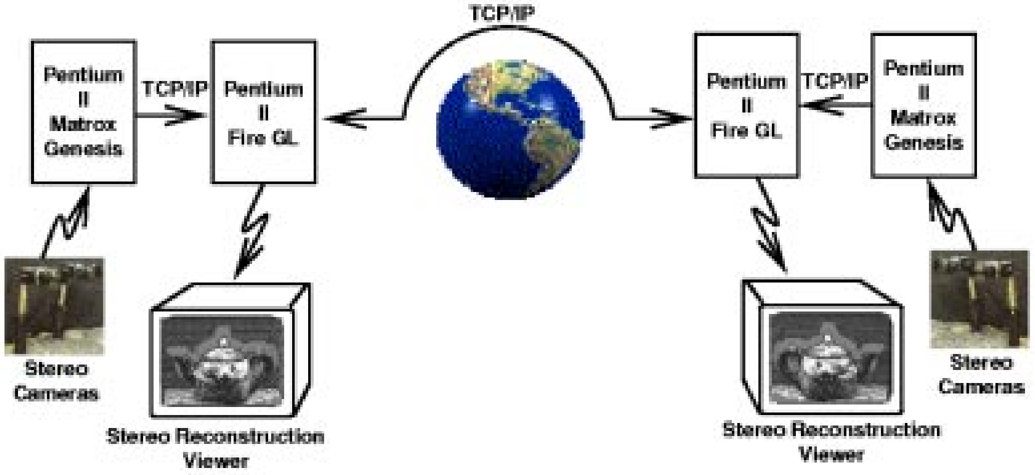
TCP/IP
Pentium II Matrox Genesis
TCP/IP
Pentium II Fire GL
Pentium II Fire GL
TCP/IP
Pentium II Matrox Genesis
Stereo Cameras
Stereo Reconstruction Viewer
Stereo Reconstruction Viewer
Stereo Cameras

Figure 1.1 First set-up for tele-immersion hardware.

The next generation of the system will integrate rendering and display technology contributed by Henry Fuchs and his co-workers at the University of North Carolina, Chapel Hill. The difference in the 3D scene acquisition will be a new surround configuration of seven cameras arranged in an arch of 120 deg. These seven cameras yield five overlapping stereo triples each of them connected to a four-Pentium-III multiprocessor workstation. The cameras will be digital and connected to the workstations via an IEEE 1394 interface.

The display system in a remote geographic location consists of a polarized stereo projector system and a wall as a display. The user's head will be magnetically tracked and the received 3D-scene will be rendered on an SGI engine. Life-size projection of the remote scene will maximize spatial augmentation and the feeling of sharing the same room.

Our innovation in this system will be a new trinocular stereo reconstruction algorithm. Our surround camera configuration produces image planes which cannot be warped into a common rectified plane. We describe in a later section a new algorithm and results based on such a non-rectifiable stereo configuration.

To minimize processing and transmission time the background is assumed to be stationary, reconstructed once initially, and then permanently subtracted from every incoming scene.

4. BINOCULAR STEREO RECONSTRUCTION

We elaborate next the main steps of the reconstruction algorithm and emphasize the factors that affect the quality of reconstruction and the processing time. Our reconstruction uses two images but it is easily extensible to a polynocular configuration. We rely on the well known stereo processing steps of matching and triangulation given that the cameras are calibrated.

Filtering. It is well known that two image patches can be matched if they contain sufficient gray-value variation. Since most of the matching steps are time consuming we want to avoid them if we know a-priori that there is not sufficient image structure to match. Therefore, we compute the image gradient at each position by convolving the image with a Gaussian derivative. A subsequent thresholding extracts the image areas with a high gradient. If the background is stationary we would like to avoid its reconstruction at each time-frame. A change detection method detects only the moving area on the image by thresholding the frame difference and post-filtering with morphological operators.

Rectification. When a 3D-point is projected onto the left and the right image plane of a fixating stereo-rig the difference in the image positions is both in horizontal and vertical directions. Given a point in the first image we can reduce the 2D-dimensional search to a 1D-dimensional if we know the so called *epipolar geometry* of the camera which is given from calibration. Because the subsequent step of correlation is area based and for reduction of time complexity we first perform a warping of the image that makes every epipolar line horizontal [Ayache and Hansen, 1988]. This image transformation is called *rectification* and results in corresponding points having coordinates (u, v) and $(u - d, v)$, in left and right rectified images, respectively, where d is the horizontal disparity.

Matching: disparity map computation. The degree of correspondence is measured by a modified normalized cross-correlation [Moravec, 1981],

$$mncc(I_L, I_R) = \frac{2\ cov(I_L, I_R)}{var(I_L)\ +\ var(I_R)}. \tag{1.1}$$

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 matching produces a correlation profile $c(u, v, d)$ where d ranges over a disparity range. The definition domain is the so called *disparity range* and depends on the depth of *working volume*, i.e. the range of possible depths we want to reconstruct. The time complexity of matching is linearly proportional to the size of the correlation window as well as to the disparity range.

We consider *all* peaks of the correlation profile as possible disparity hypotheses. This is different from other matching approaches which decide on the maximum of the matching criterion. We call the resulting list of hypotheses for all positions a *disparity volume*. The hypotheses in the disparity volume are pruned by a *selection procedure* that is based on the constraints imposed by

- Visibility: If a spatial point is visible then there can not be any other point on the viewing rays through this point and the left or right camera.
- Ordering: Depth ordering constrains the image positions in the rectified images. Both constraints can be formulated in terms of disparities without reconstructing the considered 3D-point [Yuille and Poggio, 1984, Dhond and Aggrawal, 1989].

The output of this procedure is an integer *disparity map*. To refine the 3-D position estimates, a *subpixel correction* of the integer disparity map is computed which results in a subpixel disparity map. The subpixel disparity can be obtained either using a simple interpolation of the scores or using a more general approach as described in [Devernay, 1994] which takes into account the distortion between left and right correlation windows, induced by the perspective projection, assuming that the surface can be locally approximated with a plane. The first approach is faster while the second gives a more reliable estimate of the subpixel disparity. We chose an extended version of the former which assumes preservation of the intensity value left and right. To achieve fast subpixel estimation and satisfactory accuracy we proceed as follows.

Let ϵ be the unknown subpixel correction. For corresponding pixels in the left and right images,

$$I_L(u,v) = \alpha I_R(u-d+\epsilon, v) = \alpha(I_L(u-d,v) + \epsilon \nabla I_L(u-d,v)) \tag{1.2}$$

where the coefficient α takes into account possible differences in camera gains. By taking a first order linear approximation of (1.2) over the correlation window we obtain the equivalent of a differential method for computing the optical flow. We use an FIR-filter-approximation of the image gradient appearing in the above formula. The disparity map is the input to the reconstruction procedure.

3D-Reconstruction. Each of the stereo rigs is calibrated before the experiment using a standard “strong” calibration technique [Tsai, 1987]. The calibration estimates the two 3x4 projection matrices for the left and the right camera. Given the disparity at each point and the calibration matrix the coordinates of a 3D-point can be computed.

From the disparity maps and the camera projection matrices the spatial positions of the 3D points are computed based on triangulation [Faugeras, 1993]. The result of the reconstruction (from a single stereo pair of images) is a list of spatial points.

The error in the reconstruction depends on the error in the disparity and the error in the calibration matrices. Since the working volume to be reconstructed is close to the origin of the world coordinate system the depth error due to calibration is negligible in comparison to the error in the disparities. What is mainly of concern is the number of outliers in the depth estimates usually appearing near occlusion or texture-less areas.

5. A NOVEL TRINOCULAR STEREO ALGORITHM

Reconstructions from a single stereo pair often have errors and extreme outliers due to ambiguity in matches along the epipolar line. For applications such as building detailed object models or creating models of humans for virtual environments, identifying and eliminating such points or patches is critical, but often difficult and expensive. One well known constraint for reducing these ambiguities is to add a third camera to verify hypothesized matches. The trinocular epipolar constraint in stereo vision is based on the fact that for a hypothesized match $[u, v, d]$ 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 [Dhond and Aggrawal, 1989]. A hypothesis is correct if the epipolar lines in the third camera image for the original point $[u, v]$ and the hypothesized match $[u-d, v]$, intersect.

The configurations we are interested in are similar to that depicted in Figure 1.2, where a sequence of cameras surrounds an object to be modeled or a user interacting with an augmented reality system.

We begin by independently rectifying the left and centre cameras (L and C_L) and the centre and right cameras (C_R and R), so that their epipolar lines are parallel respectively. For the right rectified camera

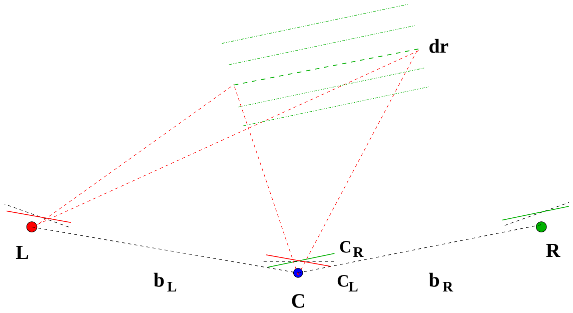

Figure 1.2 Trinocular camera triple.

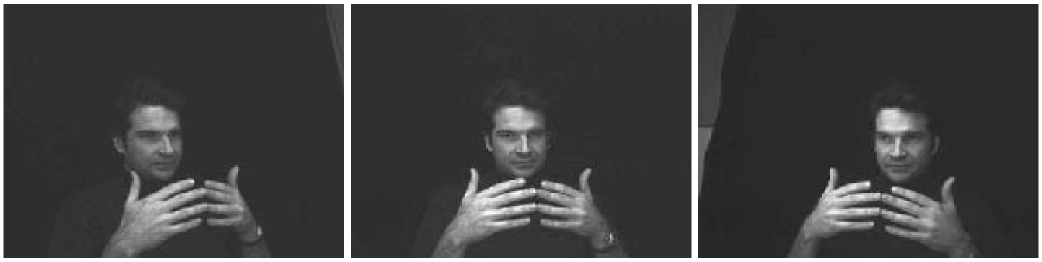

Figure 1.3 Three camera views.

pair every disparity d_R to be searched represents a plane with constant Z, which can be projected into the L and C_L images to compute the corresponding $[u_L, v_L, d_L]$ for each $[u_R, v_R, d_R]$. This straightforward application of the trinocular constraint is illustrated in Figure 1.2.

Of course for any Z-plane constructed from d_R, a range of d_L will be required to match points in the left pair. For example for the images used later, the right range $D_R = [-90, 10]$ corresponds to a left range $D_L = [-74, 67]$. Also because the two pairs are independently rectified, corresponding points in the left pair will not necessarily have $v_L = v_R$, thus all of u_L, v_L, and d_L depend on $[u_R, v_R, d_R]$. The calculation is simplified slightly by the fact that C_L and C_R are derived from the same image C and are related by the a priori rectification rotations R_{CL} and R_{CR}. We can thus precompute a lookup table of locations in C_L equivalent to those in C_R by precalculating $[u_{CL}, v_{CL}, s] = R_{CL}R_{CR}^{-1}[u_{CR}, v_{CR}, 1]^T$, for all image locations.

Our underlying matching measure is modified normalized cross correlation (MNCC) as defined in Sec. 4. Borrowing from [Okutomi and Kanade, 1993] the insight that we need to select matches based on minima (or maxima in the case of correlation) of the combined matching measure with respect to depth, we sum the MNCC values for corresponding $[u_R, v_R, d_R]$ and $[u_L, v_L, d_L]$ to obtain a correlation measure which now varies between -2 and $+2$.

Given the intrinsic and extrinsic camera parameters, and rectification matrices we can precalculate D_L the range of d_l generated by the plane implied by the current d_R. We calculate and store the right to left (C_L to L) correlation for the left pair, for all $d_L \in D_L$. This gives us a set c_L of $k = |D_L|$ planes of correlation values for the left centre image.

To evaluate a match at $[u_R, v_R, d_R]$, first we calculate $c_r = \text{MNCC}(C_R, R, d_R)$. For the left pair we calculate the location (u_{LL}, v_{LL}) of points on the depth plane in the left rectified image L. Using the precomputed lookup table we find the coordinates (u_{CL}, v_{CL}) and finally we can calculate the disparity $d_L = u_{LL} - u_{CL}$ for each point. Given the corresponding $[u_{CL}, v_{CL}, d_L]$ for each point in the centre right

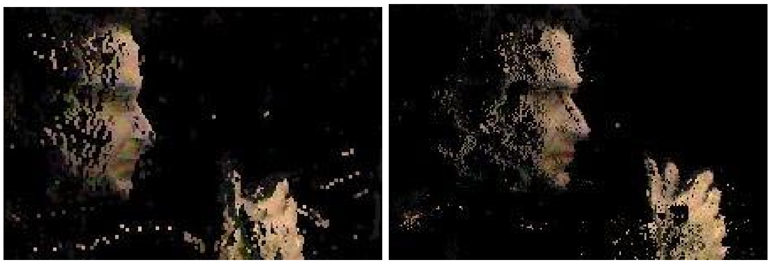

Figure 1.4 Reconstructed views for binocular (left) and trinocular (right) matching. rotated

image, we can look up the correlation value c_L at the specified location in the computed left correlation planes. We can now calculate our overall correspondence by $S_{corr} = c_L + c_R$.

In Fig. 1.4 we show results of binocular and trinocular reconstruction of the stereo triple in Fig. 1.3.

6. PERFORMANCE

We next present a listing of the timing of every algorithmic step for two exemplary parameter set-ups resulting in a frame-rate of 2Hz and 0.5Hz, respectively. The fast set-up has a quarter of the resolution of the original slow set-up as well as half of the working volume. The working volume in the slow set-up is 50cm at a distance of 1m of the camera. We do not mention the effective bandwidth of our network connection and the display speed because both of them are orders of magnitude faster than the reconstruction processing (for a local network application) and depend on the coding of the transmitted data.

Step	Fast setup	Slow setup
Total time	506ms	2080ms
Rectification	26ms	110ms
Filtering	32ms	90ms
Correlation and Selection	358ms	1460ms
Subpixel disparity	42ms	270ms
Reconstr. and Coloring	48ms	150ms

Table 1.1 Timings of each processing step in two different qualities

The real power of the system lies in the accuracy of the depth estimation without sacrificing time. We achieve a relative depth error of less than 0.1% at a distance of 1m (less than 1mm). Comparison with the performance of other stereo algorithms is difficult since we have to consider both depth accuracy and speed. Furthermore, depth accuracy is measurable only on objects with known ground-truth which are difficult to compare with human figures.

There exist considerably faster systems all of them based on rougher depth estimates. The Stereo Vision Machine II from SRI [Konolige, 1997] and the Interval stereo processor [Woodfill and Herzen, 1997] use a DSP C60 and an FPGA array, respectively, achieving a video frame rate (30Hz) processing. However, their depth accuracy is not useful for close range systems because it is based on integer disparity estimation. Pentium-II based machines are the SRI-SVM-I [Konolige, 1997] and the Point-Grey Triclops trinocular systems which achieve 12 and 3 frames per second, respectively, are also calculating only integer-valued disparities.

7. CONCLUSIONS AND THE FUTURE

We have presented a first real-time implementation of 3D-tele-immersion based on view-point independent scene acquisition. The current stereo-reconstruction uses state of the art stereo matching. We also introduced a novel trinocular algorithm which will be part of the next system release. The fusion of the two 3D worlds is asynchronous which facilitates higher flexibility in the display site. The implementation enables the tuning of quality and working volume vs. speed. The user can choose an acceptable balance among size of working volume, depth quality, and spatial resolution.

As with many other prototypes in the history of technology it opens numerous challenges for all disciplines of graphics, vision, and communication. Tele-immersion is already recognized as one of the key-applications for Internet-2. The main challenge for the vision as well as the graphics community is the issue of representation. Like the explosion of coding techniques for transmission of 2D images after the introduction of WWW we anticipate breakthroughs in problems related to representation.

The wide use of 3D-data from reconstruction raises demand for a higher quality of shape representation. We are working on the critical problems of occluding contours and specularities arising in stereo reconstruction. The dynamics of the scene necessitate shape representations that will be easily updatable using some simple assumptions on temporal coherence. Even if we use multiple cameras to obtain a surround capture we need surface parametrizations that can be also spatially registered in a simple and robust way. Last but not least, the 3D-data have to be transmitted over the network. The challenge for progressive 3D wavelet-like representations which simultaneously address the critical issues above remains open.

References

[Ayache and Hansen, 1988] Ayache, N. and Hansen, C. (1988). Rectification of images for binocular and trinocular stereovision. *Proc. of 9th International Conference on Pattern Recognition*, 1:11–16.

[Belhumeur, 1996] Belhumeur, P. (1996). A bayesian approach to binocular stereopsis. *Intl. J. of Computer Vision*, 19(3):237–260.

[Chen and Williams, 1993] Chen, E. and Williams, L. (1993). View interpolation for image synthesis. In *ACM SIGGRAPH*, pages 279-288, 1993.

[Devernay, 1994] Devernay, F. (1994). Computing differential properties of 3-D shapes from stereoscopic images without 3-D models. *INRIA, Sophia Antipolis, RR-2304*.

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

[Faugeras, 1993] Faugeras, O. (1993). *Three-dimensional Computer Vision*. MIT-Press, Cambridge, MA.

[Konolige, 1997] Konolige, K. (1997). Small vision system: Hardware and implementation. *Eighth International Symposium on Robotics Research, Hayama, Japan*.

[Kutulakos and Vallino, 1998] Kutulakos, K. and Vallino, J. (1998). Calibration-free augmented reality. *IEEE Trans. on Visualization and Computer Graphics*, 4(1):1–20.

[Macedonia and Noll, 1998] Macedonia, M. and Noll, S. (1998). Real-time 60hz distortion correction on a silicon graphics ig. *IEEE Computer Graphics and Applications*, 5:76–82.

[Matthies, 1992] Matthies, L. (1992). Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. *International Journal of Computer Vision*, 8:71–91.

[Maybank and Faugeras, 1992] Maybank, S. and Faugeras, O. (1992). A theory of self-calibration of a moving camera. *Intl. J. of Computer Vision*, 8(2):123–151.

[Moravec, 1981] Moravec, H. (1980/1981). Robot rover visual navigation. *Computer Science: Artificial Intelligence*, pages 105–108.

[Narayanan et al., 1998] Narayanan, P., Rander, P., and Kanade, T. (1998). Constructing virtual worlds using dense stereo. *Proc, Intl. Conf. Computer Vision ICCV98*, pages 3–10.

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

[Scharstein and Szeliski, 1996] Scharstein, D. and Szeliski, R. (1996). Stereo matching with non-linear diffusion. *Proc. Int. Conf. Computer Vision and Pattern Recognition*.

[Seitz and Dyer, 1996] Seitz, S. and Dyer, C. (1996). Towards image-based scene representation using view morphing. In *ACM SIGGRAPH*.

[Tomasi and Manduchi, 1996] Tomasi, C. and Manduchi, R. (1996). Stereo without search. *Proc. European Conf. Computer Vision*.

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

[Woodfill and Herzen, 1997] Woodfill, J. and Herzen, B. V. (1997). real time stereo vision on the parts reconfigurable computer. In *IEEE Workshop on FPGAs for Custom Computing Machines*.

[Yuille and Poggio, 1984] Yuille, A. and Poggio, T. (1984). A generalized ordering constraint for stereo correspondence. AI Lab Memo 777, MIT.