3-D Data Acquisition and Interpretation for Virtual Reality and Telepresence

R. Šára
Center for Machine Perception
Czech Technical University
Prague, Czech Republic
sara@cmp.felk.cvut.cz

R. Bajcsy, G. Kamberova, R. A. McKendall
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA, U.S.A.
kamberov@grip.cis.upenn.edu

Published in: *Proc. IEEE Workshop Computer Vision for Virtual Reality Based Human Communications.* Bombay, India, January 1998.

This publication can be obtained via anonymous ftp from
ftp://cmp.felk.cvut.cz/pub/cvl/articles/sara/cvvrbhc98.ps.gz

Copyright 1998 IEEE. Published in the Proceedings of CVVRHC'98, January 1998 in Bombay, India. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works, must be obtained from the IEEE. Contact: Manager, Copyrights and Permissions / IEEE Service Center / 445 Hoes Lane / P.O. Box 1331 / Piscataway, NJ 08855-1331, USA. Telephone: +Intl. 908-562-3966.

3-D Data Acquisition and Interpretation for Virtual Reality and Telepresence

R. Šára
Center for Machine Perception
Czech Technical University
Prague, Czech Republic
sara@cmp.felk.cvut.cz

R. Bajcsy, G. Kamberova, R. A. McKendall
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA, U.S.A.
kamberov@grip.cis.upenn.edu

Abstract

We present our long-term effort focused on building a 3-D model acquisition system for teleimmersion applications. In our project we stress fast processing and high fidelity of the result: it is not just geometric accuracy of the recovered geometric model but also radiometric correctness of the recovered surface texture. We suggest that single-type geometric model is not sufficient for modeling large and complex scenes. In this paper, a hierarchy of representations suitable for 3-D structure models recoverable from visual data is presented and the related problems of data acquisition and fusion, outlier identification, and models reconstruction are discussed.

1 Introduction

The goal of teleimmersion is to provide people at remote places with realistic experiences that they are all sharing virtual and 'real' objects, including their partners present 'in the same room.' This is done by creating a virtual world for each participant, one that is a dynamically updated representation of the blend of the remote and local real worlds. Teleimmersion could be realized by integrating virtual reality, computer graphics, machine perception, distributed systems, and high-performance network applications.

Creating computer models of real dynamic world requires data acquisition and interpretation. In this context, by interpretation it is usually understood geometric and radiometric modeling. The modeling is necessary to provide 'views' to the remote observer that can be *predictively* interpolated or extrapolated from a small set of primary views. The prediction is necessary since the communication channel bandwidth will always limit the size of data to be transferred and the fast sensor control loop cannot be closed over the communication link. We believe that, *unlike the image-based representations, geometry modeling is essential* for such interpolation/extrapolation.

Although theoretically possible, from our experience it follows that recovering highly structured uniform representations (e.g. surface triangulations) from visual data is not technically feasible (even with many visual sensors and/or many processing units) because of the structural complexity of usual scenes with respect to the finite resolution of usual sensors.

After discussing a hierarchy of 3-D model representations in this section, we describe our efforts in utilizing the representations in achieving near real-time 3-D reconstruction of dynamic scenes for shared virtual reality and teleimmersion applications.

We use *active* or *passive polynocular stereo.* Passive stereo relies on natural surface texture, while an active stereo system includes an uncalibrated projector of a random pattern that induces texture on objects. The combination of an uncalibrated projector and calibrated cameras (we usually use 4 or 5 cameras per 'view,' see Figure 3) gives us a great deal of flexibility. Such a system is compact and mechanically quite robust, it can be moved around, and needs no projector re-calibration on the fly. After initial calibration, cameras can be easily self re-calibrated from the correspondences found during the normal function of the system [19]. Because of the uncalibrated projector, an active stereo system is *not* a structured-light range-finder.

We see the 3-D object (a surface, or, generally, a manifold) reconstruction process as follows:

Raw reconstructed 3-D points $[x, y, z]$ obtained from one or several pairs of intensity images are first validated and outliers and redundant points are removed. Redundant points are those that can be removed without decreasing information contents of the dataset by more than a threshold. The validation process uses either an empirical noise model (for a given surface texture), or a theoretical noise model obtained from propagating random noise model through the stereo algorithm. Currently we use variance [4] and interval propagation; we plan building an empirical noise model for our projector-induced random texture.

Hyperellipsoid-like local geometric primitives are recovered by grouping non-redundant points. Points that cannot be grouped are left uninterpreted. Grouping reduces the amount of data to work with and *locally* interprets the point set (the locality is important for many reasons, most of all it allows for introducing very little assumptions about the observed objects). We prefer building models in a bottom-up way because we do not have to introduce very restrictive a priori assumptions (like surface smoothness) during the early stages of the interpretation process. This simplifies data fusion, allows for greater flexibility of the interpretation (model recovery), and allows for 'bootstrapping' the accuracy in later stages (when attention can be focused on structures that need the accuracy increase). Postponing the 'focus on accuracy' has the advantage that the dataset is already (partially) interpreted and therefore reasoning about where and how to focus is easier. Simply put, we interpret the data in a bottom-up and solve-one-problem-at-a-time manner.

The local primitives capture the local surface (manifold) orientation defined as the normal of the hyperellipsoid (it exists iff the primitive is flat). We call the primitives *fish-scales:* they cover manifolds (e.g. surfaces or curves) the same way scales cover some fish or reptiles [12]. Our fish-scales are a generalization of *oriented particles* proposed by Tonnessen and Szeliski [13], that, in turn, are a generalization of the particles of Reeves [9].

Primitives parameters are refined based on *polynocular local image dissimilarity* that, akin to *shape from texture* [16], allows for estimating the primitive orientation more precisely [20, 3]. The main difference from shape from texture is that the texture isotropy assumption is not necessary, since we can re-project textures from multiple images on the same primitive (its main hyper-plane) and compute the dissimilarity on the re-projected images. This is an example of how only a partial local interpretation can help significantly in the accuracy bootstrapping process.

The collection of local primitives can then be further interpreted as a set of free-form surfaces with boundaries (by connecting hyper-ellipsoids that are 'flat,' close, and compatible at their overlap) and/or as a set of free-form curves (by connecting hyper-ellipsoids that are elongated, close, and compatible). Some primitives may be left uninterpreted.

The full hierarchy of partial interpretations should exist at each instant of scene modeling. When new measurements (3-D points) are added to the set, those that form coherent structures are grouped to local primitives, which are in turn linked to global structures. We believe this is

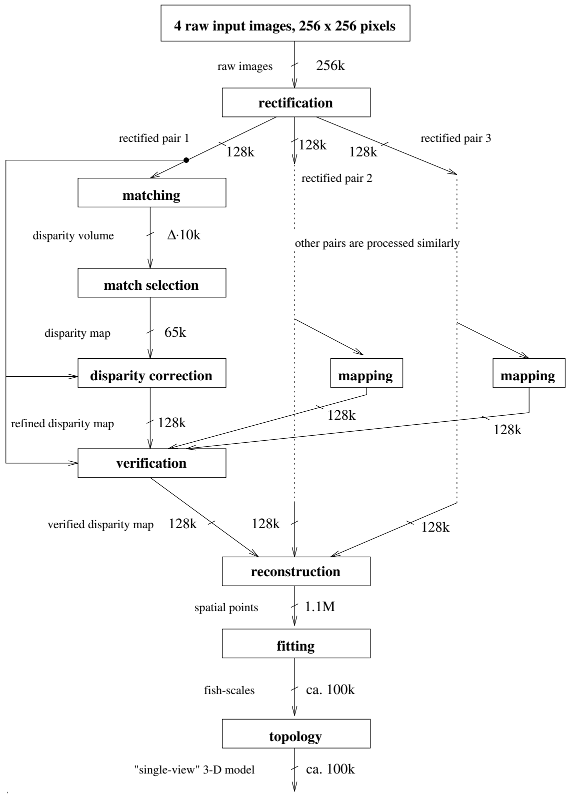


Fig. 1: The first generation of our polynocular stereo system.

a very flexible representation suitable for a large class of objects that may be encountered in natural and virtual environments. By flexible we mean that it can be *cheaply* updated from an initial guess according to the need for fidelity. A high-level interpretation is necessary in just a small (or the most important) fraction of the scene. One need not be afraid of having uninterpreted isolated spatial points in the model, we will show later that if the points are augmented with color they provide a sufficient model for quite realistic renderings.

2 Our Current and Future Research

We have build a polynocular stereo system [15], see Figure 1 in which a parallelized pipeline of simple algorithms is shown. The algorithms solve for the correspondences first and give integer-resolution disparity map after rejecting ambiguous and unreliable matches based on the weak forbidden zone constraint [18]. Once the integer disparity map is found, sub-pixel disparity correction d_ϵ is computed assuming that $|d_\epsilon| < 1$. This is done by least-squares minimization assuming affine distortion model constrained by epipolar geometry [2]. It requires no search. Matching and disparity map correction are done *pairwise* and in parallel. The reconstructed points are then *verified* in all input images for the similarity of their neighborhoods. This step

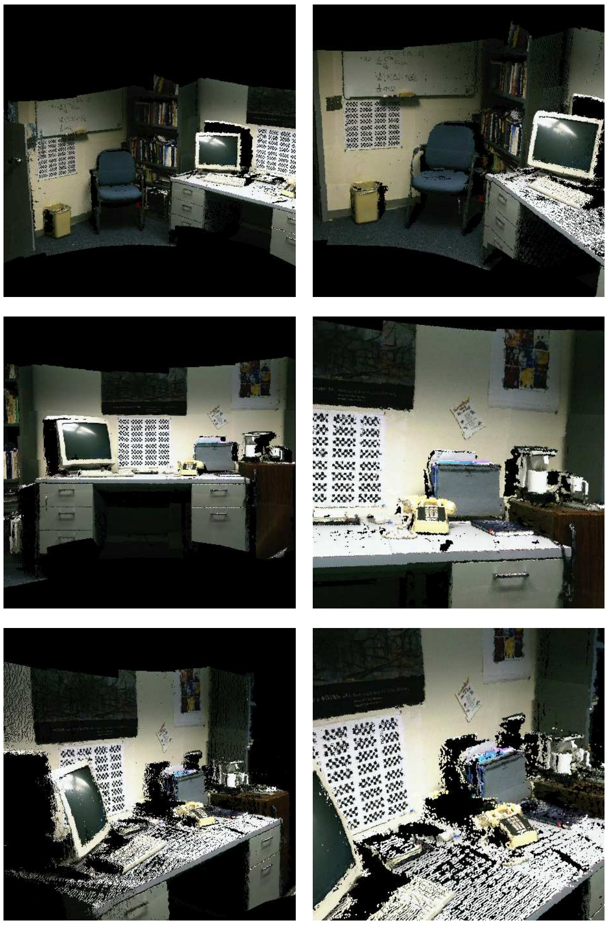


Fig. 2: A $150^\circ \times 90^\circ$ view reconstruction of an 3m×2m office. This is an unverified redundant set of colored isolated points in Euclidean space without any reconstructed connectivity. The reconstruction was built from 40 polynocular views. The camera motions were estimated from landmarks located in the scene (one of them is on the wall above the desk). Registration errors in the scene were about 1cm.

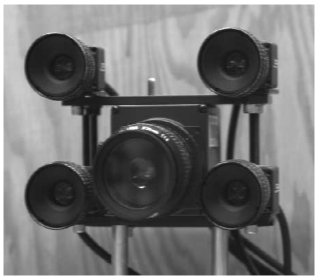


Fig. 3: Five-camera stereo rig used in our experiments (four b/w Sony XC-77RR cameras with 25mm or 16mm Cosmicar lenses; one 3-chip color Sony XC-007 camera with 25mm Canon lens). The color camera in the center is used for texture projection. The horizontal distance between the b/w cameras is 12.5cm and the vertical distance is 8cm; the horizontal pairs are slightly verging.

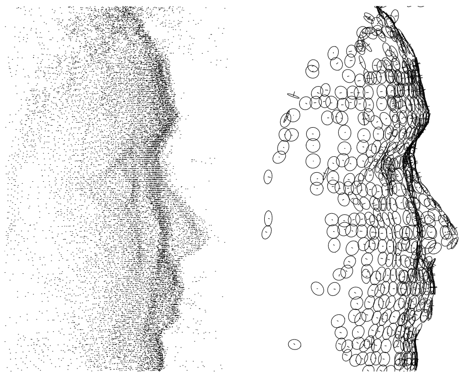


Fig. 4: Local geometric primitives (fish-scales, right) approximate the point set (left) locally. The primitives are given by their positions and orientations (short needles) and are in fact fuzzy sets of infinite extent and ellipsoidal structure.

fuses the disparity maps. An example of an uninterpreted set of colored reconstructed points in Euclidean space is shown in Figure 2, visit [11] for details.

Doing stereo matching pairwise has several important advantages: (1) surface patches visible in just one of the pairs can still be reconstructed, (2) specularities do not influence the final reconstruction too much, since the likelihood of their simultaneous visibility in all pairs is very small [1], (3) points that are difficult to match because of a large skew component in the geometric mapping between the images of a horizontal camera pair (like points on floors and ceilings) are easy to match when viewed by a vertical pair. Having four independent pairs—two horizontal and two vertical as in Figure 3—thus guarantees dense reconstructions in many more cases than in one quadrinocular stereo matching employing all four images at once.

Local disk-like geometric primitives introduced above are then fit to the cloud of points resulting from 3-D point reconstruction, see Figure 4. The primitives are recovered under the assumption that there is no hole in the surface smaller than the fish-scale diameter. The diameter is a parameter to the reconstruction procedure. The primitives are then filtered based on the goodness of fit and on their *rank* that measures their ellipticity and flatness. See [12] for details.

3-D models of manifolds are reconstructed based on proximity and mutual compatibility of the local primitives. We have defined fuzzy binary relations of relative intersection and relative inclusion of the primitives [12]. Relative intersection measures their compatibility. The most-consistent manifold is extracted from the collection of sim-

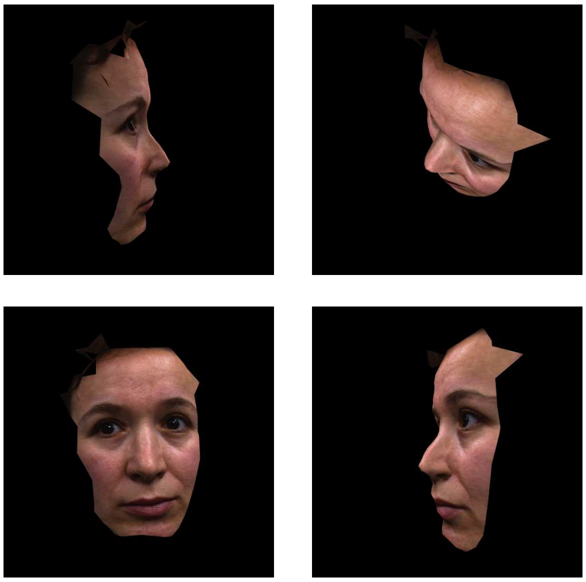
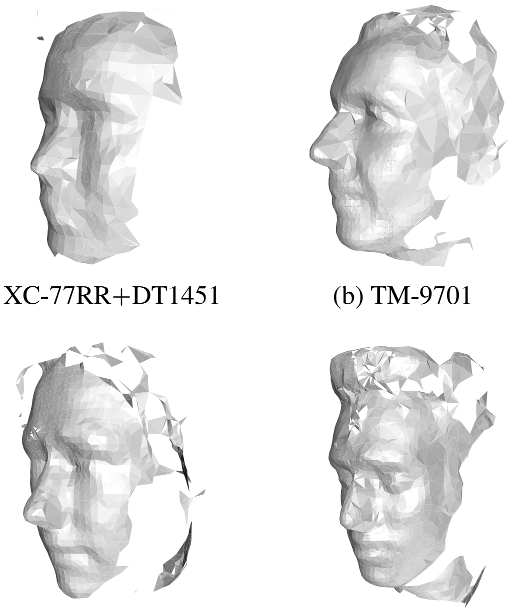


Fig. 5: Textured 3-D surface model recovered from the dataset in Fig. 4 shown textureless in Fig. 6a using an improved reconstruction algorithm.



(a) XC-77RR+DT1451 (b) TM-9701

(c) XC-77RR+DT1451 (d) TM-9701

Fig. 6: Cameras with less image noise and less digitization artifacts or strong induced texture can provide much better data for polynocular stereo. A surface reconstructed from four-camera stereo with off-the-shelf TV cameras and a conventional framegrabber (a). A different subject's face reconstructed from a similar setup of four digital cameras is clearly more complete and accurate (b). Relying on natural texture may be tricky as in (d) reconstructed under the same conditions as (b), see holes on the forehead caused by the presence of highlights on the surface. If projected random pattern is used to induce texture, a good geometric accuracy can be achieved even with the TV cameras (c). Both subjects (c) and (d) have very smooth skin. To demonstrate the geometric accuracy, models are neither textured nor smoothed. None of the point sets used for the reconstructions shown here was filtered by the rejection procedure discussed in the text.

plices defined over the centers of the primitives. Manifold consistency is computed as the sum of all pairwise compatibilities associated with the edges in the manifold. The manifold order (curve, surface) can be selected beforehand. Some results for real human faces are shown in Figures 5 and 6. Visit [15] for details and for VRML models.

Relative inclusion is a fuzzy ordinal relation. It allows for building irregular multi-resolution representations and for removing redundant primitives. This is a topic for our current research.

2.1 Understanding the Systematic and Random Errors

We have shown that radiometric correction of the images, which accounts for differences in the individual sensor characteristics, reduces significantly the systematic errors in case of scenes of weak textures [6]. For scenes of strong texture, the sources of systematic errors are the attributes of the scene: occlusions, highlights, geometry, repetitive texture, shadows, etc.

The sensor noise is the source of *random errors* in the results. We used variance and interval propagation techniques—together with Gaussian model assumption for the subpixel disparity noise—to *quantify the precision* of the results with confidence intervals in each of the coordinates of the reconstructed points [7]. The sizes of the intervals for the z-coordinate are varying per pixel, depending on the scene (typically from 0.1mm to 14mm for our test scenes). We have demonstrated the potential use of the confidence intervals for rejection of unreliable points. For the example shown in Figure 7 we used one stereo pair, intervals of 0.68 confidence, and used the mean interval size in the z-component as a threshold for the rejection unreliable points.

Currently we are exploring a rejection procedure based on the observation that integer disparity values with higher uncertainty are more likely to occur at image positions corresponding to scene areas deviating significantly from the continuity and planarity modeling assumptions of matching. In particular, when multiple stereo pairs are used, such a rejection procedure removes a lot of far outlier *and* redundant points, thus reducing the reconstructed set size significantly and still preserving the resolution. For the test images used for the reconstruction in Figure 7, the procedure removed almost 60% of the points. The reduction in the reconstructed set size also leads to a decrease in the total running time of all stages (up to and including verification) by 30% compared with the reconstruction without the use of the rejection procedure. If the full 3-D manifold representation is reconstructed, the comparison in running time is even more dramatic: almost 50% decrease in overall running time.

For a quantitative evaluation, the accuracy of the reconstruction was evaluated on a planar test scene by fitting a plane to the data (without and with the use of the rejection procedure). For the reconstruction with the rejection procedure, the residuals were by an order of magnitude smaller than the residuals of the fit to reconstructed points without the rejection

2.2 Redesigning the Match Selection

The size of the confidence intervals is related to the scene attributes. The intervals are large when: (1) the signal-to-noise ratio of the image is locally low, as in the textureless areas, in the areas overlaid by highlights where the texture contrast is low, and in dark or shadowed areas; (2) when the pixel values in the left and right image windows used for assessing the match do not correspond well with each other (as in the case of highlights, in the hair and eyebrows which does not image as regular surface texture, and near occlusions where the matches are notoriously bad in all area-based stereo matching algorithms [10]).

The experiments based on the reconstruction with rejection indicate that the stereo matching can be improved. Incorporating the rejection directly into match selection procedure is one of our intentions.

2.3 Intrinsic Surface Texture Recovery

Of course, a realistic model has to be augmented with high-fidelity *texture* that corresponds to surface albedo. It is generally possible to recover the albedo by methods like *photometric stereo* [17, 14, 5]. One of our research goals is to avoid the necessity to know the incident light distribution and/or the surface reflectance model in order to recover the albedo. We expect that partial control over light will be necessary but no light position or intensity calibration will be required.

2.4 Real-Time Parallel Implementation

Real-time implementation is our final goal. One of our motivations for a bottom-up bootstrap-like process we have just described is its easy structured parallelizability. We have demonstrated that parallel implementation of correlation-based 3-D reconstruction and fusion from a series of 2-D images is feasible using the Client-Manager-Worker model on a scalable cluster of heterogeneous com-

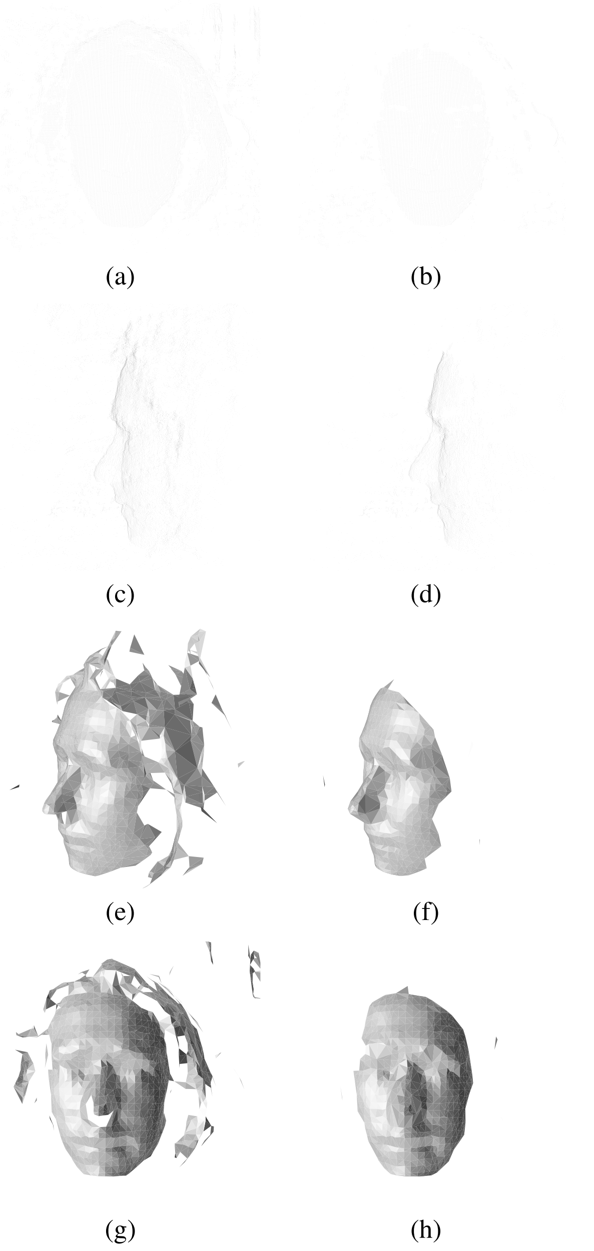


Fig. 7: Unfiltered point set from single-pair stereo (a,c) and the corresponding reconstruction (e,g). The same set filtered using a threshold on confidence intervals size (b,d) and the reconstruction (f,h). The filtered set is easier to interpret as a consistent surface: note the hole in (e,g) under the nose where few points were measured; here even far points may influence the reconstruction and the sensitivity to outliers increases. Although images are from the same experiment as in Fig. 6c, the geometric resolution is by 30% worse, since here only a single stereo pair was used as opposed to Fig. 6c. The visible Moiré pattern in (a,b) is due to regularity of the points distribution in space.

puters under the message-passing paradigm of parallel computing [8].

3 Conclusions

We have presented a part of collaborative work between the GRASP Lab and the Telepresence Group of Henry Fuchs at UNC at Chapel Hill.

In our current efforts, G. Kamberova is working on redundancy reduction in the point data-sets and their fusion without interpretation. R. Šára started a new research on recovering albedo (intrinsic texture) maps from images using physics-based methods. We also expect further progress in geometric accuracy and free-form surface model recovery.

In the project, the new effort in GRASP Lab will be towards 3-D reconstruction of dynamic scenes. In order to meet the high real-time demands and network bandwidth limitations, we are working on *incremental and local update* of the reconstructed set exploring the spatio-temporal changes in the image sequence, and on new *efficient* surface *invariants and methods for encoding them.*

Acknowledgements

This work has been supported by the following grants: Army DAAH04-96-1-0007, DARPA N00014-92-J-1647, NSF SBR89-20230. First author has been partly supported by the Czech Grant Agency under the grants 102/95/1378, 102/97/0480, 102/97/0855, and EU grant Copernicus CP941068.

References

[1] D. N. Bhat and S. K. Nayar. Stereo in the presence of specular reflection. Research Report CUCS–030–94, Department of Computer Science, Columbia University, New York, Dec. 1994.

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

[3] P. Fua. Reconstructing complex surfaces from multiple stereo views. Tech Note 550, AI Center, SRI International, Menlo Park, CA, Nov. 1994.

[4] R. Haralick. Propagating covariance in computer vision. In *Workshop on Performance Characterization of Vision Algorithms*, Cambridge, 1996.

[5] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *Journal of Optical Society of America A*, 11(11):3079–3089, Nov. 1994.

[6] G. Kamberova and R. Bajcsy. The effect of radiometric correction on multicamera algorithms. Technical Report MS-CIS-97-17, GRASP Lab, University of Pennsylvania, Philadelphia, PA, USA, 1997.

[7] G. Kamberova and R. Bajcsy. Precision of 3-D points reconstructed from stereo. GRASP Lab, University of Pennsylvania, Philadelphia, PA, USA, November 1997. Submitted for publication.

[8] R. McKendall, R. Šára, and R. Bajcsy. Scalable parallel computing for real-time 3-D reconstruction from polynocular stereo. `http://www.cis.upenn.edu/~mcken/progressII/main.html`, February 1997.

[9] W. T. Reeves. Particle systems—a technique for modeling a class of fuzzy objects. *Computer Graphics*, 17(3):359–376, July 1983. SIGGRAPH 1983.

[10] R. Šára and R. Bajcsy. On occluding contour artifacts in stereo vision. In *Proc. IEEE Computer Society Conf. CVPR*, pages 852–857, Puerto Rico, June 1997.

[11] R. Šára and R. Bajcsy. PennOffice: A room reconstruction from quadrinocular stereo vision. `http://cmp.felk.cvut.cz/~sara/PennOffice/home.html`, November 1997.

[12] R. Šára and R. Bajcsy. Fish-scales: Representing fuzzy manifolds. In *Proc. Int. Conference on Computer Vision*, Bombay, India, January 1998.

[13] R. Szeliski and D. Tonnesen. Surface modeling with oriented particle systems. *Computer Graphics (SIGGRAPH '92)*, 26(2):185–194, July 1992.

[14] H. D. Tagare and d. R. J. P. A theory of photometric stereo for a class of diffuse non-Lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2):133–152, February 1981.

[15] The authors. Accurate 3-D reconstruction from polynocular stereo. `http://cmp.felk.cvut.cz/~sara/Stereo/home.html`, 1997.

[16] A. P. Witkin. Recovering surface shape and orientation from texture. *Artificial Intell.*, 17:17–45, 1981.

[17] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980.

[18] A. L. Yuille and T. Poggio. A generalized ordering constraint for stereo correspondence. Artificial Intelligence Laboratory Memo 777, MIT, 1984.

[19] Z. Zhang and V. Schenk. Self-maintaining camera calibration over time. In *Proc. IEEE Computer Society Conf. CVPR*, pages 231–236, Puerto Rico, June 1997.

[20] V. Zýka and R. Šára. Polynocular local image dissimilarity for 3-D reconstruction. Unpublished paper, Center for Machine Perception, Czech Technical University Prague, December 1997.