BMVC99

Model Refinement from Planar Parallax

A. R. Dick R. Cipolla
Department of Engineering, University of Cambridge, Cambridge, UK
{ard28,cipolla}@eng.cam.ac.uk

Abstract

This paper presents a system for refining the accuracy and realism of coarse piecewise planar models from an uncalibrated sequence of images. First, dense depth maps are estimated by aligning a planar region of a scene in each image, approximating camera calibration, and generating dense planar parallax. These depth maps are then robustly fused to obtain incrementally refined surface estimates. It is envisaged that this system will extend the modelling capability of existing systems [3] which generate simple, piecewise planar architectural models.

1 Introduction

The acquisition of metric 3D structure from multiple uncalibrated images has been one of the most actively pursued computer vision tasks of recent years. Systems have recently been developed to automatically recover general 3D models from a large number of closely spaced images. For instance, [2] uses robustly tracked features to calibrate each camera and update an estimate of the 3D location of each tracked point in a "quasi-Euclidean" frame. More realistic models are generated in [9] from dense disparity maps calculated between adjacent viewpoints, and bi-directionally linked along the image sequence. Here self-calibration is obtained from the absolute quadric, which unlike quasi-Euclidean frame estimation does not require an initial calibration estimate. As they make no assumptions about the structure of the scene and require many images, these systems are both computationally expensive.

In practice many scenes, especially those in man-made environments, are well approximated by a small number of planar faces. Systems which exploit this constraint, such as Facade [12], can produce highly realistic models of architectural scenes from relatively few, widely spaced images at less computational expense. However, Facade requires a user to specify a coarse polyhedral model of the scene and register it in each each image. This is a laborious process and limits the complexity of the model which can be practically recovered. More recent interactive modelling systems such as PhotoBuilder [3] and [11] require only that the user specify key image features such as parallel or orthogonal lines, but the effort required of the user quickly becomes prohibitive as the desired model complexity and number of images increases.

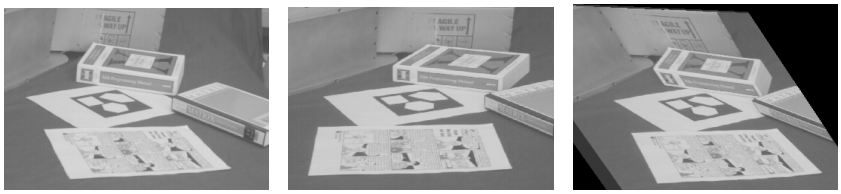
This paper proposes a system which uses planar parallax to enable such interactive systems to recover more complex and realistic models with minimal cost to the user. Given a sequence of images of a near planar scene and a planar model of the scene, it iteratively refines the accuracy of this model. It is particularly suited to architectural model

BMVC 1999 doi:10.5244/C.13.8

refinement, as images of architectural scenes are often dominated by a single approximately planar surface such as a wall. It also complements the more general 3D modelling systems, for which such a scene is a near degenerate case.

2 Theory of planar parallax

A pair of images of a planar (2D) scene is related by a 2D projective transformation, known as a *homography*. The idea of planar parallax is to align two images of a planar region of a scene by applying the homography induced by that region to one of the images (see figure 2). This decomposes the motion between frames into two components: the



(a) Left Image (b) Right Image, Unwarped (c) Right Image, Warped

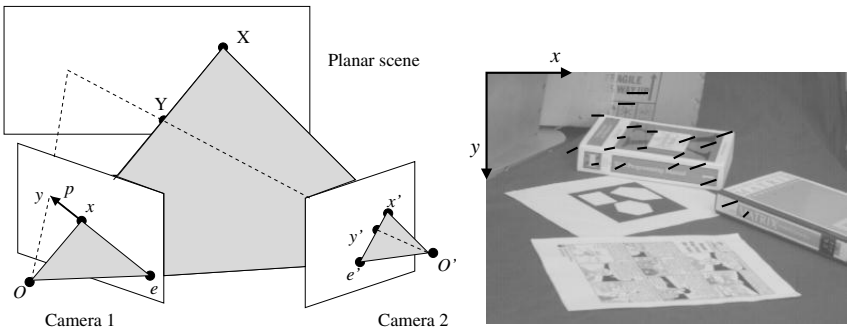
Figure 1: (c) shows the result of warping the right image (b) so that the plane of the tabletop is aligned with the left image (a). Note that all points on the tabletop are aligned while those above the table (such as the top surface of the book) are displaced by a residual parallax.

homography $\mathbf{H}$, and a set of residual motion vectors, known as a *planar parallax* field. It has been shown [4, 13] that for a plane of equation $\mathbf{n}^T\mathbf{X} = d$ in the coordinate system of the first camera, $\mathbf{H}$ is of the form

$$\mathbf{H} = \mathbf{C}[\mathbf{R}d + \mathbf{t}\mathbf{n}^T]\mathbf{C}^{-1}$$

where $\mathbf{C}$ is the camera calibration matrix (assumed to be the same for each camera) and $\mathbf{R}$ is the rotation and $\mathbf{t}$ the translation between the cameras. Thus $\mathbf{H}$ encapsulates the calibration of the cameras and the rotation between them, while the planar parallax field depends only on the translation between the cameras and depth of each point. Because the parallax field depends only on the translational motion of the camera and not its rotation, it converges at the epipole [5], as shown in figure 2. It is clear from figure 2(b) that while it is theoretically possible to compute the epipole from these vectors, in practice the estimation is very unstable, particularly when the epipole is far from the image or the parallax vectors are small.

The planar parallax decomposition has been successfully applied to many problems of computer vision; in particular it complements more general algorithms for which 2D structure is a degenerate case. For instance, novel views of approximately planar scenes are generated using parallax in [7]. Because they separate the effects of translation and



(a) Parallax geometry (b) Example parallax field

Figure 2: (a) The geometry of planar parallax. If x is the projection of a point X belonging to the planar region, its correspondence x' will be aligned with it when image 2 is warped by **H** . However if x is the projection of a non-planar point such as Y, its correspondence after warping will be given by y, resulting in the planar parallax vector p aligned with the epipolar line $\overline{xe}$. (b) Example parallax vectors. The epipole for this image pair has been independently computed to be at position (3187, -31); i.e. far to the right and just above the top of the image.

rotation, image warping and parallax have been used to simplify egomotion estimation, and to detect independently moving objects in a scene [8]. In [4, 13] planar parallax is linked to projective depth, which with camera calibration information can be related to euclidean structure.

3 Model refinement algorithm

3.1 Overview

This section details each step of our algorithm for applying planar parallax to model refinement from multiple uncalibrated images. These steps are summarised as follows:

1. One image is chosen to be the reference image, and one plane chosen as the reference plane. A homography is computed between each image of the reference plane.
2. Each camera is calibrated using these homographies.
3. A dense parallax map is estimated between the reference image and each other image.
4. A depth map is estimated and sequentially refined using the parallax maps. The final result is viewed as a triangulated, texture mapped VRML surface.

3.2 Homography definition

Our calibration method (section 3.3) requires that one image is chosen as the reference image. This image must be approximately front on to the reference plane. The reference plane is selected manually, and the homography between the reference image of the plane and each other image of the plane is then determined. As the homography is a 3×3 matrix defined up to a scale factor, it has 8 degrees of freedom; therefore it is defined by at least 4 corresponding points belonging to the plane. These are selected as the intersection of lines defined by the user, which are automatically fitted to local intensity edges to improve their accuracy.

3.3 Camera calibration

This calibration technique is intended to generate initial estimates of the projection matrices $\mathbf{P}_k$, to be refined at a later stage. It assumes a simple perspective camera model which is well approximated in practice: that the camera has zero skew, an aspect ratio of 1, and that its principal point lies at the centre of the image. Then its projection matrix has the form

$$\mathbf{P}_k = \mathbf{C}\,[\mathbf{R}_k\ |\mathbf{t}_k\ ]$$

where

$$\mathbf{C} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$\mathbf{R}_k$ is a rotation matrix and the vector $\mathbf{t}_k$ defines the translation of the camera centre in the world coordinate system. We define the world coordinate system so that the reference plane has equation $z = 0$, and the origin is given by the intersection of the optical axis of the reference camera with the plane. As the reference image is approximately front on to the reference plane, its rotation matrix $\mathbf{R}_{ref} \approx \mathbf{I}$ and its translation $\mathbf{t}_{ref} \approx [0\ 0\ d]^T$, where d is the distance of the optical centre from the plane. Hence its projection matrix has the form

$$\mathbf{P}_{ref} \approx \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & d \end{bmatrix}$$

and the world plane to reference image homography is given by the scaling

$$\mathbf{H}_{ref}^{wld} \approx \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & d \end{bmatrix}$$

Assuming that all images were taken by the same camera (and hence have the same calibration matrix $\mathbf{C}$), the pose and hence the projection matrix of any other camera k can be calculated from the homography $\mathbf{H}_k^{ref}$ between that image and the reference image. Let

$$\mathbf{P}_k = \mathbf{C}[\mathbf{r}_{1k}|\mathbf{r}_{2k}|\mathbf{r}_{3k}|\mathbf{t}_k]$$

Then

$$\mathbf{H}_k^{wld} = \mathbf{C}[\mathbf{r}_{1k}|\mathbf{r}_{2k}|\mathbf{t}_k]$$

and

$$\mathbf{H}_k^{ref} = \mathbf{H}_k^{wld}\mathbf{H}_{ref}^{wld\,-1} \approx \begin{bmatrix} r_{11k} & r_{21k} & \frac{f}{d}t_{1k} \\ r_{12k} & r_{22k} & \frac{f}{d}t_{2k} \\ \frac{1}{f}r_{13k} & \frac{1}{f}r_{23k} & \frac{1}{d}t_{3k} \end{bmatrix}$$

The focal length f can be estimated using the fact that $\mathbf{r}_1$ and $\mathbf{r}_2$, the first two columns of a rotation matrix, must be orthonormal. d can be set arbitrarily to 1, as there is an ambiguity between magnitude of the camera translation and its distance from the plane. $\mathbf{P}_k$ is then completely determined by setting $\mathbf{r}_3$ to be the cross product of $\mathbf{r}_1$ and $\mathbf{r}_2$.

3.4 Finding correspondences

Before finding a disparity map, we warp each image so that the reference plane is aligned with the reference image of the plane. This greatly simplifies the correspondence problem and facilitates the recovery of dense disparity maps, as shown in [12, 9].

To obtain a parallax field we apply a complex wavelet transform to each image and perform multi-resolution matching based on the phase of the output coefficients at each level. Due to some redundancy in the wavelet representation, similarity surfaces can be interpolated between pixel locations, which allows subpixel accuracy in matching and provides a directional confidence measure for each matched pair. More details are given in [10]. As the cameras are calibrated we weight the matching constraint to favour correspondences along epipolar lines. At each resolution the parallax field is regularised by minimising the energy functional

$$E(\{\mathbf{u}\}) = E_{sm}(\{\mathbf{u}\}) + \lambda E_{ap}(\{\mathbf{u}\})$$

where $E_{sm}(\{\mathbf{u}\})$ measures the variation of the field $\{\mathbf{u}\}$, $E_{ap}(\{\mathbf{u}\})$ is a measure of the error between $\{\mathbf{u}\}$ and the unsmoothed parallax field, weighted by the confidence of each matched pair, and λ is a scale factor controlling the relative influence of these two terms (we use λ = 0.2). [1] suggests appropriate formulae for these error terms and a method for minimising $E(\{\mathbf{u}\})$. Because it assumes that the scene is a smooth surface, this algorithm is vulnerable to surface discontinuities and occlusion. We address this problem when interpolating the disparity field from coarse to fine by considering each of the four coarse level disparities surrounding a finer level disparity as a candidate for the interpolated disparity. Rather than averaging these candidates we choose the one which results in the best match at the finer level, thereby eliminating the influence of most mismatches at the coarse level and preserving sharp changes in depth. The problem of matching failure is also tackled in the depth estimation stage, as described in section 3.5. This algorithm matches two 640×480 images in less than a minute on a Sun Ultra 1 workstation, and is used to find correspondences between the reference image and each other image in the sequence.

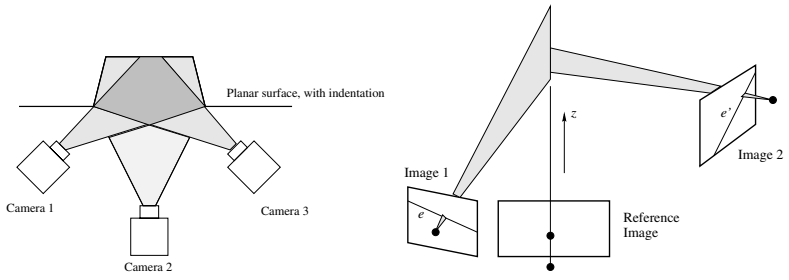
3.5 Depth estimation

Our initial esitmate of depth, derived solely from the reference image, is $z = 0$ for all points; that is, the initial structure is given by the reference plane. To update the depth map we unwarp the correspondences between a selected image and the reference image, and triangulate using the method proposed by Hartley and Sturm [6]. This is based on

the observation that matching errors arise in the image plane, and hence finds the pair of epipolar lines which minimise the total image distance to the pair of correspondences, before correcting the correspondences so that they lie on these lines. This is optimal under the assumption that correspondences are perturbed by isotropic, homogeneous Gaussian noise. By matching between the aligned images and triangulating between the unaligned images, we take advantage of both the simplified correspondence of small baseline stereo and the accurate depth triangulation of wide baseline stereo.

By triangulating between the reference image and each other image in turn we obtain a set of depth estimates for each point in the reference image. It was found that for the small number of images used, the simplest and most effective way to avoid the influence of outliers is to set the current depth estimate to be the median of the depth estimates obtained so far.

However this strategy is not reliable near depth discontinuities, where the correct match may be visible only in a minority of images (see figure 3). If a point is found to be near a possible discontinuity, we use instead only the depth estimates from the cameras which are in a position to avoid occlusion. We presently detect possible discontinuities by measuring the local variation in the current depth estimates.



(a) Partially occluded surface (b) Depth uncertainty

Figure 3: (a) The area to the left of the indentation is occluded from camera 1; in this case our algorithm uses only the depth estimate from cameras 2 and 3 for points in this region. Similarly the area to the right would be estimated only from cameras 1 and 2. (b) The variation in depth due to perturbation along epipolar line e is greater than that caused by a similar perturbation along e', as image 2 is more widely separated from the reference image than image 1.

To improve the appearance of the reconstruction, if a depth estimate is within an interval containing 0, its displacement is considered to be due to inaccuracies caused by the image resampling and matching processes, and it is set to 0. The limits of this interval are set to the depths obtained by perturbing the matched point, already lying on an epipolar line due to the triangulation process, 1 pixel in either direction along its epipolar line and triangulating again. In this way, similarly to [9], we obtain an idea of the uncertainty of depth based on camera position (see figure 3(b)).

4 Results

As this system is intended to augment the PhotoBuilder package, its major application domain is refining models of approximately planar architectural scenes. We give here two examples of our algorithm operating on sequences of images of such scenes. These images were taken with a Fuji MX-700 digital camera, and each has a resolution of 640 by 480 pixels. Both results were obtained using only the initial estimates of $\mathbf{P}_k$ generated by the calibration algorithm. These scenes are problematic for many structure from motion algorithms, because only minor depth variation is present.

Figure 4 illustrates the effect of our occlusion detection scheme on a three image sequence of a stone wreath located outside Fitzwilliam Museum, Cambridge. Because the right facing parts of the wreath are not visible from the viewpoint of camera 1, their reconstruction based on this viewpoint is poor. Similarly the left facing regions are poorly reconstructed from the viewpoint of camera 2. However the cumulative reconstruction is accurate in all regions except where the top of the wreath joins the wall, which is not visible from any viewpoint. These reconstructions are obtained by downsampling the depth map and triangulating between the points to form a VRML surface. The final reconstruction is texture mapped using the reference image. In future we hope to improve the texture mapping algorithm so that parts of the scene not visible in the reference image are rendered using an image in which they are visible.

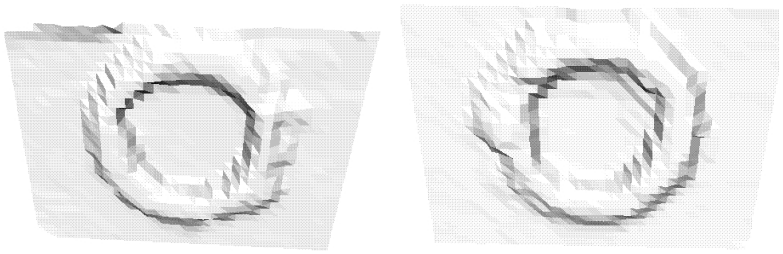
Figure 5 shows five images of a gateway, the most significant features of which are the central inset gate and a pair of shallow rectangular hollows to either side of it. Below these images are four stages in the evolution of the intially planar surface as the parallax map from each image is triangulated and incorporated in the depth estimate. Regions of minor deviation from the plane, and the rectangular indentations which exhibit only gradual depth variation, evolve gradually over the sequence. The left side of the gate, which is occluded in images 1 and 2, is not accurately reconstructed until the incorporation of images 3 and 4. Although the right side of the gate is occluded and hence poorly estimated in images 3 and 4, its cumulative estimate, shown in figure 5 (h) and (i), is not degraded. Figure 6 shows two views of a realistic texture mapped reconstruction of the gateway.

5 Conclusion

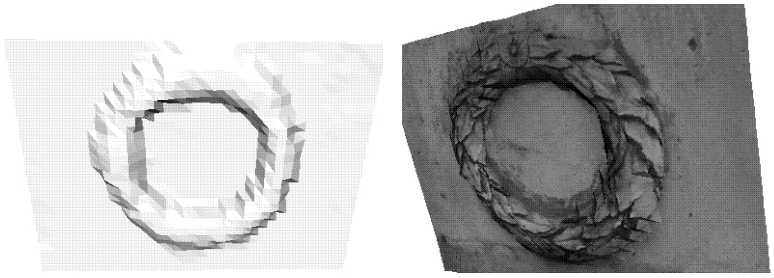
This paper has presented a system comprising a novel combination of computer vision techniques for the incremental refinement of planar model surfaces. Few images are required to obtain an accurate set of depth estimates, which greatly enhances the accuracy and realism of the resulting reconstruction at very little cost to the user. No camera calibration information is required.

Future development will focus on the complete automation of the algorithm and on improving its accuracy. For instance we plan to use the intermediate depth maps to predict more accurately the location of each pixel in each newly incorporated image, which will simplify the matching process and indicate the accuracy of the current structure estimate. It is anticipated that this system will be integrated into the PhotoBuilder package to automatically refine the piecewise planar models which it currently produces, thus offering a simple, convenient and computationally inexpensive system for the generation of extremely realistic 3D models.

(a) Reference Image (b) Image 1 (c) Image 2

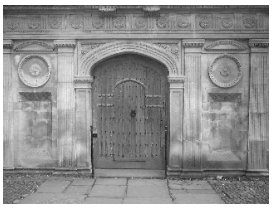


(d) Reconstruction from camera 1 only (e) Reconstruction from camera 2 only



(f) Fused reconstruction (g) Texture mapped reconstruction

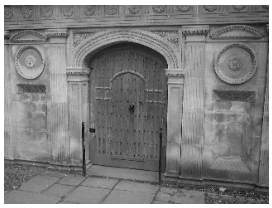
Figure 4: Stonework wreath reconstructions, shown as untextured and (g) textured VRML surfaces. Areas which are occluded from either viewpoint are reconstructed poorly from that viewpoint, but the final reconstruction is accurate for regions visible from any viewpoint.



(a) Reference Image

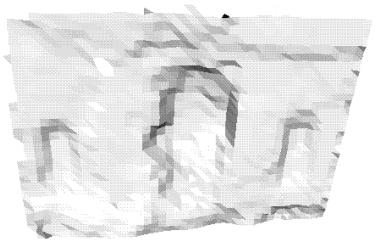
(b) Image 1

(c) Image 2

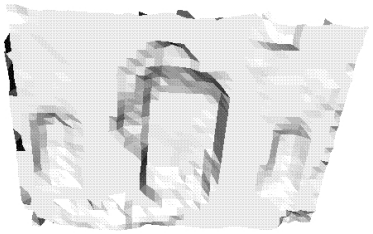


(d) Image 3

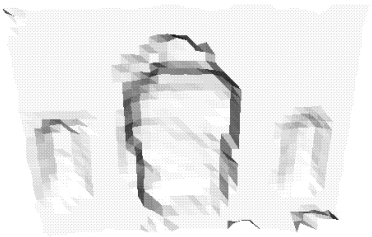
(e) Image 4



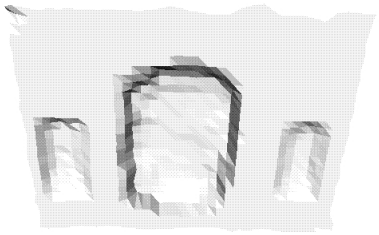
(f) Surface after incorporating image 1



(g) Surface after incorporating images 1 and 2



(h) Surface after incorporating images 1, 2 and 3



(i) Surface after incorporating images 1, 2, 3 and 4

Figure 5: The reference image and 4 other views of a gateway at Caius College, Cambridge. The major features are the inset gate and a pair of rectangular indentations to either side of it. Below these images are untextured VRML surfaces showing the evolution of the surface as each depth map is incorporated.

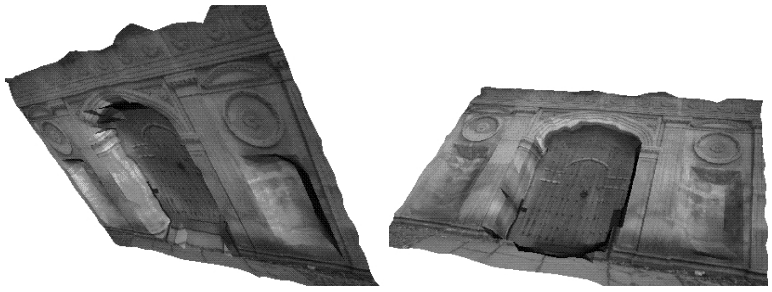


Figure 6: Two views of the refined planar model of the Caius gateway.

Acknowledgement

We wish to thank Tom Drummond for his suggestion of the camera calibration algorithm.

References

[1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, January 1989.

[2] P.A. Beardsley, A. Zisserman, and D.W. Murray. Sequential updating of projective and affine structure from motion. *International Journal of Computer Vision*, 23(3):235–259, 1997.

[3] R. Cipolla, D. Robertson, and E. Boyer. Photobuilder – 3d models of architectural scenes from uncalibrated images. In *IEEE Int. Conf. on Multimedia Computing and Systems*, 1999.

[4] R. Collins. Projective reconstruction of approximately planar scenes. In *Interdisciplinary Computer Vision: An Exploration of Diverse Applications*, pages 174–185, 1992.

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

[6] R.I. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, November 1997.

[7] M. Irani, P. Anandan, and D. Weinshall. From reference frames to reference planes: Multi-view parallax geometry and applications. In *European Conference on Computer Vision*, pages 829–846, 1998.

[8] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using region alignment. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(3):268–272, March 1997.

[9] R. Koch, M. Pollefeys, and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In *European Conference on Computer Vision*, pages 55–71, 1998.

[10] J. Magarey and N. Kingsbury. Motion estimation using a complex-valued wavelet transform. *IEEE Trans. Signal Processing*, 46(4):1069–1084, April 1998.

[11] H.Y. Shum, M. Han, and R. Szeliski. Interactive construction of 3d models from panoramic mosaics. In *IEEE Computer Vision and Pattern Recognition*, pages 427–433, 1998.

[12] C.J. Taylor, P.E. Debevec, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. Technical report, University of California at Berkeley, 1996.

[13] T. Vieville, C. Zeller, and L. Robert. Using collineations to compute motion and structure in an uncalibrated image sequence. *International Journal of Computer Vision*, 20(3):213–242, 1996.