Perceptual Grouping for Multiple View Stereo using Tensor Voting

Philippos Mordohai and Gérard Medioni
Integrated Media Systems Center
University of Southern California
{mordohai, medioni}@iris.usc.edu

Abstract*

We address the problem of multiple view stereo from a perceptual organization perspective. Currently, the leading methods in the field are volumetric. They operate at the level of scene voxels and image pixels, without considering the structures depicted in them. On the other hand, many perceptual organization methods for binocular stereo are not extensible to more images. We present an approach where feature matching and structure reconstruction are addressed within the same framework. In order to handle noise, lack of image features, and discontinuities, we adopt a tensor representation for the data and tensor voting for information propagation. The key contributions of this paper are twofold. First, we introduce "saliency" instead of correlation as the criterion to determine the correctness of matches; second, our tensor representation and voting enable us to perform the complex computations associated with multiple view stereo at a reasonable computational cost. We present results on real data.

1 Introduction

The derivation of scene descriptions from images encompasses two processes: the establishment of feature correspondences and the reconstruction of surfaces based on the depth measurements obtained by the previous process. The completion of these tasks is encumbered with inherent difficulties, such as the lack of distinct image features, measurement and quantization noise, surface discontinuities and half occlusions. Four decades ago, Julesz brought the issue of binocular stereo vision into attention, introducing random dot stereograms, and demonstrating that depth perception can occur even in the absence of monocular information [5][6]. The work of Marr and Poggio [10] and Marr [11] made a first attempt to define the problem and its fundamental constraints.

After briefly reviewing previous work on binocular and multiple view stereo in the next section, we describe the design of our algorithm in Section 3. Section 4 is an overview of tensor voting and Section 5 illustrates its application to multiple view stereo. In Section 6, we show some results, while Section 7 concludes with a discussion on future work.

2 Related Work

Due to lack of space, we refer readers to [14] for a comprehensive review of previous work on stereo. Among the numerous publications, of major significance to us are the methods presented in [2][4]. The former on account of processing in three dimensions, instead of two or one as in most binocular stereo methods, and the latter because of the integration of the feature matching and structure extraction phases.

In the last few years, volumetric approaches have become the leading methods in the field of multiple view stereo and have achieved very good results. Surveys of these methods were published in [3][13]. We propose to address the problem from a perceptual organization point of view, instead.

Volumetric methods adopt a ray-space representation of the scene and pose multiple view stereo as the computation of the visual hull of the scene. Central to many of these approaches is the notion of photo-consistency. The criterion for the existence of a voxel in the final scene model is the similarity of its appearance in all the cameras from which it is visible.

Even though volumetric methods are currently the state of the art, they suffer from some inherent limitations. First and foremost among them is the inability to handle concavities in the scene. Good approximations can be obtained with a large set of input views, but the treatment of concavities within a visual hull framework is flawed. Even in the absence of acute concavities, accurate reconstruction of the scene requires a large number of

* This research was funded in part by the U.S. National Science Foundation under grant No. IRI-9811883. and in part by Geometrix, Inc.

views, while the general placement of cameras further increases the computational cost [13]. Volumetric methods are by nature iterative, require careful initialization of the scene model, and cannot recover from erroneous initializations. Finally, they are susceptible to pixel quantization noise and require fine voxelization of the space to compute the intersections between rays and voxels with the necessary accuracy [1].

Perceptual methods are not confounded by these limitations. Since the problem is posed as the reconstruction of the positions of the selected features in 3-D space, concavities do not present additional difficulties, as long as the points are visible in at least two views. A perceptual grouping framework enables the propagation of information among tokens and imposes constraints, such as the "matter is cohesive" principle [11], therefore, it can utilize information from the images more efficiently and achieve reconstructions of the same accuracy with less views.

The limitations of current approaches based on perceptual organization are that either their representation is inadequate for generalization to multiple views, or that computational complexity increases exponentially with the number of views. As a result they have failed to match the quality of the reconstructions produced by the volumetric methods.

3 Implementation Considerations

We propose to pose the problem of multiple view stereo as a perceptual organization problem using tensor voting. At these initial stages of our research, we want to ensure that our framework does not suffer from any fundamental weaknesses. Tensor representation and voting enable us to perform all processing in three dimensions after combining the initial information from all views. This makes our approach applicable to scenes that contain multiple layers, transparency, or full 360° views, which would have been impossible had we adopted a representation such as the 2 ½-D sketch. The computational cost is manageable, since voting is performed locally and complexity is linear with respect to the number of views.

We tried to develop an approach that does not suffer from design deficiencies such as the requirement that features appear on all images, special topology of the camera positions, or the treatment of some views as privileged. Our approach requires that features be visible in at least two of the images, and treats all images equally. At the current stage, camera calibration information has to be provided, but the literature provides methods for self-calibration such as [17].

The novelty of our method comes from the fact that we use surface saliency instead of cross-correlation to determine the correctness of matches. Instead of making the decisions at the matching stage based on cross-correlation, we delay them until saliency information is available. High cross-correlation values between image intensities are indications of potential matches, but not very reliable as a criterion for resolving the correctness of matches. A consequence of the use of saliency is the integration of feature matching with surface extraction.

4 Brief Overview of Tensor Voting

The use of a voting process for feature inference from sparse and noisy data was formalized into a unified tensor voting framework [12]. This methodology is non-iterative and robust to considerable amounts of outlier noise. The only free parameter is the scale of voting, which essentially defines the size of the neighborhood of each point. The input data is encoded as second-order symmetric tensors, and constraints, such as proximity, co-linearity and co-curvilinearity are propagated by voting within the neighborhood. We propose to incorporate first-order information, namely a polarity vector, in the representation, in order to detect depth discontinuities.

The representation with first-order and second-order symmetric tensors can convey the multiple roles of a point in space simultaneously. We are able to represent points on smooth surfaces, surface intersections and boundaries, curves, curve junctions and endpoints, as well as outliers with the same representation. A second-order symmetric tensor in 3-D has the form of an ellipsoid, or equivalently of a 3x3 matrix. Given its eigensystem, the tensor can be decomposed into a "stick", a "plate" and a "ball" component as follows (see also Fig. 1):

$$S=\lambda_1\cdot e_1e_1^T+\lambda_2\cdot e_2e_2^T+\lambda_3\cdot e_3e_3^T=$$

$$=(\lambda_1-\lambda_2)e_1e_1^T+(\lambda_2-\lambda_3)(e_1e_1^T+e_2e_2^T)+\lambda_3(e_1e_1^T+e_2e_2^T+e_3e_3^T)$$

Surface saliency is encoded by the difference between the two largest eigenvalues, with a surface normal given by e_1. The difference between the second and third eigenvalue corresponds to curve saliency, with a curve normal on the plane defined by e_1 and e_2, or, equivalently, a tangent parallel to e_3. Finally, junction saliency is the encoded by smallest eigenvalue and has no preference of orientation.

During the voting process, the second-order tensor at

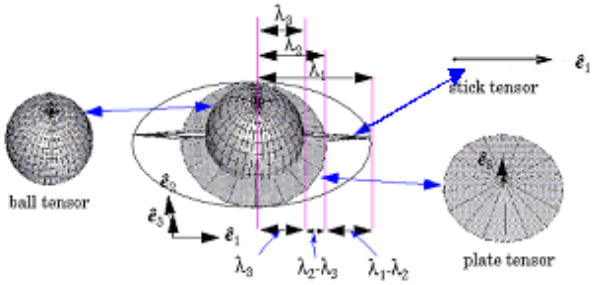


Figure 1. Tensor decomposition

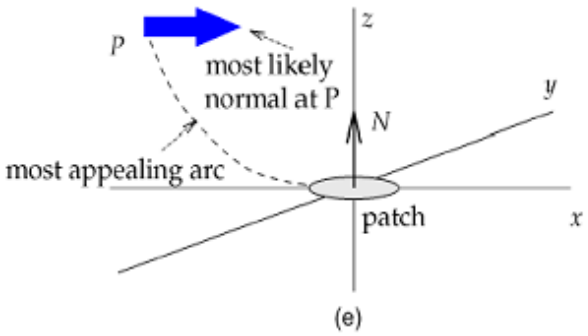


Figure 2. Vote generation

each input site casts votes to its neighboring input sites (sparse vote) or to all locations within its neighborhood (dense vote). The votes are also second-order symmetric tensors and their orientation corresponds to the normal orientation the receiver would have, if the voter and receiver were in the same perceptual structure. Their saliency (size) decays with respect to distance and curvature according to the following equation:

$$V(d,\rho) = e^{-\frac{d^2+c\rho^2}{\sigma^2}}$$

Where d is the distance between the caster and receiver, ρ is the curvature of the smooth path between the two locations (see Fig. 2), σ is the scale, and c is a constant.

The accumulation of these tensor votes is performed by tensor addition, which is equivalent to the addition of 3x3 matrices. After voting is complete, the eigensystem at each site is analyzed and the likelihood of a point belonging on a smooth structure or not is determined.

The mechanism for first-order voting is exactly the same as in the case of second-order tensor votes, with the difference being that the votes cast are vectors instead of tensors. The first-order votes have the same magnitude, as defined by the saliency decay function, but they are cast along the tangent of the smoothest path instead of the normal. Therefore, besides the saliency tensor, every location also has a polarity vector associated with it.

First-order vote accumulation is performed by vector addition. Potential surface boundaries and curve endpoints can de detected based on the large vector sum they have accumulated since most of their neighbors are on one side. The exterior among these points, are the correct boundaries and can be marked as such.

Both second and first order votes are pre-computed and stored in three-dimensional voting fields to reduce the computational cost. In large datasets the probability of repeating the same direction and saliency computations is high. Therefore, we compute the first and second order votes generated at all grid locations in the voting neighborhood by the elementary stick, plate and ball tensors and store them in look-up tables. This results in an increase in storage requirements, which is clearly outweighed by the decrease in computation time.

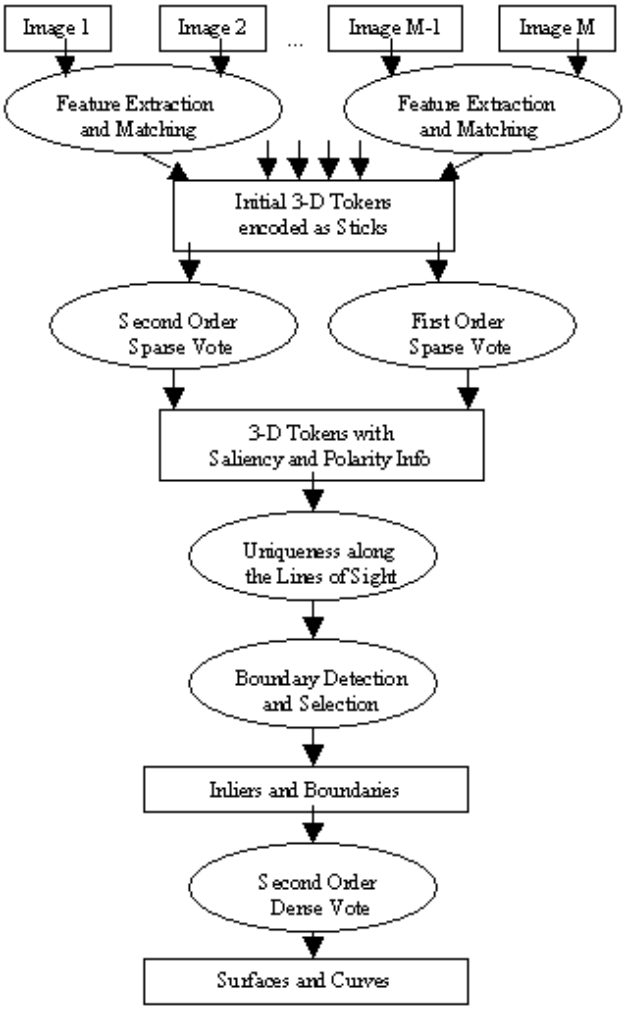


Figure 3. The flowchart of our approach

5 Tensor Voting for Multiple View Stereo

In this section, we describe our algorithm for multiple view stereo. Starting from a set of images, we extract features, identify potential correspondences, perform first and second order sparse tensor voting to detect inliers, boundaries and outliers, and, finally, perform dense tensor voting to extract surfaces and curves. Even though our algorithm is far from being finalized, our preliminary results are encouraging and we can only expect improvement as the processing modules, especially structure extraction from the output of the sparse vote, get more advanced. A flowchart of the algorithm can be seen in Fig. 3.

The first stage of processing entails the detection and matching of features in the images. More specifically, we begin by extracting points of high intensity variance, from

each image. Our method can work with both sparse and dense sets of features [8]. In this case, we use almost all pixels as inputs to the next stage by keeping the minimum requirement in intensity variance low.

Feature matching is performed using cross-correlation in rectangular windows between consecutive images. Applying the same scheme between non-consecutive images only introduces redundancy without producing additional matches. To overcome the well-documented failures of correlation-based matching, we identify all potential matches for every feature as local peaks of correlation, and use them as input to the tensor voting process.

The potential matches from the previous stage are reconstructed in 3-D space, using the known calibration information, and serve as the inputs for the sparse tensor vote. All data are integrated into one dataset that contains all the matches generated from every pair of images. We believe that processing all the available information at the same time can produce superior results compared to methods that are based on the fusion of surfaces generated by multiple executions of a binocular stereo technique. The latter approach suffers from inaccuracies that inevitably occur at the borders of each partial set of surfaces.

We initialize the tensor at each point as a stick pointing towards the midpoint of the optical centers of the cameras that generated the match, thus encoding an orientational preference for a surface visible from the appropriate angle. If more potential matches fall within the same voxel, indicating that the match is confirmed by more than two views, the tensors are added resulting in increased initial saliency. First-order tensors are initialized to zero, since no polarity information is available.

Sparse tensor voting is then performed and first and second order votes are collected at each location as described in the previous section. When saliency information is available, we enforce the uniqueness constraint along the lines of sight and select the most salient match for each feature. Points with high surface saliency are most likely to be inliers of smooth surfaces, points with high curve saliency lie on curves or surface orientation discontinuities, while points with high junction saliency are curve junctions.

At the same time, first-order polarity information allows us to detect points on depth discontinuities and curve endpoints. Points with high surface saliency and high polarity in a direction orthogonal to the surface normal are at or very close to the boundaries of the surface. Points with high curve saliency and high polarity in a direction parallel to the curve tangent are near the endpoints of the curve. Outliers tend to have high polarity values but no definite preference in orientation, since they receive contradicting second-order votes from points that lie mostly on one side.

Points that have been identified as potential surface boundaries and curve endpoints at the previous stage undergo a selection process during which the actual boundaries are identified. This final selection is performed using a simple rule that selects the outermost point among all neighboring potential boundary points whose polarity vectors are almost parallel.

Finally, combining dense tensor voting with a marching process [9] implemented as in [15], we extract surfaces and curves as the local maxima of surface and curve saliency respectively, while junctions can be extracted as the local maxima of junction saliency without any propagation. Since calculating votes for every location in the volume containing the data points is pointless and impractical, surface and curve extraction begins from seeds, locations with high saliency, and voting is performed only towards the directions indicated by the surface normals and curve tangents. Surfaces are extracted with sub-voxel accuracy, as the zero-crossings of the first derivative of surface saliency and curves are extracted in an analogous way.

6 Results

We have demonstrated the validity of our approach on binocular stereo in a variety of examples that include random dot stereograms, synthetic and real scenes ranging from aerial images to images of faces and classical stereo pairs. The input images and output surfaces and curves can be seen in [7][8].

For the multiple view case, we have obtained some preliminary encouraging results. Despite the evident aesthetic imperfections, they have demonstrated the capability to infer structure from extremely noisy data sets, as well as very accurate estimation of surface

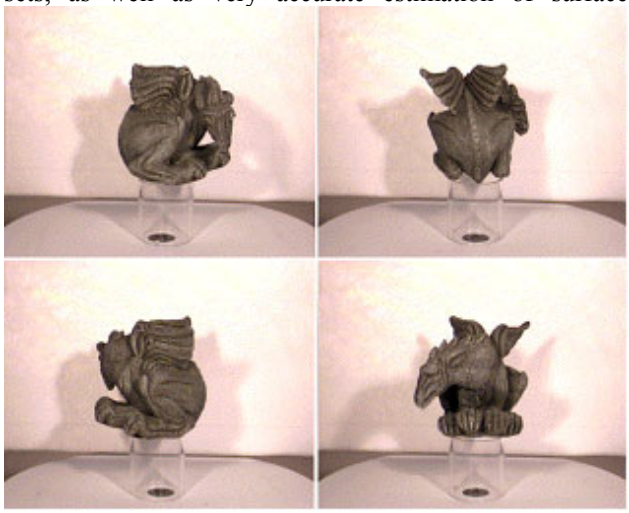
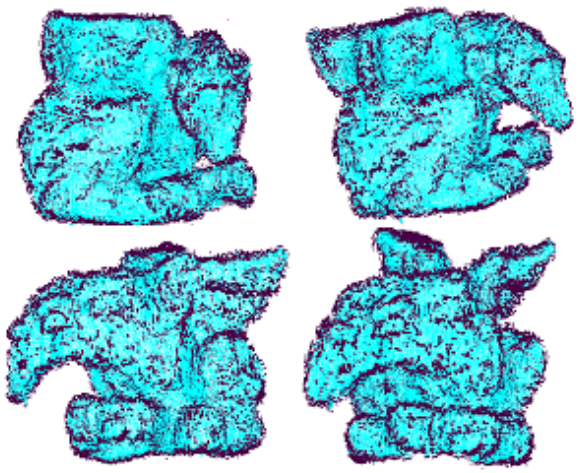
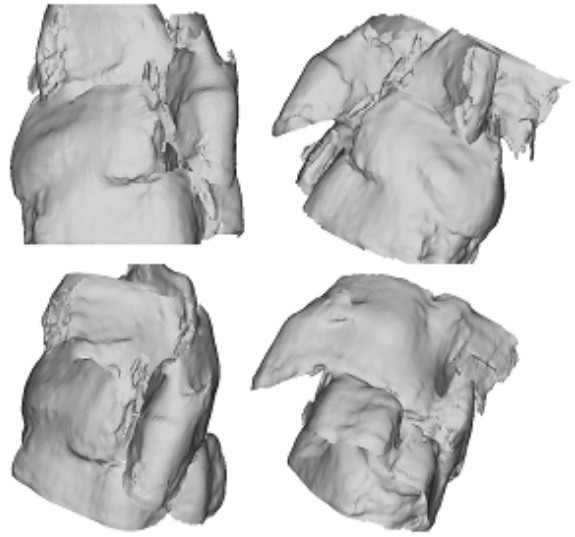


Figure 4. Four images of the Dragon sequence



(a) Output of the sparse vote with boundaries marked in black



(b) Output surfaces after the dense tensor vote

Figure 5. Output of the Dragon Sequence

normals and detection of depth and orientation discontinuities.

For the "dragon" sequene the input consists of 36 images from a turn-table sequence of a complicated object with concavities and significant self-occlusions, four of which can be seen in Fig. 4. The output of the sparse tensor vote can be seen in Fig. 5(a). Points colored light blue are surface inliers while points marked in black are the detected boundaries. The extracted surfaces can be seen in Fig. 5(b).

The next example is again a turntable sequence, the "lighthouse". It differs from the previous example in that many of the surfaces are planar, whereas the "dragon" consists solely of non-planar surfaces, and in that there are significant variations in texture form region to region. The difference in the quality and quantity of features

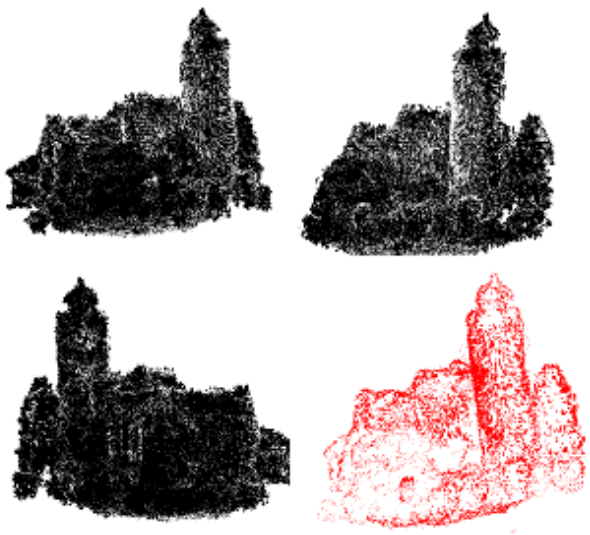
Figure 6. Four views of the Lighthouse sequence

between regions in the images has considerable effects in the matching stages. The buildings have distinct features that can be easily matched, while obtaining good correspondences for points on the rock is not an easy task. Four frames of the input sequence can be seen in Fig. 6. The input data set containing all candidate matches can be seen in Fig. 7, while Fig. 8(a) contains some views of the results of the sparse tensor vote. Note that the bottom right sub-figure contains only the detected discontinuities. Finally, Fig. 8(b) contains the results of the dense tensor vote.

7 Future Work

Our future work will focus on the development of an approach that addresses the multiple view stereo problem in a general way from a perceptual organization standpoint. Our approach can handle a large class of scenes, as demonstrated in the case of binocular stereo, and since the way it is set up does not conflict with the additional difficulties associated with the multiple view configuration, we can only expect superior results when depth and orientation discontinuity detection and curvature inference, according to [16], are better integrated and more thoroughly studied in this context. A more advanced feature matching technique would enable

Figure 7. Noisy input to the sparse vote



(a) Views of the output of the sparse vote

(b) Views of the output of the dense vote

Figure 8. Output of the Lighthouse Sequence

us to begin with less noisy data sets and allow us to capture details more accurately.

Our initial experiments have indicated that tensor voting in multiple scales is necessary for this problem. A weakness of our current implementation is that the selection of a single scale for the entire data set warrants a trade off between the level of detail in regions of high data density, and sufficient information propagation in sparser regions. Data of varying density appear in scenes that contain regions rich in texture, as well as textureless surfaces. Lower density in the data can be observed, for instance, in occluded parts of the scene that are visible from a small number of cameras. Voting with a large scale to enforce continuity and closure in these regions would unnecessarily smooth out the details in other parts. Therefore, we propose to apply a small scale initially, in order to capture details, where this is possible, and, then, progressively larger scales to achieve sufficient information propagation in sparser areas.

References

[1] A. Broadhurst, T. Drummond, and R. Cipolla, "A probabilistic Framework for Space Carving", *Proc ICCV,* pp. 388-393, 2001.

[2] Q. Chen and G. Medioni, "A Volumetric Stereo Matching Method: Application to Image-Based Modeling", *Proc. CVPR*, pp. 29-34, 1999.

[3] C.R. Dyer, "Volumetric Scene Reconstruction from Multiple Views", in *Foundations of Image Analysis* (L. S. Davis, ed.), *Kluwer*, 2000.

[4] W. Hoff and N. Ahuja, "Surfaces from Stereo: Integrating Feature Matching, Disparity Estimation, and Contour Detection", *IEEE Trans. on PAMI,* vol. 11, no. 2, pp. 121-136, 1989.

[5] B. Julesz, "Binocular depth perception of computer-generated patterns", *Bell System Technical Journal,* vol. 39, p. 1125-62, 1960.

[6] B. Julesz, "Dialogues on Perception", *MIT Press,* 1995.

[7] M.S. Lee and G. Medioni, "Inferring Segmented Surface Description from Stereo Data", *Proc. CVPR*, pp. 346-352, 1998.

[8] M.S. Lee, G. Medioni, and P. Mordohai, "Inference of Segmented Overlapping Surfaces from Binocular Stereo", to appear in *IEEE Trans. on PAMI,* 2001.

[9] W.E. Lorensen and H.E. Cline, "Marching Cubes: A High Resolution 3-D Surface Reconstruction Algorithm", *Computer Graphics,* vol. 21, no. 4, pp. 163-169, 1987.

[10] D. Marr and T. Poggio, "A Theory of Human Stereo Vision", *Proc. Roy. Soc. London,* vol. B204, pp. 301-328, 1979.

[11] D. Marr, "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information", *W.H. Freeman and Co.,* 1982.

[12] G. Medioni, M.S. Lee, and C.K Tang, "A Computational Framework for Feature Extraction and Segmentation", *Elsevier,* 2000.

[13] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer, "A Survey of Methods for Volumetric Scene Reconstruction from Photographs", Proc. *International Workshop on Volume Graphics,* 2001.

[14] R. Szeliski, "Stereo Algorithms and Representations for Image-Based Rendering", *in Proc. British Machine Vision Conference,* pp. 314-328, 1999.

[15] C.K. Tang and G. Medioni, "Inference of Integrated Surface, Curve, and Junction Descriptions from Sparse 3-D Data", *IEEE Trans. on PAMI,* vol. 20, no. 11, pp. 1206-1223, 1998.

[16] C.K. Tang, G. Medioni, "Curvature-Augmented Tensorial Framework for Integrated Shape Inference from Noisy, 3-D Data", to appear in *IEEE Trans. on PAMI,* 2001.

[17] Z. Zhang, R. Deriche, L.T. Luong, and O. Faugeras, "A Robust Approach to Image Matching: Recovery of the Epipolar Geometry", Artificial Intelligence Journal, vol. 78, pp. 87-119, 1995.