3D Reconstruction through Segmentation of Multi-View Image Sequences

Carlos Leung and Brian C. Lovell
Intelligent Real-Time Imaging and Sensing Group
School of Information Technology and Electrical Engineering
The University of Queensland, Brisbane, Queensland, 4072, Australia

Abstract

We propose what we believe is a new approach to 3D reconstruction through the design of a 3D voxel volume, such that all the image information and camera geometry are embedded into one feature space. By customising the volume to be suitable for segmentation, the key idea that we propose is the recovery of a 3D scene through the use of globally optimal geodesic active contours. We also present an extension to this idea by proposing the novel design of a 4D voxel volume to analyse the stereo motion problem in multi-view image sequences.

1. Introduction

The reconstruction of a dynamic, complex 3D scene from multiple images is a fundamental problem in the field of computer vision. While numerous studies have been conducted on various aspects of this general problem, such as the recovery of the epipolar geometry between two stereo images [10], the calibration of multiple camera views [30], stereo reconstruction by solving the correspondence problem [24], the modelling of occlusions [9], and the fusion of stereo and motion [14], more work needs to be done to produce a unified framework to solve the general reconstruction problem.

Given a set of images of a 3D scene, in order to recover the lost third dimension, depth, it is necessary to compute the relationship between images through correspondence. By finding corresponding primitives such as points, edges or regions between the images, such that the matching image points all originate from the same 3D scene point, knowledge of the camera geometry can be combined in order to reconstruct the original 3D surface.

One approach to the correspondence problem involves the computation of a disparity map, where each pixel in the map represents the disparity of the matching pixels between two images. The optimisation of a cost function is a common approach in order to obtain the disparity map [8, 21, 22]. Taking advantage of the epipolar constraint, which enables the search area to collapse from a 2-dimensional image to 1-dimensional epipolar lines, along with the ordering [28], uniqueness and continuity constraint [18], algorithms have been proposed which compute the disparity map to sub-pixel accuracy. However, when factors such as noise, lighting variation, occlusion and perspective distortion are taken into account, stereo disparity algorithms are still challenged to model accurately discontinuities, epipolar line interactions and multi-view stereo [6, 11].

Roy and Cox [22] and more recently Kolmogrov [15] developed an algorithm for solving the multi-view stereo correspondence problem. By stacking the candidate matches of range disparity along each epipolar line into a cost function volume, maximum flow analysis and graph cuts are used in order to determine the disparity surface. While these approaches to stereo analysis provide a more accurate and coherent depth map than the traditional line-by-line stereo, these methods remain dependent on and sensitive to the uniqueness and the accuracy of the matching correspondence stage. Although the optimisation of the cost function is performed in a three-dimensional space, the computation of a disparity surface remains only a 2.5-D sketch of the scene [18].

While the aforementioned techniques operate in 1 or 2D space, there also exists a class of stereo algorithms that operate in 3D scene space. Introduced by Collins [5] and Seitz and Dyer [23], these algorithms, instead of using disparity to compute the depth of an image point, directly project each image into a 3D volume, such that the locations of 3D world points are inferred through analysis of each voxel's relationship in 3D space. Kutulakos and Seitz recently proposed the Space Carving Algorithm aimed at solving the N-view shape recovery problem [17]. The photo hull, the volume of intersection of all views, is determined by computing the photo-consistency of each voxel through projections onto each available image. While these approaches produce excellent outcomes, apart from the fact that they require a vast number of input images, improvements can be made by imposing spatial coherence, replacing the voxel-based analysis with a surface orientated technique.

Classical active contours such as snakes [12] and level sets [1] have mainly been applied to the segmentation problem in image processing. The recent introduction of fast implicit active contour models [16], which use the semi-implicit additive operator splitting (AOS) scheme introduced by Weickert et al. [26], is an improved version of the geodesic active contour framework [4]. Given such advancements in active contour analysis, multi-dimensional segmentation is becoming not only more robust and accurate, but computationally feasible. The application of surface evolution and level set methods to the stereo problem was pioneered by Faugeras and Keriven [7]. Although Faugeras' approach is limited to binocular stereo and the epipolar geometry, their novel geometric approach to the stereo problem laid the foundation for a new set of algorithms that can be used to solve the 3D reconstruction problem.

In this paper, we will present two new techniques for 3D scene reconstruction. Firstly, we propose a new approach to 3D reconstruction through the use of globally optimal geodesic active contours. In order to formulate the 3D reconstruction problem suitable for segmentation analysis, we explicitly describe the design of a 3D voxel feature space, which integrates all the information available from each camera view into one unified volume for processing. Rather than solving the correspondence problem between the images by computing disparity and matching feature primitives, and instead of using photo-consistency constraints to determine the colouring of each voxel, our approach projects and integrates all the feature information about each image into one voxel volume. By collapsing the voxel space into a metric space, segmentation algorithms can then be applied to directly reconstruct the complex 3D scene.

Secondly, we propose a new approach for the recovery of 3D models and its motion from multi-view image sequences. While there are many studies in the area of stereo and motion analysis from stereo rigs, we propose the use of a 4D voxel volume to recover not only 3D and motion information from stereoscopic image sequences, but an algorithm capable of processing multi-view image sequences. By augmenting the design of our 3D voxel feature volume to a 4-D feature space, we present a novel approach to the analysis of stereo motion for the reconstruction of dynamic 3D scenes.

Section 2 of this paper will explain in detail the design of the 3D voxel volume and how segmentation can be used to compute the 3D scene. Section 3 will further describe how the 3D voxel volume can be extended to solve the stereo motion problem. The application of 4D reconstruction to analysis multi-view image sequences will be presented. Finally, section 4 will provide a summary of the proposed techniques and directions for future research.

2. 3D Reconstruction

A common approach to stereo reconstruction is the optimisation of a cost function, computed by solving the correspondence problem between the set of input images. The matching problem involves establishing correspondences between the views available and is usually solved by setting up a matching functional for which one then tries to find the extrema. By identifying the matching pixels in the two images as being the projection of the same scene point, the 3D point can then be reconstructed by triangulation, intersecting the corresponding optical rays. Our proposed method differs from this approach by projecting all the images into a common space prior to analysing the correspondence between the images. The matching problem is then solved not as a correspondence problem between images, but as a matching functional, computed for each voxel in the volume. This functional is optimised through segmentation to recover the 3D structure of the scene.

Prior to the construction of the 3D voxel volume, the camera geometry of the images needs to be computed through camera calibration. Knowledge of the camera geometry not only enables the construction of the projection matrix, but also allows the computation of the polyhedral intersection of the camera views. Although solving the bounding region of interest of the images is similar in idea to solving the space carving problem, our proposed method differs greatly from space carving. Rather than deciding on the likelihood of a photo-consistent match for each voxel in 3D space, our method does not perform any computation at the projection stage. Instead, all the feature information is stored inside each voxel and is dependent on the solution to the segmentation problem in order to decide on the 3D surface of the scene.

Assuming a pinhole camera model, the 3D voxel volume is created by projecting all of the images into a 3D polyhedron, such that each voxel contains a feature vector of all the information contained in each camera view. For example, the feature vector can include the RGB values of the voxel's projection into each camera image, the gradient of the image, and even information relating to the projected pixel's neighborhood. A metric volume can then be derived from this voxel space to become the input to the segmentation stage. Cost functions such as the variance between the projections or even a probability density function can be used. Furthermore, by altering the resolution of the voxel volume, the segmentation can output either a dense or a sparse reconstruction of the 3D scene.

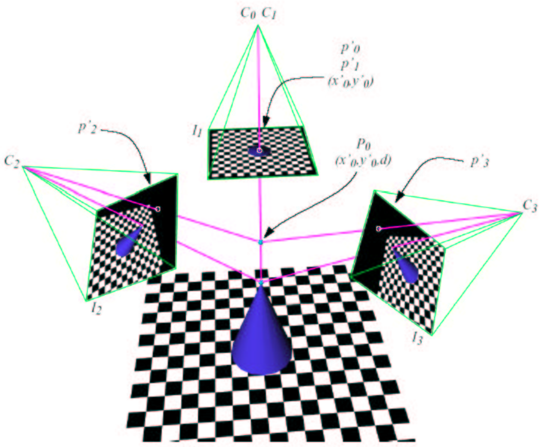

Figure 1. Multi-View Projection. 3D point P_0 is projected onto Images I_1, I_2, I_3. (Image courtesy of S. Roy and I. J. Cox, Figure 1 [22])

2.1. Voxel Volume

We briefly present the well-studied general framework of projective geometry required in the construction of the 3D voxel volume [10]. A set of n input images $I_1, \ldots, I_n$ of a 3D scene are projected from n cameras $C_1, \ldots, C_n$, as depicted in Figure 1 with $n = 3$. In our formulation, we will assume a pinhole camera model and that all surfaces are Lambertian (i.e. the intensity of a 3D point is independent of viewing direction). The projective coordinate of a 3D point P_w in world space is expressed with homogeneous coordinates as

$$P_w = [\begin{array}{cccc} x_w & y_w & z_w & 1 \end{array}]^T$$

while the projective image coordinate of a pixel in image I_i is

$$p_i = [\begin{array}{ccc} x_i & y_i & z_i \end{array}]^T$$

such that the corresponding pixel coordinate p'_i of the projected point p_i can be obtained by applying a homogenising function H where

$$H(\left[\begin{array}{c} x \\ y \\ z \end{array}\right]) = \left[\begin{array}{c} x/z \\ y/z \end{array}\right] \qquad (1)$$

Given the volume of interest of the 3D space for reconstruction, we can obtain each voxel's feature vector by projecting every P_w in the 3D volume onto each of the n images available. With f features for every pixel in the image, each voxel will contain an $f \times n$ matrix, such that the collection of all voxels will contain all the information all the images. In other words, given the 3D voxel volume, all the processing and analysis can be achieved without the need of the original images. We define 4 matrices to describe the projection from a voxel in the volume to a pixel in the image.

Given a volume of M voxels, each voxel will be indexed by its voxel coordinates, $v_m = [v_a, v_b, v_c, 1]^T$, where v_a, v_b and v_c will range from 1 to the dimensions of the volume. The extra parameter appended at the end of the voxel coordinate is for consistency with the augmented homogeneous coordinate in projective space. To transform from voxel coordinates to 3D world coordinates, we compute

$$P_w = \boldsymbol{V} v_m$$

where

$$\boldsymbol{V} = \left[\begin{array}{cc} \boldsymbol{I_3 k} & \boldsymbol{t_v} \\ \boldsymbol{0^T} & 1 \end{array}\right]$$

and $\boldsymbol{I_3}$ is the 3×3 identity matrix, $\boldsymbol{t_v}$ the translation vector for specifying the world coordinate of the voxel volume's origin, and $\boldsymbol{k} = [k_x, k_y, k_z]^T$ is the stride in each voxel dimension, i.e. the number of units between each voxel in 3D world space. The choice of $\boldsymbol{k}$ and $\boldsymbol{t_v}$ is dependent on the resolution desired for the voxel volume and the origin of the volume of interest respectively.

From world coordinates, the classical 3×4 perspective projection matrix, $\boldsymbol{P}$, can be applied to obtain the projection of the 3D world point in image coordinates. In the case where we define the optical centre of the base camera, C_0, to coincide with the origin of the world coordinate system, the projection matrix will be simplified to be

$$\boldsymbol{P_0} = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

subsequently, we can define a transformation matrix $\boldsymbol{W_i}$

$$\boldsymbol{W_i} = \left[\begin{array}{cc} \boldsymbol{R_i} & \boldsymbol{t_i} \\ \boldsymbol{0^T} & 1 \end{array}\right]$$

with rotation $\boldsymbol{R_i}$ and translation $\boldsymbol{t_i}$ to define the position and orientation of camera C_i relative to the base camera, C_0. The relative positions and orientations of each camera i is determined by a calibration procedure. Thus for C_i, the projective projection matrix will be

$$\boldsymbol{P_i} = \boldsymbol{P_0 W_i}$$

From the image coordinates, the pixel coordinates of a projective point can be recovered up to a scaling factor, given knowledge of the internal parameters of the camera. Neglecting radial distortion from calibration, a matrix of intrinsic parameters can be computed such that

$$\boldsymbol{A} = \left[\begin{array}{ccc} -f_x & 0 & o_x \\ 0 & -f_x/\alpha & o_y \\ 0 & 0 & 1 \end{array}\right]$$

where f_x is the focal length in effective horizontal pixel size units, α the aspect ratio (i.e. the vertical to horizontal pixel size), and (o_x, o_y) the image centre coordinates.

Combining all the described matrices, each voxel can be projected into each image i by computing

$$p_i = \boldsymbol{A P_i V} v_m$$

From the obtained projective pixel coordinates, the actual pixel coordinates can be obtained by applying the homogenising function described in Eq. 1.

2.2. The Correspondence Problem

The construction of the voxel volume is purely based on image projections and thus does not require the solution to the correspondence problem in order to produce a dense reconstruction. Unlike algorithms that use disparity maps to guide the 3D reconstruction, dense feature correspondence or area-based matching of every pixel is no longer necessary. However, while the computation of the 3D volume does not require dense correspondences, feature correspondence is needed in establishing the camera geometry and for the purpose of camera calibration. The accuracy of the projections and the reliability of the volume are highly dependent upon robust and accurate camera calibrations. As noted by Medioni [19], in multi-view stereo, there is an imperative need for camera calibration and consistency of matches between multiple-views. Therefore full calibration information needs to be provided with the image set or techniques similar to [31] must be used to obtain the calibration parameters.

The construction of our voxel volume is similar to the concept proposed by Kimura, Saito and Kanade [13] in recovering the 3D geometry from the camera views. Their method is, however, restricted to three images since it is dependent on dense feature correspondences between the input images in order to model the 3D surface and cannot overcome the problem of occlusion since there are no matching points in those regions. Algorithms that depend on dense feature correspondence have much difficulty modelling occlusions. Our proposed method overcomes this problem by redefining the problem, using a new approach that does not depend on pixel to pixel feature correspondence. One of the advantages of our approach is that occlusion does not need to be explicitly modelled. Occluded regions visible in a limited number of images are still projected validly into 3D space for analysis, with the major difference being that less images project to that region. The occluded regions, however, can still be modelled, only that the 3D reconstruction for those region depends on less data. This scheme subsequently also allows for occluded region to be iteratively improved as more images of the occluded scene are available.

2.3. Segmentation

The development of algorithms that can provide globally optimal solutions to segmentation problems makes its application in image processing very attractive. By designing a volume appropriate for maximum-flow analysis, the minimum-cut associated with the maximum flow can be viewed as an optimal segmentation. While Roy and Cox have demonstrated a version of maximum-flow to analyse stereo images, a more computationally feasible method was recently proposed by Sun [24]. A two-stage dynamic programming (TSDP) technique was introduced to obtain efficiently a 3D maximum-surface, which enables the computation of a dense disparity map.

In our voxel volume formulation, since our projected volume enables us to work directly in true 3D coordinates, we aim to output a 3D surface representative of the complete 3D scene rather than using a disparity map to obtain a 2.5-D sketch of the scene [18]. Formulating the 3D reconstruction problem as a segmentation problem has many advantages over the use of the classical dynamic programming technique. In segmentation, optimisation is performed along a surface rather than along a line. This subsequently provides segmentation methods with the advantage of outputting contours that wrap back on themselves, while dynamic programming will have difficulty following these concave surfaces. Rather than reformulating dynamic programming or similar techniques in order to model occlusions and concavity, we propose the use of segmentation to approach 3D reconstruction from a new point of view.

Active contours have been demonstrated to be a useful tool in the segmentation problem. Geodesic active contours that use a variational framework have been shown to obtain locally minimal contours [4]. Fast implicit active contour models, that use the semi-implicit additive operator splitting (AOS) scheme introduced by Weickert et al. [16, 26], and shortest path algorithms [3], have been used to avoid the variational framework producing optimal active contours. By formulating a volume appropriate for 3D segmentation, we propose the use of a form of geodesic active contours recently introduced by Appleton and Talbot [2] which has been demonstrated to be globally optimal. By choosing a positive scalar metric, g, such that g can be assured to be always greater than zero, the minimisation of the energy functional E, can be formulated to describe the segmentation

$$E(C) = \int_C g(C(s))ds$$

where C is the segmentation contour.

3. 4D Reconstruction

The fusion of stereo and motion has been recognised by many researchers as a means of providing additional information that were not previously obtainable through their independent analysis. Waxman and Duncan pioneered the analysis of stereo motion by considering binocular image flow [25]. Many studies have subsequently tackled this problem through the use of Kalman filtering, optical flow and feature tracking [14, 20]. While these methods have demonstrated reasonable success, they are limited by problems inherent in the correspondence problem, as described in section 2.2. Similar to our proposed approach in the use of segmentation, rather than reformulating optical flow or Kalman filtering to model stereo motion, we propose the use of a 4D voxel volume in order to analyse the stereo dynamics in stereoscopic image sequences.

By embedding the design of our 3D voxel feature volume into a 4-D feature space, we present a novel approach to the analysis of stereo motion for the reconstruction of dynamic 3D scenes. Given a set of images captured over different time frames, we can compute the camera geometry and projection parameters for each image through the use of the many calibration techniques developed, such as Zhang's four point algorithm for stereo rig analysis [29]. From the set of projection matrices computed, we can construct a voxel volume for each time frame. Since geodesic active contours can be applied to segment multi-dimensional volumes, similar to the analysis of our 3D voxel volume, we can compute a segmentation in 4D in order to produce a 3D surface in time. The use of this 4D voxel volume also has the advantage of not only recovering the 3D and motion information from stereoscopic image sequences, but is capable of processing multi-view image sequences. The computational feasibility of multi-dimensional segmentation makes this 4D approach to stereo motion an attractive alternative to the analysis of dynamic, complex 3D scenes.

4. Summary and Future Directions

We have proposed two novel approaches to the 3D reconstruction problem through the design of a 3D and 4D feature voxel volume. While current techniques depend heavily upon the solution to the correspondence problem in order to guide the 3D analysis, by taking advantage of the recent developments in segmentation and the introduction of globally optimal algorithms, our method reformulates the computation of correspondence as a segmentation problem. Furthermore, we present a novel approach to the analysis of stereo motion by transforming the problem into a 4D segmentation analysis.

The success of this approach is dependent on the accuracy of the construction of the 3D and 4D voxel volume. Subsequently, the shortcomings of this method are related to its sensitivity to errors in camera calibration and projection. However, with the many developments and studies completed in this area, the error can be minimised. The accuracy of this method also increases as the number of input images increases, making this approach well suited for analysing multi-view image sequences.

The size, shape and location of the voxel volume is also currently manually estimated. Although a difficult and complex problem, a dramatic improvement to the algorithm will be the direct computation of a polyhedral volume of interest. Similar to solving the space carving problem, the polyhedral volume can be obtained by computing the intersections of all camera's field of view. Assuming a pinhole camera model, each camera projection will be a rectangular pyramid, thus the bounding polyhedron will be the solution to the problem of intersecting multiple rectangular pyramids of varying orientations.

The design of a multi-dimensional voxel volume also lays the foundation for 3D or even 4D recognition. A ball for example in 3D space would occupy a spherical volume, while a ball in trajectory can be recognised as a cylindrical tube with hemispherical ends in 4D space. Previous works by Xu [27] have attempted to unify stereo, motion and object recognition into one approach by observing their common use of feature correspondence. Using our new proposed approach, feature correspondence is replaced with a voxel volume that contains all feature information. Thus, through the analysis of a multi-dimensional feature volume, it is possible to design a unified framework for multi-view, motion and object recognition.

Acknowledgements

We would like to thank Ass. Prof. Peter Kootsookos, IRIS Group, School of ITEE, UQ, and Dr. Changming Sun, CSIRO Mathematical and Information Sciences, for the many inspirational discussions and expert advices throughout this research.

References

[1] D. Adalsteinsson and J. A. Sethian. A fast level set method for propagating interfaces. *Journal of Computational Physics*, 118(2):269–277, 1995.

[2] B. Appleton and H. Talbot. Globally optimal geodesic active contours. *Journal of Mathematical Imaging and Vision*, 2002. Submitted.

[3] M. Buckley and J. Yang. Regularised shortest-path extraction. *Pattern Recognition Letters*, 18(7):621–629, 1997.

[4] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[5] R. T. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, pages 358–363, 1996.

[6] I. Cox, S. Hingorani, S. Rao, and B. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996.

[7] O. Faugeras and R. Keriven. Variational principles, surface evolution, pde's, level set methods and the stereo problem. *Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis of the IEEE Transactions on Image Processing*, 7(3):336–344, March 1998.

[8] P. Fua. From multiple stereo views to multiple 3d surfaces. *International Journal of Computer Vision*, 24(1):19–35, 1997.

[9] D. Geiger, B. Ladendorf, and A.Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14:211–226, 1995.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 1998.

[11] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *ECCV*, pages 232–248, Freiburg, Germany, June 1998.

[12] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1998.

[13] M. Kimura, H. Saito, and T. Kanade. 3d voxel construction based on epipolar geometry. In *ICIP*, volume 3, pages 135–139, October 1999.

[14] R. Koch. 3-d surface reconstruction from stereoscopic image sequences. In *ICCV*, pages 109–114, Cambridge, MA., USA, June 1995.

[15] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, volume 3, pages 82–96, 2002.

[16] G. Kühne, J. Weickert, M. Beier, and W. Effelsberg. Fast implicit active contour models. In L. V. Gool, editor, *Pattern Recognition*, Lecture Notes in Computer Science. Springer, Berlin, 2002. To appear.

[17] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. Technical Report TR692, Computer Science Dept., U. Rochester, 1998.

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

[19] G. Medioni. Binocular and multiple-view stereo using tensor voting. Technical report, USC IMSC, 2001.

[20] T. Moyung and P. Fieguth. Incremental shape reconstruction using stereo image sequences. In *ICIP*, 2000.

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

[22] S. Roy and I. J. Cox. A maximum-flow formulation of the n-camera correspondence problem. In *ICCV*, pages 492–499, 1998.

[23] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *CVPR*, pages 1067–1073, 1997.

[24] C. Sun. Fast stereo matching using rectangular subregioning and 3d maximum-surface techniques. *International Journal of Computer Vision*, 47(1/2/3):99–117, May 2002.

[25] A. M. Waxman and J. H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):715–729, Nov. 1986.

[26] J. Weickert, B. ter Haar Romeny, and M. A. Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Trans. Image Proc.*, 7(3):398–410, 1998.

[27] G. Xu. Unification of stereo, motion, object recognition via epipolar geometry. In *ACCV*, volume I287-291, 1995.

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

[29] Z. Zhang. Motion and structure of four points from one motion of a stereo rig with unknown extrinsic parameters. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(12):1222–1227, December 1995.

[30] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[31] Z. Zhang, R. Deriche, L. T. Luong, and O. Faugeras. A robust approach to image matching: Recovery of the epipolar geometry. In *ECCV*, pages 179–186, 1994.