Rigid Body Segmentation and Shape Description from Dense Optical Flow Under Weak Perspective

Joseph Weber and Jitendra Malik

Abstract—We present an algorithm for identifying and tracking independently moving rigid objects from optical flow. Some previous attempts at segmentation via optical flow have focused on finding discontinuities in the flow field. While discontinuities do indicate a change in scene depth, they do not in general signal a boundary between two separate objects. The proposed method uses the fact that each independently moving object has a unique epipolar constraint associated with its motion. Thus motion discontinuities based on self-occlusion can be distinguished from those due to separate objects. The use of epipolar geometry allows for the determination of individual motion parameters for each object as well as the recovery of relative depth for each point on the object. The algorithm assumes an affine camera where perspective effects are limited to changes in overall scale. No camera calibration parameters are required. A Kalman filter based approach is used for tracking motion parameters with time.

Index Terms—Optical flow, epipolar constraint, fundamental matrix, shape from motion, motion segmentation, scene partitioning problem.

1 Introduction

VISUAL motion can provide us with two vital pieces of information: the segmentation of the visual scene into distinct moving objects and shape information about those objects. We will examine how the use of epipolar geometry under the assumption of rigidly moving objects can be used to provide both the segmentation of the visual scene and the structure of the objects within it.

Epipolar geometry tells us that a constraint exists between corresponding points from different views of a rigidly moving object (or camera). This epipolar constraint is unique to each motion. Optical flow provides a dense set of correspondences between frames. Therefore the unique epipolar constraint can be used to find objects undergoing separate motions given the optical flow. Typically the epipolar constraint is used for large displacement motions, but it is equally valid for optical flow fields which we assume represent small inter-frame displacements.

An algorithm will be outlined for segmenting the scene while simultaneously recovering the motion of each object in the scene. This algorithm makes the assumption that the scene consists of connected piecewise-rigid objects. The image then consists of connected regions, each associated with a single rigid object.

Once the motion of rigidly moving objects has been determined, scene structure can be obtained via the same epipolar constraint. The scene structure problem becomes analogous to stereopsis in that object depth is a function of distance along the epipolar line. Dense correspondences such as those in optical flow can lead to rich descriptions of the scene geometry.

The epipolar geometry will be examined in the context of an af-

- *J. Weber is with the Engineering Department, The California Institute of Technology, Pasadena, CA. 91125. E-mail: weber@vision.caltech.edu.*
- *J. Malik is with Computer Science Division, University of California at Berkeley, Berkeley, CA. 94720-1776. E-mail: malik@cs.berkeley.edu.*

Manuscript received Mar. 9, 1995; revised Apr. 1, 1996. Recommended for acceptance by A. Singh.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P96040.

fine camera model where perspective effects are limited to uniform changes in scale. Under weak perspective, the epipolar constraint equation becomes linear in the image coordinates, thus allowing a least-squares solution for the parameters of the constraint. Different regions of the image representing independently moving rigid objects can then be segmented by the fact that they possess different epipolar constraints on their motion in the image plane. Once the parameters of the constraint equation have been recovered, they can be used to describe the three dimensional rigid motion that each object in the scene has undergone.

2 Review of Past Work

Early work on segmentation via motion looked for discontinuities in the displacement field [17], [2] or piecewise affine partitions of the field [1], [12]. Since under general perspective projections the motion field is continuous as long as the depth of the viewed surface is continuous, discontinuities in the flow field signal depth discontinuities. Unfortunately, the flow field is difficult to recover at discontinuities. At locations of depth edges, motion will introduce regions of occlusion and disocclusion which are often not explicitly modeled in optical flow routines. Optical flow techniques based on derivatives of the image function assume continuous or affine flow and will fail at these regions.

The fact that epipolar geometry implies a linear constraint between the projected points of a rigid body as it undergoes an arbitrary rigid transformation has been used for years in photogrammetry [5] and more recently in structure from motion algorithms [11], [15], [16]. Motion parameters and shape descriptions can also be obtained from correspondences between two views under weak perspective projection, modulo a relief transformation such as depth scaling [8]. Algorithms under this model were implemented by Shapiro et al. [14] and Cernuschi-Frias et al. [4].

In Section 6 we will see that we can formulate the segmentation of the optical flow field into a *scene partitioning problem* [10]. The segmentation problem is formulated in terms of a cost functional which attempts to balance a number of model constraints. These constraints include terms for fitting a model to the data while simultaneously minimizing the number of distinct regions.

There are stochastic [6], region-growing [7], and continuation [10], [3] methods for finding solutions to the scene partitioning problem when it is described in terms of a cost functional. Our solution will use the region-growing method described in [19] to solve for the partition. This method uses a statistic-based region growing algorithm which assumes the solution is piecewise continuous in image coordinates.

3 Projections and Rigid Motions

3.1 The Weak Perspective Camera

The weak perspective camera projection can be written as:

$$\mathbf{x} = M\mathbf{X} + \mathbf{p} \tag{1}$$

where $\mathbf{X}$ is the 3D world coordinate point and $\mathbf{x}$ its 2D image projection. The 2×3 matrix M rotates the 3D world point into the camera's reference frame, scales the axes and projects onto the image plane. The vector $\mathbf{p}$ is the image plane projection of the translation aligning the two frames. The simplest form of the matrix M occurs when the world and camera coordinates are aligned and the camera's aspect ratio is unity. In this case M can be written

$$M = \frac{f}{Z_{ave}}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{2}$$

where Z_{ave} is the average depth of the scene. This transformation is a valid approximation to a real camera only if the variance of the depth in the viewed scene is small compared to Z_{ave}.

A rigid transformation of the world points that takes the point $\mathbf{X}$ to $\mathbf{X}'$ can be written as

$$\mathbf{X}' = R\mathbf{X} + \mathbf{T} \tag{3}$$

where R is a rotation matrix with unit determinant.

Eliminating the depth component Z between equations for $\mathbf{x}$ and $\mathbf{x}'$ we obtain the linear constraint

$$ax' + by' + cx + dy + e = 0 \tag{4}$$

where

$$\begin{aligned} a &= -R_{23},\ b = R_{13} \\ c &= sR_{23}R_{11} - sR_{13}R_{21} \\ d &= sR_{23}R_{12} - sR_{13}R_{22} \\ e &= R_{23}t'_x - R_{13}t'_y \end{aligned} \tag{5}$$

and the vector $\boldsymbol{t}'$ is $sM(\mathbf{T} - R(\mathbf{p}0)^T)$. The scale factor $s = Z_{ave}/Z'_{ave}$ is the fractional change in average depth between frames. More details can be found in [20].

Equation (4) can be written in terms of a special form of the Fundamental Matrix [5].

$$(x', y', 1)F(x, y, 1)^T = 0 \tag{6}$$

3.2 Koenderink and van Doorn Rotation Representation

A rotation in space can be expressed in a number of representations: Euler angles, axis / angle pair, quarternions etc. A particularly useful representation for vision was introduced by Koenderink and van Doorn [8]. In this representation, the rotation matrix is the composition of two specific rotations: the first about the viewing direction (cyclorotation) and the second about an axis perpendicular to the viewing direction at a given angle from the horizontal.

Using this representation in the formation of the Fundamental Matrix as in (4) we find that

$$\begin{aligned} a &= \sin(\rho)\cos(\phi),\ b = \sin(\rho)\cos(\phi) \\ c &= -s\sin(\rho)\cos(\theta - \phi) \\ d &= s\sin(\rho)\sin(\theta - \phi) \\ e &= -\sin(\rho)\left(t'_x\cos(\phi) + t'_y\sin(\phi)\right) \end{aligned} \tag{7}$$

Equations (8) are identical to the ones used in Shapiro et al. [14].

We can invert (8) to find the motion parameters s, ϕ, and θ given the elements of the Fundamental Matrix (a, b, c, d, e). In the next section we explain how to estimate these given the optical flow.

4 Solving for the Fundamental Matrix

The epipolar constraint (4) requires point correspondences between frames. Equating point displacements with optical flow (u, v), we get $(x', y') = (x, y) + (u, v)$ and

$$au + bv + c'x + d'y + e = \vec{\mathbf{x}} \cdot \vec{\mathbf{n}} + e = 0 \tag{8}$$

with $\vec{\mathbf{x}} = \left(u, v, x, y\right)$ and $\vec{\mathbf{n}} = \left(a, b, c', d'\right)$. The epipolar constraint equation elements (4) are related to the primed values by $c' = c + a$, $d' = d + b$.

The affine epipolar constraint equation forces the optical flow to lie on a line in velocity space. Because of noise, the measured optical flow may not lie on the line dictated by the epipolar constraint. We can use weighted least squares to solve for the pa-

rameters $(\vec{\mathbf{n}}, e)$ by minimizing the weighted distance in velocity space between the measured optical flow and the constraint line. The weighting factor for each error term, w_i, comes from the error covariance of the measured optical flow, $\mathrm{var}(\vec{v}^T\vec{v}) = \Omega_{\vec{v}}$.

The following minimization is similar to the one in Shapiro et al. [14]. We define a Lagrange multiplier, λ, on the constraint $a^2 + b^2 = 1$. This constraint can be written as $\|Q\vec{\mathbf{n}}\|^2 = 1$ with an appropriate diagonal matrix Q. The function to be minimized is then

$$\min_{(\vec{\mathbf{n}},e)} \sum_i w_i(\vec{\mathbf{x}}_i \cdot \vec{\mathbf{n}} + e)^2 - \lambda(\|Q\vec{\mathbf{n}}\|^2 - 1) \tag{9}$$

where the summation is over all points with optical flow measurements.

The minimization over e can be done immediately by setting $e = -\bar{\mathbf{x}} \cdot \vec{\mathbf{n}}$ where $\bar{\mathbf{x}} = \sum_i w_i\vec{\mathbf{x}}_i \,/ \sum_i w_i$ is the weighted centroid of the 4D points $\vec{\mathbf{x}}_i$. After substituting for e and differentiating we obtain

$$(W - \lambda Q)\vec{\mathbf{n}} = 0 \tag{10}$$

where the measurement matrix W is $\sum_i w_i(\vec{\mathbf{x}}_i - \bar{\mathbf{x}}_i)^T(\vec{\mathbf{x}}_i - \bar{\mathbf{x}}_i)$. Since Q has only two nonzero entries, finding the value of λ which causes $(W - \lambda Q)$ to drop rank involves only a quadratic equation in λ. The solution $\vec{\mathbf{n}}$ is the vector which spans the null space of $(W - \lambda Q)$. The resulting value of λ is equal to the weighted quadratic error in velocity space.

5 The Case of Affine Flow

The solution for the Fundamental Matrix elements in (11) requires that the matrix $W - \lambda Q$ have rank three, i.e., the null space has dimension one. Multiple solutions can exist if the optical flow is affine in image coordinates. In this case, a linear relationship exists between (u, v) and (x, y), and thus W drops rank. The Fundamental Matrix cannot be uniquely determined. The nontrivial causes of affine flow are either coplanarity of the observed points, or if the object motion is a rotation which contains no rotation in depth (the rotation ρ in the Koenderink and van Doorn representation).

Since the optical flow is corrupted by noise, a criterion must be developed for deciding if a region contains affine flow. The symmetric matrix $W - \lambda Q$ should have rank three and therefore have three positive, nonzero singular values. A region is designated as containing affine flow via a ratio of singular values. A threshold on this ratio is used to label regions as containing affine flow. The magnitude of the threshold comes from the variance estimate produced by the optical flow algorithm used.

6 Segmenting via a Region-Growing Method

We wish to partition the scene into distinct regions, each region being labeled by a unique Fundamental Matrix. We define a cost functional which balances the cost of labeling each pixel with a penalty for having too many different labeling. We define as a total cost functional

$$E(\vec{\mathbf{n}};\alpha) = \sum_i D(\vec{\mathbf{n}}_i) + P(\vec{\mathbf{n}}_i;\alpha) \tag{11}$$

where the summation i is over all pixels in the image. The vector $\vec{\mathbf{n}}_i$ is the estimate of the Fundamental Matrix at pixel i. In terms of the standard form of a cost functional [3], $D(\vec{\mathbf{n}})$ represents a *goodness of fit* term which attempts to keep the estimate close to the data, and $P(\vec{\mathbf{n}};\alpha)$ is a *discontinuity penalty term* which tries to limit the frequency of discontinuities. The $D(\vec{\mathbf{n}})$ term is the weighted sums of squared distances in velocity space with a Lagrange multiplier as defined in the Section 4. The penalty term attaches a fixed cost α for each pixel bordering a discontinuity.

To solve this partitioning problem we will use the region-growing method described in [19]. The algorithm begins by forming small initial patches of size 4×4 pixels. Each of these patches then computes its solution, $\vec{\mathbf{n}}$, and error, D_R. For a small value of the boundary penalty α, all regions which can be combined when a statistic, F', is below a fixed confidence level are merged. Newly formed regions are tested for affine flow solutions. The value of α is increased allowing for more regions to be merged. This continues until we reach the final value of α.

7 Recovering Depth

Once we have recovered the elements of the Fundamental Matrix for a region of the image plane, we can attempt to recover the depth of each image point. From Section 3.1 we find that up to an unknown scale factor:

$$Z = \frac{1}{\sqrt{a^2 + b^2}}(bx' - ay' + dx - cy) + Z_c \tag{12}$$

where $Z_c = -\mathbf{d}^T\mathbf{t}'' \,/\, (s\|\mathbf{d}\|^2)$ with $\boldsymbol{d} = (R_{13}R_{23})^T$. Z_c is a constant for each object. Therefore, up to an additive constant and unknown scale, the depth of each imaged point can be computed given the elements of the matrix F.

In the case of affine flow, we know that the object is either undergoing pure translation or is rotating about an axis parallel to the optical axis. In either case, no depth information can be obtained under orthographic or weak-perspective projection. Consequently depth recovery would have to rely on other cues.

8 Object Tracking

In order to track the segmented objects, the algorithm takes the present segmentation and forms a prediction of the segmentation for the next flow field. The segmentation algorithm is run using this prediction image to fill in the unassigned regions. This is repeated for each new optical flow field.

The proposed scheme avoids having to run the entire segmentation algorithm from scratch at each new frame since it uses the previous segmentation as a prediction. However, this method requires a correct initial segmentation. If two objects are labeled as a single object in the initial segmentation they may remain so in subsequent frames.

We can use the information in each new frame to increase the accuracy of both the shape and motion of each independently moving object. We adopt a Kalman filter approach in which the motion parameters are modeled as a slowly varying process. The work by Soatto et al. [13] addresses the case of estimating the elements of the Fundamental Matrix in a Kalman Filter framework. Although their work was for the full Fundamental Matrix, it is easily adapted to the simpler affine form.

9 Experimental Results

The algorithm was tested on a number of synthetic and real image sequences. The optical flow was computed using the multi-scale differential method of Weber and Malik [20]. Flow fields were about 80% dense with most estimates missing from discontinuous flow regions. These regions violate the constancy assumption used by the differential method.

9.1 Sequence 1

A synthetic sequence was created consisting of two texture-mapped cubes rotating in space. The magnitude of the optical flow ranged from zero to about five pixels/frame. For the first 10 frames of the sequence, the cubes were rotating about fixed but different rotation axes. For the second 10 frames these axes were switched. The rotation axes used, as well as a sample image and optical flow field are shown in Fig. 1.

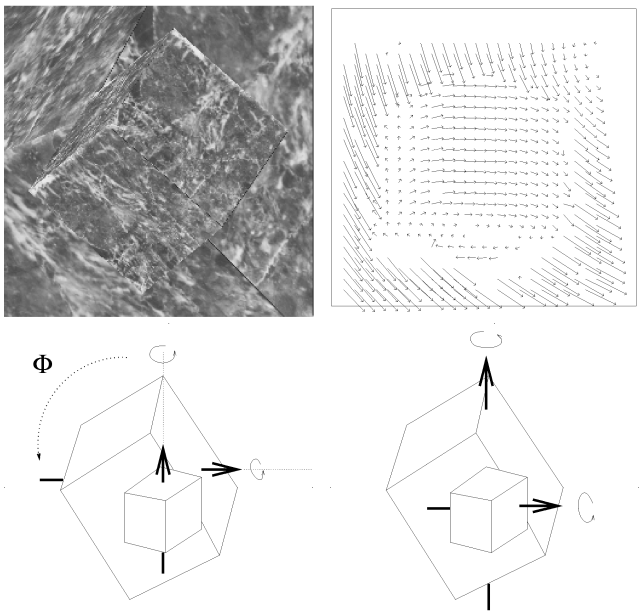

Fig. 1. Two independently rotating texture-mapped cubes were created on a Silicon Graphics workstation. A single frame from the sequence and a sample optical flow field is shown on the top row. No flow estimates were available at the boundaries of the two cubes because such regions violate the constancy assumption used by the differential method. For the first 10 frames, the cubes rotated with rotation axes indicated in the bottom left figure. For the second 10 frames, the rotation axes were as indicated in the bottom right figure.

The segmentation algorithm found two separate moving objects for each frame. The initial segmentation along with the initial depth recovered for the smaller cube is shown in Fig. 2.

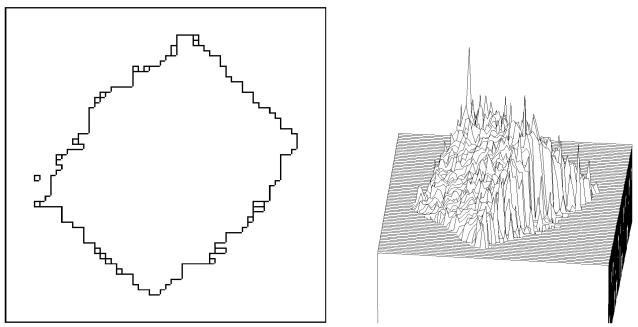

Fig. 2. The boundary between the two independently moving objects found by the segmentation algorithm and the pixel depths of the smaller cube.

The estimated angle ϕ as a function of frame number for each cube is shown in Fig. 3. The original estimate is good because of the density of the optical flow. Subsequent frames do not show much improvement. The Kalman Filter successively tracks the change in rotation axis which occurs at frame 10.

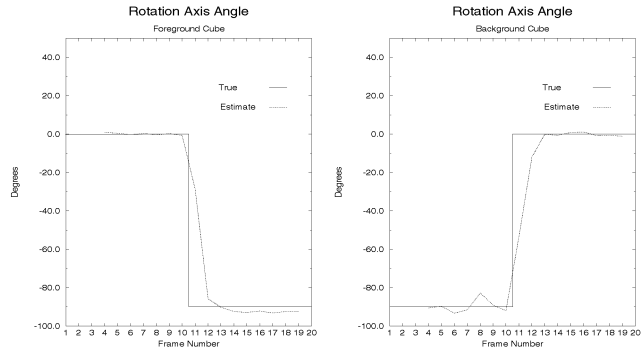

Fig. 3. The recovered value of the angle the rotation axis of each cube makes in the image plane as a function of frame number. After 10 frames, the rotation directions were switched.

9.2 Sequence 2

The algorithm was run on a real sequence consisting of a cube placed on a rotating platen.[1] The background was stationary. The displacements between frames are very small in this sequence, with the largest displacement on the cube itself being only 0.5 pixel. The background had zero flow and was labeled as affine. An image from the sequence, the computed optical flow and recovered depth map are shown in Fig. 4.

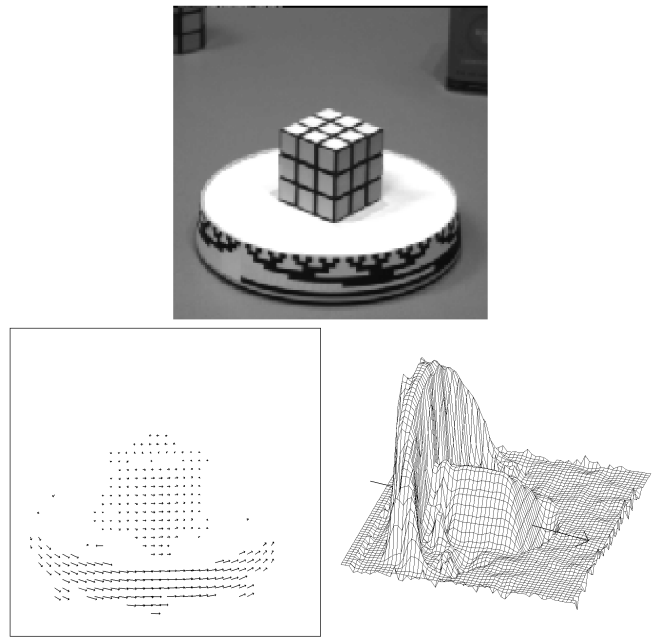

Fig. 4. A single frame of a Rubik's Cube on a rotating platen. The optical flow and recovered depth map as seen from a side view are shown as well.

In this case, the rotation axis of the cube makes an angle of 90 degrees in the image plane and was recovered as such to within a few degrees.

9.3 Sequence 3

The next image sequence contains large planar regions which produce regions of affine flow. A frame from the sequence, an example optical flow recovered and the segmentation are shown in Fig. 5. This sequence demonstrates the algorithm's ability to identify regions of affine flow. The boundaries appear irregular be-

1. This sequence was produced by Richard Szeliski at DEC.

cause no shape priors are used in the segmentation algorithm.

Affine regions will be labeled as distinct if the difference in affine parameters is larger than the expected variance in flow due to noise. The threshold used in the segmentation algorithm is bounded by this noise variance.

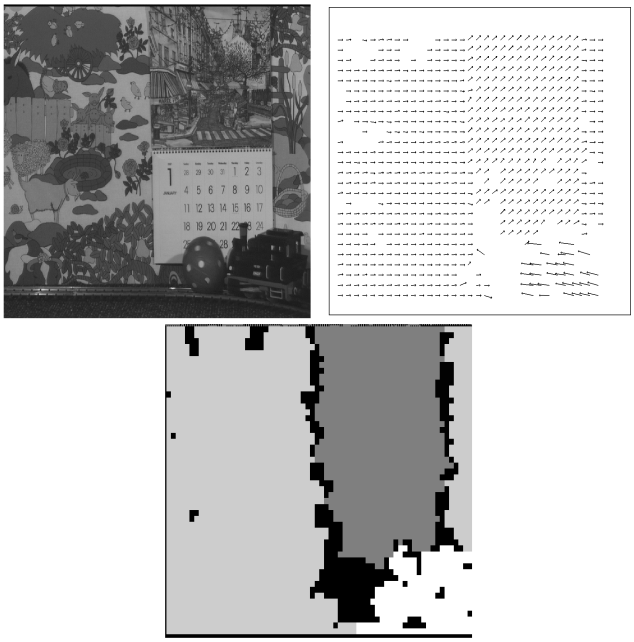

Fig. 5. A single frame of the "mobile" sequence from RPI. The background consists of planar translating patterns while a toy train traverses the foreground. An example optical flow recovered is also shown. The labeled image is shown as well. The background parts (in gray) were identified as undergoing pure translational motion by the singular value ratio test. The black and white regions (corresponding to the train, rotating ball, and transition regions) were not labeled as affine.

10 Discussion

We have shown that using just the optical flow, it is possible to segment an image into regions with a consistent rigid motion and determine the motion parameters for that rigid motion. Furthermore, the relative depth of points within the separate regions can be recovered for each point displacement between the images.

The recovery requires no camera calibration but does make the assumptions of an affine camera: i.e., perspective effects are small. The special form of epipolar geometry for the case considered here has its epipoles at infinity. Perspective dominant motions can not be fit by the motion parameters. The region-growing algorithm used for the simultaneous region formation and motion parameter estimation was not dependent on this particular form of the geometry. If a recovery of the full perspective case was required, the same algorithm could be used. However, the calculation of the Fundamental Matrix from small displacements such as found in optical flow is not stable [18], [9]. This is one of the fundamental limitations of using optical flow with an algorithm based on the epipolar constraint.

Acknowledgments

This research was partially supported by the PATH project MOU 83. The authors wish to thank Paul Debevec for creating the synthetic image sequence.

References

[1] G. Adiv, "Determining Three-Dimensional Motion and Structure From Optical Flow Generated by Several Moving Objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 384-401, 1985.

[2] M. Black and P. Anandan, "Constraints for the Early Detection of Discontinuity From Motion," *Proc. Nat'l Conf. AI*, pp. 1060-1066, Boston, 1990.

[3] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, Mass.: MIT Press, 1987.

[4] B. Cernuschi-Frias, D.B. Cooper, Y.P. Hung, and P.N. Belhumeur, "Toward a Model-Based Bayesian Theory for Estimating and Recognizing Parameterized 3D Objects Using Two or More Images Taken From Different Positions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 1028-1052, 1989.

[5] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, Mass.: MIT Press, 1993.

[6] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721-741, Nov. 1984.

[7] S.L. Horowitz and T. Pavlidis, "Picture Segmentation by a Directed Split-and-Merge Procedure," Proc. Second Int'l Conf. Pattern Recognition, pp. 424-433, 1974.

[8] J.J. Koenderink and A.J. van Doorn, "Affine Structure From Motion," J. Optical Soc. America A, vol. 8, no. 2, pp. 377-385, 1991.

[9] Q.-T. Luong, R. Deriche, O.D. Faugeras, and T. Papadopoulo, "On Determining the Fundamental Matrix: Analysis of Different Methods and Experimental Results," Technical Report RR-1894, INRIA, 1993. A shorter version appeared in the *Israeli Conf. Artificial Intelligence and Computer Vision*.

[10] Y.G. Leclerc, "Constructing Simple Stable Descriptions for Image Partitioning," *Int'l J. Computer Vision*, vol. 3, pp. 72-102, 1989.

[11] H.C. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections," *Nature*, vol. 293, pp. 133-135, 1981.

[12] H.-H. Nagel, G. Socher, H. Kollnig, and M. Otte, "Motion Boundary Detection in Image Sequences by Local Stochastic Tests," *Proc. Third European Conf. Computer Vision*, vol. 2, pp. 305-316, Stockholm, 1994.

[13] S. Soatto, R. Frezza, and P. Perona, "Recursive Motion Estimation on the Essential Manifold," *Proc. Third European Conf. Computer Vision*, vol. 2, pp. 61-72, Stockholm, 1994.

[14] L.S. Shapiro, A.P. Zisserman, and M. Brady, "Motion From Point Matches Using Affine Epipolar Geometry," *Proc. Third European Conf. Computer Vision*, vol. 2, pp. 73-84, Stockholm, 1994.

[15] R.Y. Tsai and T.S. Huang, "Uniqueness and Estimation of Three-Dimensional Motion Parameters of Rigid Objects With Curved Surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 13-27, 1984.

[16] C. Tomasi and T. Kanade, "Shape and Motion From Image Streams Under Orthography: A Factorization Method," *Int'l J. Computer Vision*, vol. 9, no. 2, pp. 137-154, 1992.

[17] W. Thompson, K. Mutch, and V. Berzins, "Dynamic Occlusion Analysis in Optical Flow Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 4, pp. 374-383, 1985.

[18] J. Weng, N. Ahuja, and T. Huang, "Optimal Motion and Structure Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 864-884, 1993.

[19] J. Weber, "Scene Partitioning via Statistic-Based Region Growing," *IS and T SPIE Symp. Electronic Imaging: Science and Technology*, San Jose, Calif., 1995.

[20] J. Weber and J. Malik, "Rigid Body Segmentation and Shape Description From Dense Optical Flow Under Weak Perspective," *Proc. Fifth ICCV*, pp. 12-20, Boston, 1995.

[21] J. Weber and J. Malik, "Robust Computation of Optical Flow in a Multi-Scale Differential Framework," *Int'l J. Computer Vision*, vol. 14, no. 1, 1995.