Estimating the Structure of Textured Surfaces Using Local Affine Flow

Andrew Calway
Department of Computer Science
University of Bristol, UK
andrew@cs.bris.ac.uk

Abstract

This paper describes a novel approach for recovering the structure and motion of a rigid textured surface from an image sequence. Camera focal length is also recovered, yielding metric estimates of the structure without the need for pre-calibration. The key innovation is the use of local *affine flow parameters* as the measurements within an extended Kalman filter (EKF) estimation framework, in contrast to feature correspondences or optical flow used in previous approaches. This enables surface normals to be recovered in addition to depth, unlike a feature correspondence scheme, but without the computational limitation of an optical flow approach. The method is based on equating the affine parameters to a local linearisation of the 2-D motion field and using the EKF to provide recursive estimates of the 3-D structure and motion. Experiments on both synthetic and real sequences demonstrate that the approach has considerable potential.

1 Introduction

The problem of how to recover 3-D rigid structure and motion in a scene from 2-D motion observed in an image sequence has been the subject of considerable research in computer vision. Although diverse, this research can be broadly divided into two classes depending on the type of measurements used: either feature correspondences tracked over time; or optical flow fields obtained from adjacent frames. The former has received the most attention and considerable advances have been made in developing algorithms using both calibrated and uncalibrated cameras [5, 11, 6].

Nevertheless, there are drawbacks to using feature correspondences. Solving the correspondence problem for scenes not containing well-defined features such as corners, etc, can be problematical. This is the case when dealing with textured surfaces, for example, in which feature location from frame to frame is often ambiguous. Feature-based techniques also result in sparse point-wise structure information which needs to be interpolated if full 3-D structure is to be obtained. Dense structure can be obtained following determination of epipolar geometry, as in [14] for example, although such methods rely on sufficient translations between frames [6].

In contrast, optical flow measurements are better suited to textured surfaces and incremental motion between frames, and have the potential to provide dense structure in-

BMVC 2000 doi:10.5244/C.14.10

formation. Tracking is, however, more difficult when dealing with optical flow, requiring intermediate segmentation for example, and the large number of data points leads to a significant increase in computation. Consequently, although the computational problems can be reduced to some extent [18], optical flow schemes have been based on either sampling the flow field or limiting the estimation to pairs of frames [1, 4, 12]. These and other optical flow approaches also assume knowledge of the camera parameters, ie focal length, and hence require a pre-calibration step if metric estimates are to be obtained.

The method described here is designed to address these problems. The key innovation is the use of *affine flow parameters* as the measurements in an extended Kalman filter (EKF) estimation framework. These are computed and tracked within local windows in a sequence and approximate the 2-D motion fields generated by surface patches moving in the scene. The advantage of using them in preference to dense optical flow measurements is twofold: the number of 'affine patches' will be relatively small in a typical sequence; and they provide an effective way of tracking motion fields [16, 10, 17]. Despite this, direct use of affine flow parameters for recovering 3-D information has received surprisingly little attention. The notable exceptions are Negahdaripour and Lee [13] and Meyer [9], who describe closed-form solutions using the formulations of Longuet-Higgins and Pradzny [7] and Subbarao and Waxman [15]. However, these are 2-frame deterministic algorithms which are sensitive to noise, relying on affine parameter differences [13, 9] and flow time derivatives [9]. They also require knowledge of the camera parameters.

The noise sensitivity problem is tackled here by using an EKF to provide recursive estimates of the structure and motion in a similar way to that of Murray and Shapiro [12]. However, it differs in two important respects. Firstly, a local linearisation of the motion field about the centres of the projected surface patches is used in order to equate with the affine flow parameters. Secondly, the method is based on a camera geometry that decouples focal length from depth, as used by Azarbayejani and Pentland in their feature correspondence technique [2], enabling estimation of the focal length and so providing metric estimates of the structure without the need for pre-calibration. These are, for each patch, a depth and surface normal for the corresponding point on the 3-D surface, in addition to the global rectilinear and angular velocities of the surface. The result is a richer description of the structure than that provided by a feature correspondence technique, but obtained without the computational disadvantages of an optical flow scheme.

The next section describes the camera, motion and structure models used, followed in Section 3 by details of the estimation procedure. Results of experiments performed on synthetic and real sequences are presented in Section 4. The paper concludes with a discussion on the performance of the technique and directions for future work.

2 Camera, Motion and Structure Models

We assume that a rigid surface in a 3-D scene is moving relative to a stationary camera. Points in the scene are defined with respect to the camera reference frame, with the z-axis aligned with the optical axis and the image plane lying in the xy-plane. As in [2] and as illustrated in Fig 1, the coordinate system origin is at the point where the optical axis cuts the image plane. For this model, perspective projection of a 3-D point $\mathbf{X} = (X_1, X_2, X_3)$ onto an image point $\mathbf{x} = (x_1, x_2)$ is given by

$$x_i = X_i/(1 + \beta X_3) \tag{1}$$

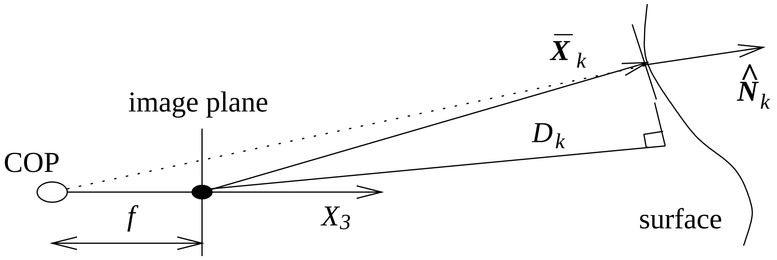

Figure 1: Camera and surface model

where $\beta = 1/f$ is the inverse focal length. In contrast to the usual approach in which the origin is at the centre of projection, this model decouples the representation of depth from that of the camera, an essential element when trying to estimate both depth and focal length [2]. The 2-D motion $\dot{\mathbf{x}} = (\dot{x}_1, \dot{x}_2)$ induced by the motion of a 3-D point is then

$$\dot{x}_i = (\dot{X}_i - \beta x_i \dot{X}_3)/(1 + \beta X_3) \tag{2}$$

and after expressing the 3-D motion in terms of the instantaneous rectilinear and angular velocities, $\mathbf{T} = (T_1, T_2, T_3)$ and $\mathbf{\Omega} = (\Omega_1, \Omega_2, \Omega_3)$ respectively, ie

$$\dot{\mathbf{X}} = \mathbf{\Omega} \times \mathbf{X} + \mathbf{T} \tag{3}$$

we arrive at an alternate form of the basic motion equations:

$$\begin{aligned} \dot{x}_1 &= \Omega_2 \beta x_1^2 - \Omega_1 \beta x_1 x_2 - \Omega_3 x_2 + (T_1 - \beta x_1 T_3 + \Omega_2 X_3)/(1 + \beta X_3) && (4) \\ \dot{x}_2 &= -\Omega_1 \beta x_2^2 + \Omega_2 \beta x_1 x_2 + \Omega_3 x_1 + (T_2 - \beta x_2 T_3 - \Omega_1 X_3)/(1 + \beta X_3) && (5) \end{aligned}$$

Note that because of the difference in origin, these equations differ from those for the usual camera model [7] in that depth and angular velocity are no longer decoupled.

The structure model is based on a local planar approximation. If $\hat{\mathbf{N}}_k$ denotes the unit normal of a surface at a point $\bar{\mathbf{X}}_k$, then combining the equation of the tangent plane, ie

$$\hat{N}_{k1} X_1 + \hat{N}_{k2} X_2 + \hat{N}_{k3} X_3 = D_k \tag{6}$$

with the projection model in eqn (1), we can approximate the variation in depth about the projected point $\bar{\mathbf{x}}_k$ as

$$\alpha_k(\mathbf{x}) \approx \frac{D_k - \hat{N}_{k1} x_1 - \hat{N}_{k2} x_2}{\beta(\hat{N}_{k1} x_1 + \hat{N}_{k2} x_2) + \hat{N}_{k3}} \tag{7}$$

where $D_k = \bar{\mathbf{X}}_k . \hat{\mathbf{N}}_k$ is the perpendicular distance of the surface point from the origin as shown in Fig. 1. Replacing X_3 with this expression in equations (4) and (5) then gives a non-linear expression for the motion field about $\bar{\mathbf{x}}_k$ in terms of the spatial coordinates, 3-D motion, surface normal and focal length. Denoting this expression by $\mathbf{v}_k(\mathbf{x})$, where k indicates the dependence on the local planar structure, we can now obtain a six parameter affine approximation to the motion field by linearising about the projection centre $\bar{\mathbf{x}}_k$, ie

$$\dot{\mathbf{x}} \approx \mathbf{v}_k(\bar{\mathbf{x}}_k) + \nabla \mathbf{v}_k(\bar{\mathbf{x}}_k)(\mathbf{x} - \bar{\mathbf{x}}_k) \tag{8}$$

where $\nabla \mathbf{v}_k(\mathbf{x})$ is the Jacobian of $\mathbf{v}_k(\mathbf{x})$. Note that the affine parameters are defined in terms of a non-linear function of the 3-D parameters. It is measurements of these affine parameters that we use as the data in the estimation process described in the next section.

3 Recursive Estimation

We assume that in an image sequence we can identify K ‘affine patches’ corresponding to a single surface moving with rigid motion, where the 6 affine parameters define an approximation to the local 2-D motion field within each patch. Moreover, that we have tracked the patches over time as the surface moves. Given these measurements, we wish to use equation (8) to obtain estimates of the instantaneous 3-D motion of the surface and the 3-D position and surface normal associated with each patch.

This is done using an extended Kalman filter [3, 2, 12]. The state vector consists of the 6 motion parameters, the inverse focal length and $3K$ structure parameters, ie

$$\mathbf{s} = (T_1, T_2, \beta T_3, \mathbf{\Omega}, \beta, \mathbf{N}_0, \ldots, \mathbf{N}_{K-1}) \tag{9}$$

where

$$\mathbf{N}_k = (\hat{N}_{k1}/\hat{N}_{k3}, \hat{N}_{k1}/\hat{N}_{k3}, D_k/\hat{N}_{k3}) \tag{10}$$

We use the product βT_3 to reflect the dependence on focal length of the sensitivity to motion along the optical axis [2] and normalise the structure parameters by $\hat{N}_{k3}$ (which will always be significantly greater than zero for visible patches) to minimise the number of state variables by eliminating the redundant degree of freedom. The measurement vector contains the $6K$ affine parameters and hence the measurement $\mathbf{z}(t)$ at time t is related to the state $\mathbf{s}(t)$ by

$$\mathbf{z}(t) = \mathbf{h}(\mathbf{s}(t)) + \mathbf{w}(t) \tag{11}$$

where $\mathbf{w}(t)$ is zero mean with covariance R. The non-linear observation model $\mathbf{h}(\mathbf{s})$ is defined by equation (8), ie for $0 \leq k < K$

$$\begin{array}{lll} h_{6k+1}(\mathbf{s}) = \frac{\delta v_{k1}}{\delta x_1} & h_{6k+2}(\mathbf{s}) = \frac{\delta v_{k1}}{\delta x_2} & h_{6k+5}(\mathbf{s}) = v_{k1} \\ h_{6k+3}(\mathbf{s}) = \frac{\delta v_{k2}}{\delta x_1} & h_{6k+4}(\mathbf{s}) = \frac{\delta v_{k2}}{\delta x_2} & h_{6k+6}(\mathbf{s}) = v_{k2} \end{array} \tag{12}$$

where v_{ki} and $\delta v_{ki}/\delta x_j$ are evaluated at $\bar{\mathbf{x}}_k$.

The state dynamics for the filter are defined by the equation

$$\mathbf{s}(t+1) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{e}(t) \tag{13}$$

where $\mathbf{e}(t)$ is zero mean with covariance Q and $\mathbf{f}(\mathbf{s})$ is the state transition function. No prior knowledge of the dynamics is assumed and hence we use an identity transition for the motion and focal length, ie $f_i(\mathbf{s}) = s_i$ for $1 \leq i \leq 7$. The motion states then define the evolution of the normals and depths as derived by Murray and Shapiro [12]. Details of this are not repeated here; it suffices to note that it yields a non-linear state transition in the state variables. Recursive estimation of the state then proceeds as for a standard EKF [3], with the main complications being the need to compute the Jacobians of the measurement and state models, ie $\nabla \mathbf{h}$ and $\nabla \mathbf{f}$, and to determine suitable values for the model covariances Q and R. Although the former are somewhat involved, they are readily obtained using a mathematical package such as Maple. The resulting filter does not require excessive computation, the main part being the inversion of a $6K \times 6K$ matrix, and hence, since the system is overdetermined for $K \geq 3$, real-time implementation is possible on a high-performance workstation. The model covariances were arrived at through empirical means using standard EKF design procedures [3].

4 Experiments

4.1 Point Data Sets

To ensure the EKF was working correctly, we first tested the estimation process using synthetic point data. This consisted of clusters of 3-D points on a moving sphere projected onto the image plane of a stationary camera. We used 8 clusters and the surface was translated along all three axes while also rotating about both the x and y axes using a sinusoidal variation (with respect to the camera reference frame). The focal length was set to 1.0 and all points were visible over all frames using a 53^o field of view. Each cluster occupied approximately 6.5^o corresponding to 11% of the image plane. Affine flow parameters relating corresponding clusters were then obtained for each 'frame' using a standard least-squares fit, ie denoting the ith projected point of the kth cluster at time t by $\mathbf{x}_{ki}(t)$, the flow parameters are given by the 2×2 matrix $A_k(t)$ and the 2×1 vector $\mathbf{b}_k(t)$ which minimise

$$e_k(t) = \sum_{i=0}^{N-1} ||\mathbf{v}_{ki}(t) - A_k(t)\mathbf{x}_{ki}(t) - \mathbf{b}_k(t)||^2 \tag{14}$$

where $\mathbf{v}_{ki}(t) = \mathbf{x}_{ki}(t+1) - \mathbf{x}_{ki}(t)$ denotes the motion of the ith point at time t. These were then used to fill the measurement vector for the EKF, ie from eqns (11) and (12)

$$\begin{array}{lll} z_{6k+1}(t) = A_{k11}(t) & z_{6k+2}(t) = A_{k12}(t) & z_{6k+5}(t) = b_{k1}(t) \\ z_{6k+3}(t) = A_{k21}(t) & z_{6k+4}(t) = A_{k22}(t) & z_{6k+6}(t) = b_{k2}(t) \end{array} \tag{15}$$

In the experiments, we added Gaussian noise to each of the parameters, and Figure 2 shows the variation of the 6 noisy measurements over time for one of the clusters where the noise standard deviation for each parameter was set to approximately 10% of its maximum (absolute) value. For these measurements, the evolution of the true and estimated 3-D parameters over 200 frames is shown in Fig. 3, where the structure is shown as the slant α and tilt τ of the surface normal and the depth D. Note that the motion and inverse focal lengths converge rapidly, whilst the structure converges within 100 frames (the translations and depths converge to within a common scale factor). Convergence is rapid for all the states when uncorrupted measurements are used.

4.2 Texture Mapped Surfaces

To test the estimator using measurements obtained from an image sequence we first synthesised 3-D scenes consisting of textured surfaces moving with respect to a stationary camera. The geometry was similar to that used in the first experiment and the surfaces were translated in the xy-plane and rotated about the three axes. We experimented with the three surfaces shown in Fig. 4 - a plane (top), a Gaussian (middle) and a saddle type surface (bottom) - and texture mapped a tree bark image onto each to generate image sequences consisting of frames of size 256×256 pixels. Sixteen 32×32 pixel regions were then tracked through each sequence by evolving Gaussian windows according to affine flow parameters computed within each region using a weighted least-squares fit over an optical flow field. The optical flow was computed using the Lucas and Kanade algorithm

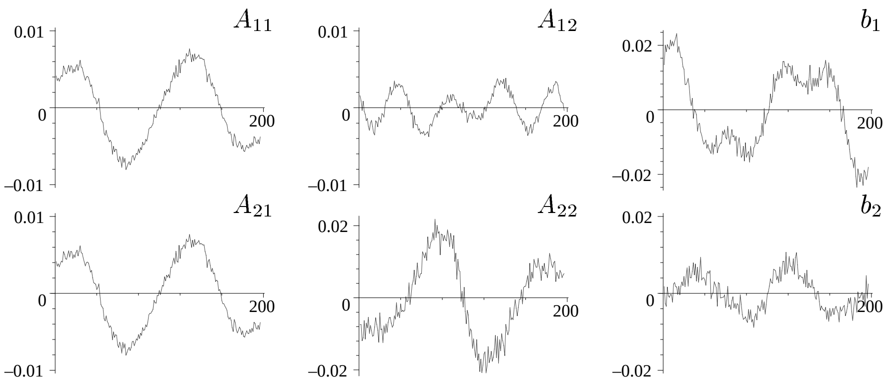

Figure 2: Noisy affine flow parameters for a cluster used in the point data experiment.

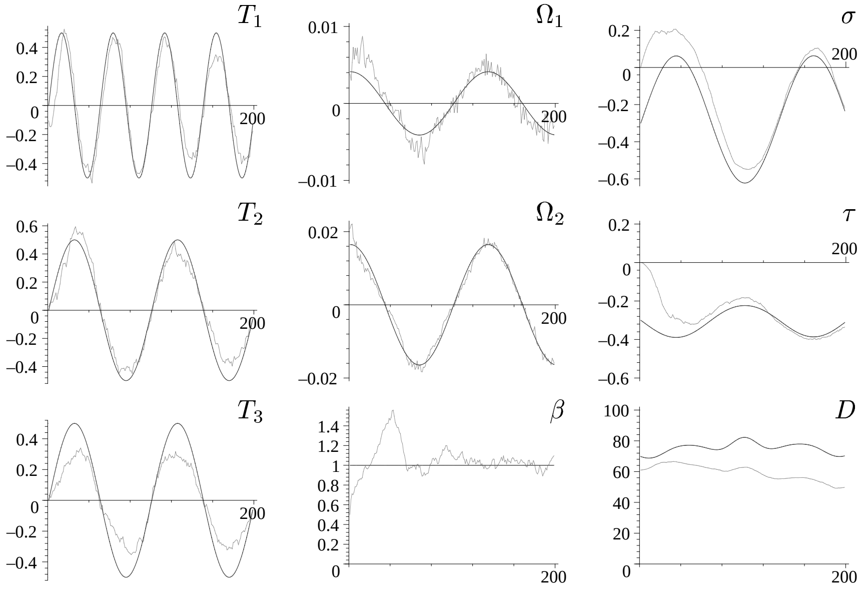

Figure 3: Motion, inverse focal length and structure estimates compared with the known groundtruth (dark lines) for the point data experiment.

[8]. The use of Gaussian windows was deliberate - they are closed under affine transformation and hence represent the evolution of the projected affine patches [17]. Fig. 4 shows the structure estimates from the filter overlaid on the corresponding frames, where the depths and surface normals are represented by perspective projections of oriented ‘platelets’ with needles indicating the normal direction. The top left image shows the initial state (normals pointing towards the camera and arbitrary depth values) and the central and right columns show the state of the platelets at frames further on in the sequence. The filter captures the surface structure well in each case, showing clearly the variation in depth and surface normal (this is better appreciated when viewing the sequences). Motion and inverse focal length estimates also converge quickly as shown in Fig. 5.

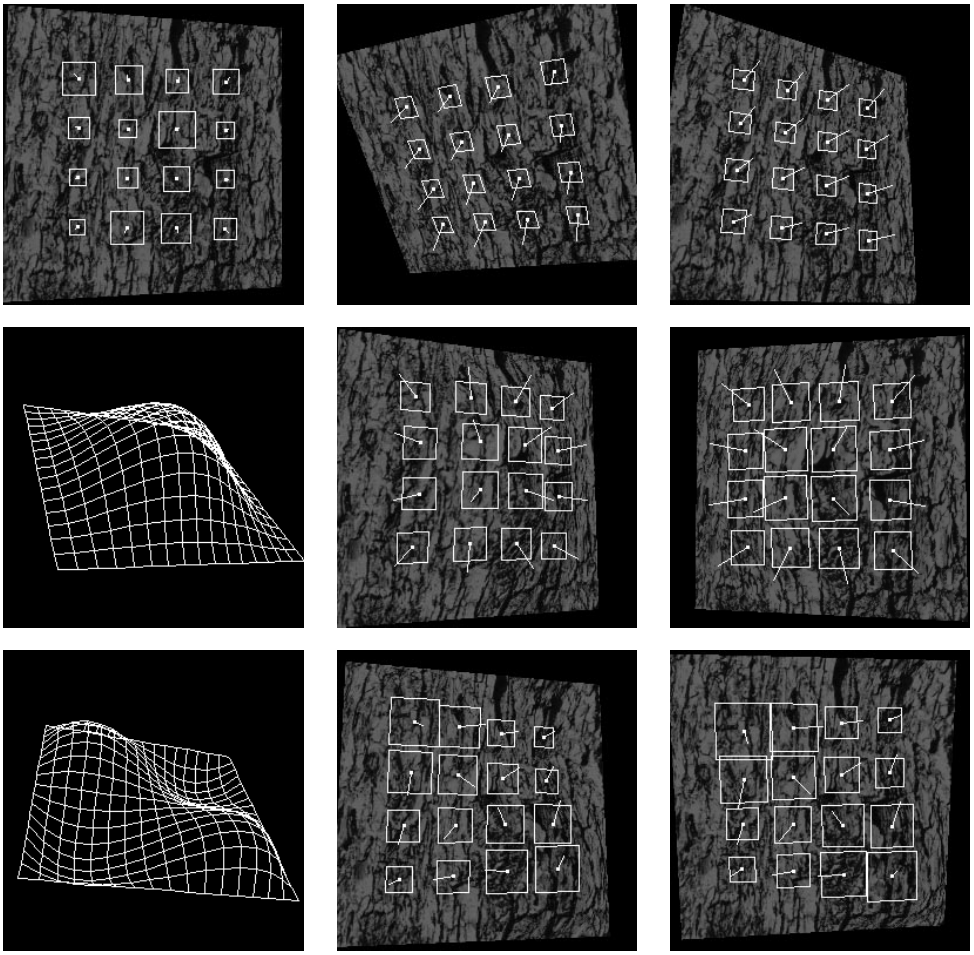

Figure 4: Texture mapped surfaces with overlaid ‘platelets’ indicating the estimated surface normals and depths.

4.3 Basket Sequence

The estimator was also tested on a real image sequence of an upturned basket being rotated on an office chair in front of a stationary video camera. Four frames from the sequence are shown in Fig. 7. Twenty regions were selected by hand and then tracked over the sequence as in the previous experiment. The computed affine flow parameters for one of the regions are shown in Fig. 6. Fig. 7 shows the initial structure state (top left) with normals pointing towards the camera and equal depths, and the evolved structure for frames further on in the sequence. In this experiment we iterated the filter three times (forwards, backwards and forwards through the sequence) to remove the initial transient estimates. Again, the filter clearly captures the surface structure. As shown in Fig. 8, the motion was predominantly cyclic translation in X and rotations about the Y and Z axes. Note also the convergence of the inverse focal length to a steady value of around 0.2.

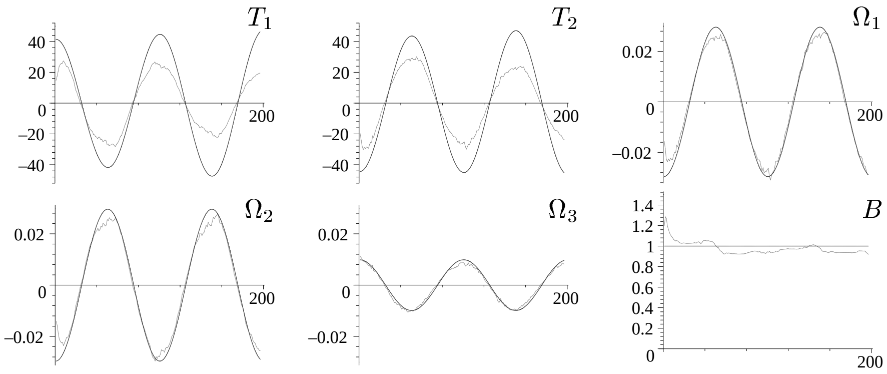

Figure 5: Motion and inverse focal length estimates compared with the known groundtruth (dark lines) for the texture mapped plane.

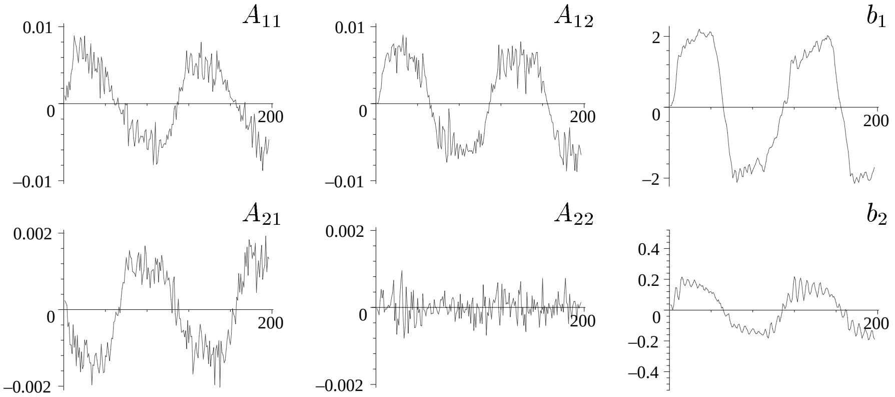

Figure 6: Affine flow parameters computed for one of the regions in the basket sequence

5 Conclusions

This work has demonstrated that affine flow parameters can be used successfully to recover 3-D structure and motion, giving the potential for a denser structure description via the surface normals than available from a feature based approach, whilst avoiding the high computational demands of an optical flow scheme. The method performed well when the affine model provided a good approximation of the local motion fields (particularly for near-view textured surfaces), even when the measurements contained significant amounts of noise. Moreover, its ability to recover stable estimates of the focal length was encouraging and appears to provide a robust method for obtaining metric estimates without the need for pre-calibration.

We are currently carrying out further investigation of the filter performance, particularly its convergence and accuracy properties given known (or estimated) uncertainty in the affine flow measurements. A key issue for its use in practice will clearly be an effective method for first identifying and then tracking appropriate sets of ‘affine patches’ corresponding to a single rigid surface in a sequence. A method for doing this was

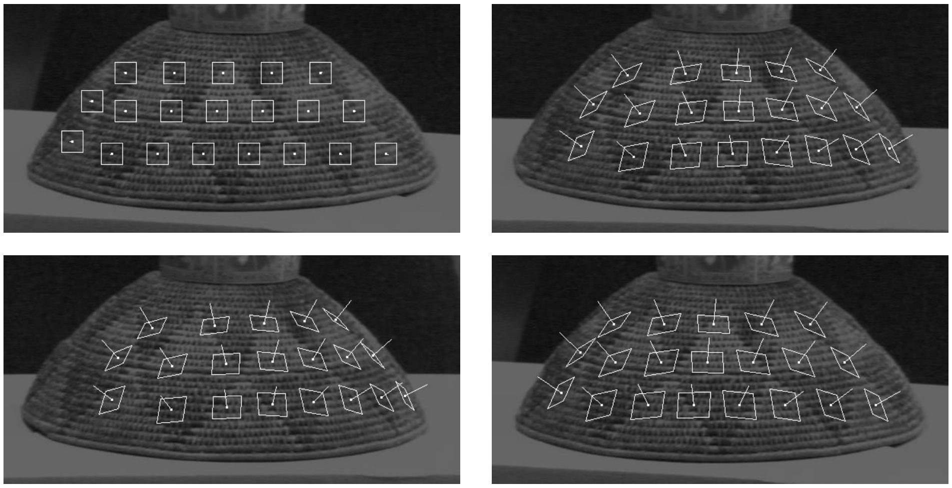

Figure 7: Frames and estimated platelets for the basket sequence.

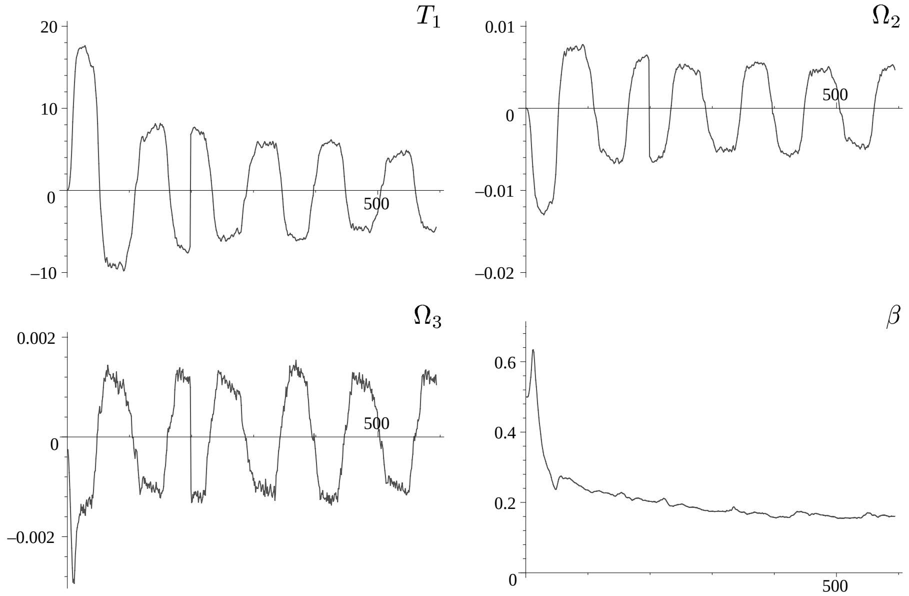

Figure 8: Motion and inverse focal length estimates obtained for the basket sequence.

described in [17] and we are now developing an integrated scheme which combines the tracking with the 3-D estimation. We are also looking at ways in which the method can be combined with a suitable feature based technique in order to provide a hybrid scheme which can adapt to the characteristics of the scene being viewed, recovering surface normal information when and if affine flow measurements become available.

Acknowledgement

This work was supported by the UK EPSRC. Many thanks to Stefan Krüger for discussions and help with generating and capturing the sequences.

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans on Patt Analysis and Machine Intell*, 7(4):384–401, 1985.

[2] A. Azarbayejani and A. P. Pentland. Recursive estimation of motion, structure and focal length. *IEEE Trans on Patt Analysis and Machine Intell*, 17(6):562–575, 1995.

[3] J. V. Candy. *Signal Processing: The Model Based Approach*. McGraw-Hill, Singapore, 1986.

[4] D. J. Heeger and A. D. Jepson. Subspace methods for recovering rigid motion i: algorithm and implementation. *Int Journal of Computer Vision*, 7(2):95–117, 1992.

[5] T. S. Huang and A. Netravali. Motion and structure from feature correspondences: a review. *Proc IEEE*, 82(2):252–268, 1994.

[6] A. Jebara, T Azarbayejani and A. P. Pentland. 3d structure from 2d motion. *IEEE Signal Processing Magazine*, 16(3):66–84, 1999.

[7] H. C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proc Royal Society, London*, B-208:385–397, 1980.

[8] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc Int Joint Conf on Artificial Intelligence*, pages 674–679, 1981.

[9] F. G. Meyer. Time-to-collision from first-order models of the motion field. *IEEE Trans on Robotics and Automation*, 10(6):792–798, 1994.

[10] F. G. Meyer and P. Bouthemy. Region-based tracking using affine motion models in long range sequences. *CVGIP: Image Understanding*, 60(2):119–140, 1994.

[11] R. Mohr and B. Triggs. Projective geometry for image analysis. Tutorial, International Symposium of Photogrammetry and Remote Sensing, Vienna, 1996.

[12] D. W. Murray and L. S. Shapiro. Dynamic updating of planar structure and motion: the case of constant motion. *Computer Vision and Image Understanding*, 63(1):169–181, 1996.

[13] S. Negahdaripour and S. Lee. Motion recovery from image sequences using only 1st order optical-flow information. *Int Journal of Computer Vision*, 9(3):163–184, 1992.

[14] R. Pollefeys, M Koch and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *Int Journal of Computer Vision*, 32(1):7–25, 1999.

[15] M. Subbarao and A. M. Waxman. Closed form solutions to image flow equations for planar surfaces in motion. *CVGIP*, 36:208–228, 1986.

[16] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Trans on Image Processing*, 3(5):625–638, 1994.

[17] R. Wilson, P. Meulemans, A. Calway, and S. Kruger. Image sequence analysis and segmentation using g-blobs. In *Proc IEEE Int Conf on Image Processing*, pages 483–487, 1998.

[18] Y. Xiong and S. A. Schafer. Dense structure from a dense optical flow sequence. *Computer Vision and Image Understanding*, 69(2):222–245, 1998.