Reactive Zoom Control while Tracking Using an Affine Camera

Ben Tordoff and David Murray
Department of Engineering Science, University of Oxford
Parks Road, Oxford OX1 3PJ, UK

Abstract

This paper describes a method for visual control of the zoom setting of an active camera during tracking. The method assumes an affine projection, and tracking is achieved using affine transfer, a process which is fundamentally invariant to zoom. However, the form of the projection matrices determined to achieve transfer allows the relative scale between the affine bases in different views to be determined, and hence controlled at unity. Unlike its projective equivalent, the method requires no self-calibration, the zooming camera may move quite generally, and no restriction need be applied to the nature of the scene being viewed.

The performances of 3D and 2D versions of the algorithms are examined using synthetic data under varying levels of noise and under varying degrees of degeneracy in motion and structure. Real-time results are presented from imagery of laboratory scenes and off-line results obtained from an outdoor surveillance video sequence.

Keywords: active vision, zoom control, affine transfer tracking, affine structure.

1 Introduction

The motions of a camera-lens combination used to track objects under surveillance, or to follow the action in a televised sporting event, are typified by substantial rotations, small or zero translations, and zooming in and out. By comparison with the considerable attention lavished on automated tracking of the scene motion, automatic zoom control is a rather unexplored area. However, it is an area which progress over the last few years in the theory of structure from motion and self-calibration of cameras lays open to practical investigation.

Under human operator control, camera zooming is initiated in two ways, as illustrated in figure 1. The first we call purposeful zooming, where some higher level process indicates that it would be valuable either to zoom in to collect more object detail, or to zoom out to obtain surrounding context. The other way zooming is used is more reactive. In this case, the camera operator adjusts the zoom to preserve the “image size” of the target object as it moves away from or towards the camera. All of us will have watched broadcasts where skilled camera-work has enhanced the information flow to the viewer (and vice versa): there seems every reason to suppose that a similar autonomous capability will be of benefit to a computer vision system. This paper explores methods of achieving the second, reactive, type of zoom control under affine viewing conditions appropriate for most surveillance applications.

But first image size must be considered more carefully. Were the camera observing a spinning disc with rotation axis a constant distance from the camera, it would be quite inappropriate for the zoom lens to oscillate in and out. Under perspective projection the

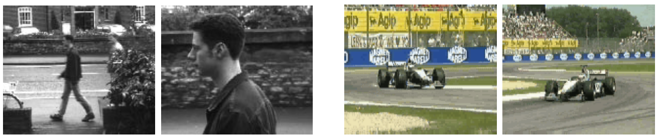

Figure 1: Purposeful zooming (left), compared with reactive zooming (right). In the former zooming depends on higher intent – the recognition of a passer-by; in the latter the scene is effectively in charge – the cameraman maintains the apparent size of the object.

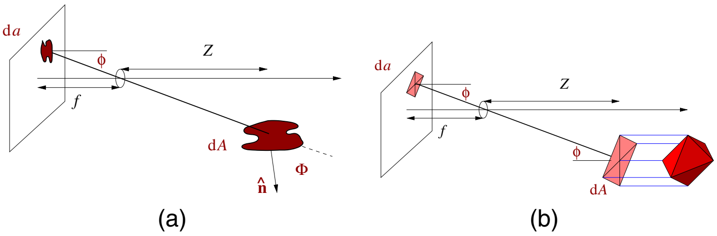

Figure 2: (a) Imaging geometry for the perspective case, used to argue for the preservation of f/Z. (b) Weak perspective projection modelled as projection onto a fronto-parallel plane before scaling.

simplest sketch of a fixated plane, such as the disc, is enough to convince one that the ratio of focal length f to depth Z of the fixation point should be preserved by changing f when Z changes. A slightly more elaborate explanation for the case of a non fixating camera (figure 2a) is that, rather than preserving image area, one should preserve the ratio ρ of image area to scene area projected along the ray direction. The solid angles subtended by the image and scene patches at the optic centre are equal and so it is

$$\rho = \frac{\mathrm{d}a\ \cos\phi}{\mathrm{d}A\ \cos\Phi} = \left(\frac{f}{Z}\right)^2$$

that should be preserved, where ϕ, Φ are the angles between the ray and the image and surface normals, and Z is the depth of the patch.

In earlier work [12] we applied this constraint to perspective cameras, where the camera must be calibrated. Self-calibration methods for special cases of motion and structure are of help here. We utilized that of Agapito *et al* [1] requiring the camera motion to be purely rotational (common for wall-mounted cameras). If the object is planar, the method of Malis and Cipolla [7] allows calibration from the target itself. However, maintaining the calibration over extended periods is burdensome. In our case [12], lack of sufficient image data for calibration while tracking often necessitate use of a calibration look-up table refreshed by self-calibration during "idle" moments.

In most visual tracking, object relief is small compared to the object's distance from the camera and the image projection can be approximated as affine. The burden of calibration is eased and the zooming camera may move quite generally while viewing an arbitrary scene. In this case, f/Z is again the appropriate measure to be preserved, with Z interpreted as the depth of the fronto-parallel plane onto which rays are projected orthogonally before being projected perspectively (fig. 2b). For the general affine camera this measure generalises to the scaling of the affine bases, as is made clear later.

2 Review: Tracking using Affine transfer

Tracking in our work is achieved using affine transfer [9, 2], a method which takes advantage of the viewpoint invariance of single image features and the collective temporal coherence of a cloud of such features, without requiring features to exist through entire sequences. The method is fundamentally invariant to zoom [4, 5], and thus independent of errors in zoom and whether the zoom control is reactive or purposive. Furthermore, transfer also provides tolerance to features appearing at and disappearing from the edge of the image as a wider or narrow view is taken. Overlaying the zoom-variant control process on top of a zoom-invariant tracking competence seems attractive from an architectural standpoint. The optimisation with respect to the Frobenius norm required to achieve transfer with more than the minimal point set was shown in [9] to be identical to Tomasi and Kanade's factorisation method [11], and so it is convenient to use the latter's standard formulation.

An unregistered image point p in frame i, $\mathbf{x}_{ip}$, is projected from the homogeneous scene point $\mathbf{X}_p$ as $\mathbf{x}_{ip} = \mathtt{M}_i\mathbf{X}_p + \mathbf{t}_i$, and registered points formed by subtracting the centroid $\mathbf{x}_{ip} \leftarrow \mathbf{x}_{ip} - \mathbf{c}_i$ where $\mathbf{c}_i = \sum \mathbf{x}_{ip}/P$. From the P point correspondences established over I frames, affine structure and motion is recovered in batch mode by singular value decomposition of the $2I \times P$ registered measurement matrix

$$\mathtt{U\Sigma V}^\top \leftarrow \mathtt{W} = \begin{pmatrix} \mathbf{x}_{1,1} & \cdot\cdot & \mathbf{x}_{1,P} \\ : & & : \\ \mathbf{x}_{I,1} & \cdot\cdot & \mathbf{x}_{I,P} \end{pmatrix}$$

and by imposing its rank-3 property in zero-noise to find the optimal affine projection matrices and structure from the ordered columns of $\mathtt{U}$ and $\mathtt{V}$

$$\begin{pmatrix} \mathtt{M}_1 \\ : \\ \mathtt{M}_I \end{pmatrix} = (\sigma_1\mathbf{u}_1 \quad \sigma_2\mathbf{u}_2 \quad \sigma_3\mathbf{u}_3) \ ; \qquad \begin{pmatrix} \mathbf{X}_1{}^\top \\ : \\ \mathbf{X}_P{}^\top \end{pmatrix} = (\mathbf{v}_1 \quad \mathbf{v}_2 \quad \mathbf{v}_3) \ .$$

For affine transfer, given registered gaze points $\mathbf{g}_i$ in the first $i = 1, \ldots, I-1$ frames, the 3D position of the gaze point $\mathbf{G}$ is found using the pseudo-inverse from

$$\mathbf{G} = \begin{pmatrix} \mathtt{M}_1 \\ : \\ \mathtt{M}_{I-1} \end{pmatrix}^+ \begin{pmatrix} \mathbf{g}_1 \\ : \\ \mathbf{g}_{I-1} \end{pmatrix}$$

and the gaze point transferred to the latest frame F using $\mathbf{g}_I = \mathtt{M}_I\mathbf{G}$. In practice we use the minimum $I = 3$ frames.

3 Recovering relative scale

3.1 Scale from Euclidean constraints in three views

In inhomogeneous coordinates using registered image points and assuming no pixel skew, one sound affine projection of a Euclidean structure is $\mathbf{x} = \mathtt{M}_E\mathbf{X}_E$ with

$$\mathtt{M}_E = S \begin{pmatrix} 1 & 0 \\ 0 & 1/\alpha \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \end{pmatrix} \ ,$$

where α is the aspect ratio and the R's are elements of the 3×3 rotation matrix between the world frame and the camera frame. For a weak perspective camera, $S = f/\bar{Z}$, where $\bar{Z}$ is the mean depth of the points in the camera frame.

The affine structure is related to the Euclidean structure by an affine transformation, $\mathtt{H}\mathbf{X}_E = \mathbf{X}$, and similarly each $\mathtt{M}$ is related to its Euclidean counterpart by $\mathtt{MH} = \mathtt{M}_E$. But the form of $\mathtt{M}_E$ ensures that for the i-th frame [11]

$$\mathtt{M}_i\mathtt{HH}^\top\mathtt{M}_i{}^\top = \begin{pmatrix} p & q & r \\ s & t & u \end{pmatrix}_i \begin{pmatrix} h_1 & h_2 & h_3 \\ h_2 & h_4 & h_5 \\ h_3 & h_5 & h_6 \end{pmatrix} \begin{pmatrix} p & s \\ q & t \\ r & u \end{pmatrix}_i = S_i^2 \begin{pmatrix} 1 & 0 \\ 0 & 1/\alpha^2 \end{pmatrix} . \qquad (1)$$

Matrix $\mathtt{HH}^\top$ is symmetric and hence has 6 degrees of freedom. Assuming we know the aspect ratio, for I frames there are I values of S_i but only $(I-1)$ d.o.f., as the overall scale is unknown. For each frame the above system gives three equations linear in the parameters, and so $3I \geq 6 + I - 1$ or $I \geq 3$ to solve. The minimum, $I = 3$, is conveniently also the minimum required to achieve 3D affine transfer for tracking. Note too that as only the motion given by the projection matrices $\mathtt{M}$ is used, one can use the cheaper eigen-decomposition of $\mathtt{U\Sigma}^2\mathtt{U}^\top \leftarrow \mathtt{WW}^\top$ rather than the SVD of $\mathtt{W}$ to recover scale. Again this computation is all that is required to maintain fixation [9].

To recover both relative scales and the transformation $\mathtt{HH}^\top$, we set $S_1 = 1$, construct the matrix

$$\mathtt{D} = \begin{pmatrix} p_1^2 & 2p_1q_1 & 2p_1r_1 & q_1^2 & 2q_1r_1 & r_1^2 & 0 & 0 \\ p_1s_1 & p_1t_1 + q_1s_1 & p_1u_1 + r_1s_1 & q_1t_1 & q_1u_1 + r_1t_1 & r_1u_1 & 0 & 0 \\ s_1^2 & 2s_1t_1 & 2s_1u_1 & t_1^2 & 2t_1u_1 & u_1^2 & 0 & 0 \\ p_2^2 & 2p_2q_2 & 2p_2r_2 & q_2^2 & 2q_2r_2 & r_2^2 & -1 & 0 \\ p_2s_2 & p_2t_2 + q_2s_2 & p_2u_2 + r_2s_2 & q_2t_2 & q_2u_2 + r_2t_2 & r_2u_2 & 0 & 0 \\ s_2^2 & 2s_2t_2 & 2s_2u_2 & t_2^2 & 2t_2u_2 & u_2^2 & -\frac{1}{\alpha^2} & 0 \\ p_3^2 & 2p_3q_3 & 2p_3r_3 & q_3^2 & 2q_3r_3 & r_3^2 & 0 & -1 \\ p_3s_3 & p_3t_3 + q_3s_3 & p_3u_3 + r_3s_3 & q_3t_3 & q_3u_3 + r_3t_3 & r_3u_3 & 0 & 0 \\ s_3^2 & 2s_3t_3 & 2s_3u_3 & t_3^2 & 2t_3u_3 & u_3^2 & 0 & -\frac{1}{\alpha^2} \end{pmatrix} ,$$

and solve the following equation for $S_{2,3}$ and $\mathbf{h} = (h_1, \ldots, h_6)^\top$ using SVD:

$$\mathtt{D} \begin{pmatrix} \mathbf{h} \\ S_2^2 \\ S_3^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \frac{1}{\alpha^2} \\ \mathbf{0}_6 \end{pmatrix} .$$

The transformation $\mathtt{H}$ can be recovered from $\mathtt{HH}^\top$ by Cholesky decomposition, although this is unnecessary for zoom control.

3.2 Scale from epipolar geometry and two views

Although transfer requires three camera matrices, it is possible to recover structure and scale from just two [6, 10]. Here, matrices $\mathtt{M}_2$ and $\mathtt{M}_3$ are the obvious choices. Shapiro *et al* describe how the spacing of the parallel epipolar lines under affine projection are characteristic of the scale, at least once the images are corrected for known aspect ratio [10]. From the entries in the affine fundamental matrix

$$\mathbf{x}_3^\top \mathtt{F}_{23}^A \mathbf{x}_2 = \mathbf{x}_3^\top \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{pmatrix} \mathbf{x}_2 \ ,$$

where $e = 0$ if the image points are registered, the relative scale is recovered as

$$S_3/S_2 = \sqrt{(c^2 + \alpha^2 d^2)/(a^2 + \alpha^2 b^2)}$$

where α is again the aspect ratio[1]. The elements of the fundamental matrix can be found directly from the camera matrices as determinants of four 3×3 minors involving rows 3 to 6 of the joint projection matrix [8] to the left of the '|' in

$$\left(\begin{array}{c|c} \mathtt{M}_1 & \mathbf{x}_1 \\ \mathtt{M}_2 & \mathbf{x}_2 \\ \mathtt{M}_3 & \mathbf{x}_3 \end{array} \right) \begin{pmatrix} \mathbf{X} \\ -\lambda \end{pmatrix} = \mathbf{0}_6 \ .$$

These determinants describe the epipoles $\mathbf{e}_{23} = (d, -c, 0)^\top$ and $\mathbf{e}_{32} = (b, -a, 0)^\top$, where the ratio of coefficients in x and y directions gives the direction of epipolar lines, and the magnitude of the epipole their (relative) spacing. The scale change from, for example, image 2 to image 3 is expressed in terms of the rows $\mathbf{m}_i^j$ of the projection matrices $\mathtt{M}_i$

$$\frac{S_3}{S_2} = \frac{\|\mathbf{e}_{23}\|}{\|\mathbf{e}_{32}\|} \qquad e_{23}^j \overset{\text{def}}{=} \begin{cases} \det \begin{pmatrix} \mathbf{m}_2^1 \\ \mathbf{m}_2^2 \\ \mathbf{m}_3^j \end{pmatrix}, & j \in \{1, 2\} \\ 0, & j = 3 \end{cases}$$

3.3 Scale from image-based methods

Although we have earlier argued against image-based methods, it is nonetheless worthwhile exploring how such approximations perform.

One approximation is to suppose that as the 3D affine transformation $\mathtt{H}$ is the same for all three views in a batch, its contribution to scale cancels out between images. That is, taking ratios of the determinants of the matrices in Equation (1)

$$\frac{\det(\mathtt{M}_i \mathtt{H} \mathtt{H}^\top \mathtt{M}_i^\top)}{\det(\mathtt{M}_j \mathtt{H} \mathtt{H}^\top \mathtt{M}_j^\top)} = \left(\frac{S_i}{S_j} \right)^4 \approx \frac{\det(\mathtt{M}_i \mathtt{M}_i^\top)}{\det(\mathtt{M}_j \mathtt{M}_j^\top)} \ .$$

The determinant of $\mathtt{M}_i \mathtt{M}_i^\top$ gives the square of the area spanned by the rows of $\mathtt{M}_i$. The approximation is equivelent to scene-based scale recovery when $\mathtt{H}$ is close to identity or $\mathtt{M}_i$ and $\mathtt{M}_j$ are related only by an image (z-axis) rotation.

As the introductory example of the spinning disc showed, there are situations where using relative area — and hence the ratio of determinants — is less than ideal. Better would be to approximate the change in a single dimension of the object, ideally along the projection of the rotation axis in the image.

The individual singular values of $\mathtt{M}_i \mathtt{M}_i^\top$ describe the dimension of the affine bases in the maximum direction and a direction perpendicular to it. The 2-norm of a matrix is equal to its largest singular value, and so the squared change in maximum dimension is given by the ratio of 2-norms

$$\left(\frac{S_i}{S_j} \right)_{\text{size}} \approx \left(\frac{\|\mathtt{M}_i \mathtt{M}_i^\top\|_2}{\|\mathtt{M}_j \mathtt{M}_j^\top\|_2} \right)^{1/2} \ .$$

[1] The expression in [10] omits explicit mention of the aspect ratio.

This distance measure obviously changes with rotation unless the axis of rotation happens to coincide with the major affine basis. However, as only one dimension is considered this method can cope with planar objects turning side-on to the camera, the situation which causes problems for all the previously described methods.

4 Experiments using Synthetic Data

In the experiments using synthetic data, a cluster of 20 scene points is distributed randomly within a cube and moved by incremental random rotations and to random depths before projection using weak perspective onto the image plane. Gaussian noise was added with standard deviation equal to some percentage of the spread of points in the image. (This recipe simulates a constant uncertainty in pixels were the camera to be re-zoomed in between frames.)

Figure 3a shows the recovered scale factor S_3/S_1 plotted against the inverse ratio of mean depths $\bar{Z}_1/\bar{Z}_3$. Points should lie on the unit slope line, and from this graph all methods appear to perform well over a large range of relative scale changes.

The relative robustness of the various methods as the level of noise is increased is illustrated in figure 3b. Here we take the ratio of recovered scale to actual scale for each of a 1000 trials at every noise value, and determine the mean μ and standard deviation σ of the error when compared to $\bar{Z}_1/\bar{Z}_3$.

If the inter-frame rotation becomes small, or the object has little depth variation, then the rank of the measurement matrix, $\mathrm{rank}(\mathtt{W})$, drops to two. Figure 3c shows the case of the object rotation decreasing to zero with image noise fixed. As one would expect, purely image-based measures are most accurate with little rotation, and the scene-based methods work equally well across the full range. Note that the errors seen for the image-based methods are not random but correspond directly with changes in the object orientation — for instance a spinning object gives regular oscillations.

The final test of figure 3d has fixed noise and rotation but the 3D point structure is reduced from being a cube to being a plane. These results, and examination of the singular values, suggest that effective planarity is achieved when the third dimension is less than 10% of the other two. For thicknesses below 10% the scene-based methods degrade rapidly giving unpredictable (σ large) and biased ($\mu \neq 0$) results. As expected, the 2-norm method is little affected as only one object dimension is required. We now examine this situation in detail by considering purely planar objects.

5 The planar case

5.1 Scene-based methods

Without loss of generality, the scene points are assumed to lie on a plane $\mathbf{X}_E = (X, Y, 0)$. We write the rotation into the camera frame as $\mathtt{R}$ and recover planar affine motion [9] using Tomasi and Kanade. The planar transfer process is analogous to the 3D method, but we cannot complete the analogy for scale recovery. We reach the point

$$\mathtt{M}_i\mathtt{H}\mathtt{H}^\top\mathtt{M}_i^\top = S_i^2 \begin{pmatrix} 1 & 0 \\ 0 & 1/\alpha \end{pmatrix} \mathtt{N}_i \begin{pmatrix} 1 & 0 \\ 0 & 1/\alpha \end{pmatrix}$$

but, unlike the 3D case, matrix $\mathtt{N}_i$ is not the product of a rotation with its transpose and hence not an identity matrix.

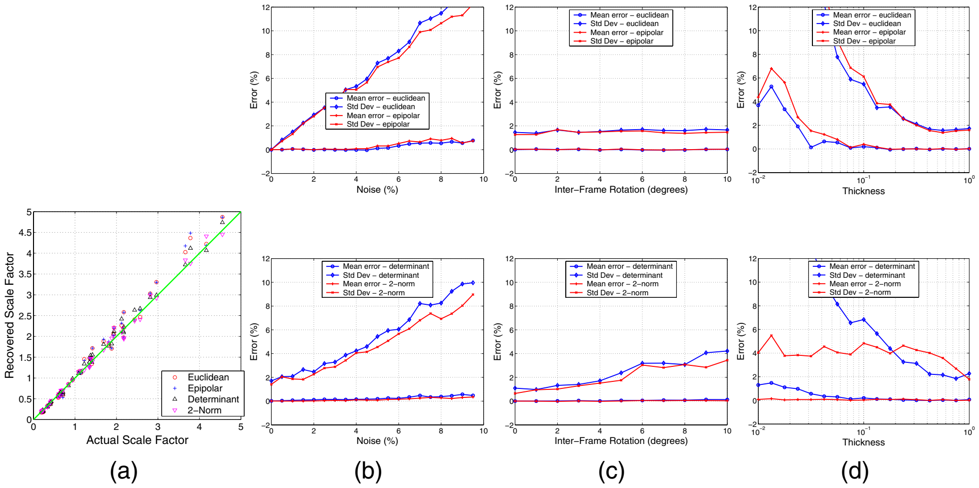

Figure 3: (a) A scatter plot of recovered scale factor vs actual scale factor for 1% image noise. (b) The mean value μ and standard deviation of the error in the ratio of recovered to actual scale for 1000 tests of each scene (top) and image (bottom) method as a function of increasing noise. (c) As (b) but with fixed noise of 1% and as the inter-frame rotation is decreased. (d) As (b) but with fixed noise of 1%, rotation of 5° per frame and as the scene is reduced from a cube (unity thickness) to a plane (zero thickness).

One can analyse $\mathtt{N}$ directly using the elements of the 3×3 rotation matrix, but an alternative is to parameterise the rotation as a success of rotations by angles γ around the z-axis, β around the y-axis, and δ around the x-axis. Then the matrix is

$$\mathtt{N}_i = \begin{pmatrix} \cos^2\beta & -\sin\delta\sin\beta\cos\beta \\ -\sin\delta\sin\beta\cos\beta & \sin^2\delta\sin^2\beta + \cos^2\delta \end{pmatrix}_i ,$$

actually independent of γ. Although there are range constraints on the various elements, the only[2] hard constraint is that the matrix's larger eigenvalue $\lambda^U(\mathtt{N}_i)$ is unity. It is no longer possible to recover the transformation $\mathtt{H}$.

We fare no better with the epipolar method, as the 3×3 minors of the joint projection matrix now each have zero determinant. Although a fundamental matrix can be calculated between views, the magnitude of the elements no longer gives the scale as in [10].

The reason both scene-based methods fail is that image matches alone are no longer sufficient to constrain the motion. For instance, it is possible for a rotation about the image y-axis plus scaling to give the same point motions as a rotation about the image x-axis plus an opposite scaling . Unless either object structure or motion is known we must resort to image-based measures.

5.2 Image-based methods

For a truly planar object it is clear that as the object turns side-on to the camera its area drops to zero, and the determinant of $\mathtt{MM}^\top$ is an unsuitable measure (as was seen in figure

[2] The three different elements of the symmetric matrix are given by two angle parameters.

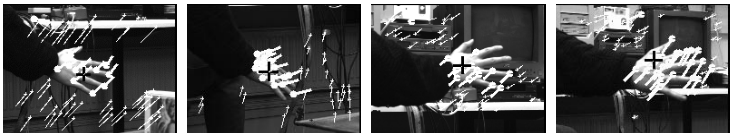

Figure 4: Corner features detected, individually tracked, segmented and then collectively fixated during zooming.

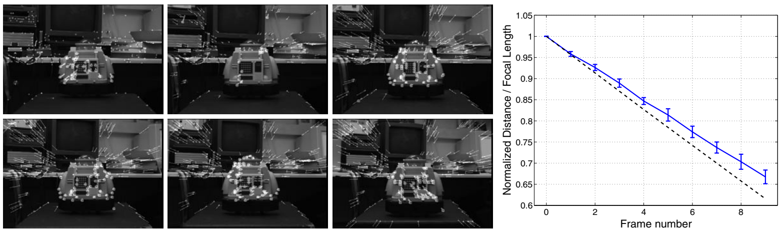

Figure 5: The toy robot moves towards the camera, but the camera zooms out to maintain f/Z. The background can be seen shrinking away from the foreground subject. The graph shows the change in normalised focal length and depth, the latter measured by hand.

3d)). The 2-norm method however does remain valid and stable, because when the projection of plane degenerates into a line the largest singular value reflects the line's length. Moreover, as only the largest singular value is taken into account in both 2D and 3D 2-norm methods, there is no difference between taking the 2-norm of degenerate 3D affine projection matrices and taking the 2-norm of 2D projection matrices. The conditions under which the scale recovered in this way accurately reflects the scene-based scale are the same as those for the 3D case (z-axis rotation plus scaling, or $\mathtt{HH}^\top$ equal to identity).

As a final point we note that tracking using affine-transfer does not degrade when the motion or object is planar, and there is little advantage to switching to planar affine-transfer [9].

6 Results from real imagery

The synthetic results above indicate that image-based measures give improved robustness to large noise, but will deviate from true f/Z tracking when the target turns. Ideally we would like to be able to switch to a scene-based method during turns, and use image-based methods during planar motion. At frame rate (25Hz) where scene motion between frames will be small, the motion is likely to appear planar except during swift turns. Detecting structural planarity is important and requires use of the 2-norm method, but may be difficult as both structural and motive planarity cause the third singular value to be small.

The methods described have been incorporated into a real-time affine-transfer based tracking system, enabling study of results both on- and off-line, and recovering scale with optional feedback to the zoom-lens.

As input $\mathbf{x}_i$ into the affine transfer algorithm we use corners features detected [3] and tracked individually over time, and segmented between foreground and background using a modified version of the MLESAC algorithm [13]. Figure 4 shows video-rate feature recovery, segmentation, and affine transfer tracking of a waving hand. The camera is

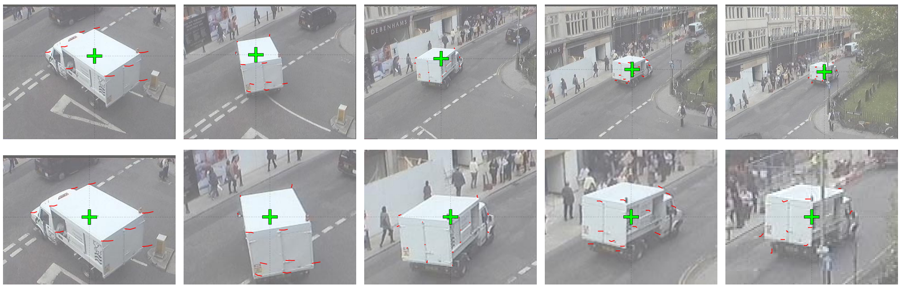

Figure 6: Top row: stills from a surveillance sequence during which the actual zoom setting was fixed. Corner features were detected, tracked and segmented, and the foreground points fed to the 3D affine transfer and scale recovery algorithm. The cross marks the fixation point. Along the lower row are images which are digitally zoomed using the scale factor recovered.

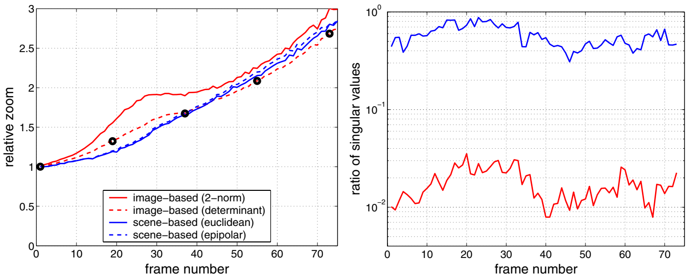

Figure 7: Left: the recovered scale varying over time, with circles marking the stills shown above. Right: the logarithm of the 2nd and 3rd singular values as a function of time, indicating that most of the motion might as well be regarded as planar. The truck is turning between frames 10 and 40, but the third singular value does not rise enough to switch methods.

zooming during this, but arbitrarily and not under control.

In contrast the sequence in figure 5 shows the camera lens automatically zooming out to compensate as the toy robot approaches the camera. The motion was repeated 10 times, and the error bars show one standard deviation either side of the mean for each position.

The last reported experiment in figure 6 shows the results from zoom control as a truck is tracked approaching and leaving a road junction. The top row of figures shows a number of stills from a sequence as captured from a non-zooming surveillance camera, with the fixation point shown.

Figure 7(left) shows the relative scale factor recovered from the imagery, and the lower row of figure 6 shows images digitally zoomed by the relative scale from the determinant method. The right graph of figure 7 shows the 2nd and 3rd singular values of the registered measurement matrix, indicating that motion is effectively planar throughout the sequence. When the van is turning, scale recovery should be via either of the scene-based methods, but detecting this case is not possible from the third singular value. If more frames were used in the affine-transfer the rotation would be easier to detect, but at the expense of reducing the number of features existing in all frames.

7 Discussion and conclusions

In this paper we suggest that automatic zooming may be as valuable as automated tracking in enhancing the visual information content of video supplied to a computer vision system. We have developed the geometric constraint which compensates for depth motion by zooming, a constraint which is applicable to both perspective and affine projections.

We have described algorithms for 3D and 2D automatic zoom control under affine projection, and have demonstrated their performance using synthetic data and using real imagery. A real time implementation of the entire process has been demonstrated using a laboratory bound active stereo rig, and the applicability to surveillance footage shown using off-line processing, post-capture.

Use of the affine approximation is justifiable in most surveillance and broadcasting environments, and it allows recovery of relative scale without knowledge of camera position or internal calibration. The invariance of the affine-transfer method to scale changes means that the tracking and zooming competences can be treated as independent, even though the zoom control part uses information from the tracking part.

In the strictly affine planar case there is not enough information to correctly determine scale, but we have shown that a simple approximation still permits reliable scale recovery. However, for real-time applications the short time between frame capture often makes the object motion appear planar, making detection of structural planarity difficult.

References

[1] L. de Agapito, E. Hayman, and I. Reid. Self-calibration of a rotating camera with varying intrinsic parameters. In *Proc 9th British Machine Vision Conf, Southampton*, pages 105–114, 1998.

[2] S. M. Fairley, I. D. Reid, and D. W. Murray. Transfer of fixation using affine structure: extending the analysis to stereo. *Int. Journal of Computer Vision*, 29(1):47–58, 1998.

[3] C. G. Harris and M. Stephens. A combined corner and edge detector. In *Proc 4th Alvey Vision Conf, Manchester*, pages 147–151, 1988.

[4] E. Hayman, I.D. Reid, and D.W. Murray. Zooming while tracking using affine transfer. In *Proc 7th British Machine Vision Conf, Edinburgh*, 1996.

[5] E. Hayman, T. Thórhallsson, and D. Murray. Zoom-invariant tracking using points and lines in affine views: An application of the affine multifocal tensors. In *Proc 7th Int Conf on Computer Vision, Corfu*, pages 269–277, 1999.

[6] J. Koenderink and A. J. van Doorn. Affine structure from motion. *Optical Society of America - Annals*, 8(2):377–385, 1991.

[7] E. Malis and R. Cipolla. Multi-view constraints between collineations: application to self-calibration from unknown planar structures. In *eccv2000*, volume 2, pages 610–624, June 2000.

[8] L. Quan and T. Kanade. Affine structure from line correspondences with uncalibrated affine cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):834–845, 1997.

[9] I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *Int. Journal of Computer Vision*, 18(1):41–60, April 1996.

[10] L.S. Shapiro, A. Zisserman, and M. Brady. 3d motion recovery via affine epipolar geometry. *Int. Journal of Computer Vision*, 16:147–182, 1995.

[11] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. Journal of Computer Vision*, 9(2):137–154, 1992.

[12] B. Tordoff and D.W. Murray. Reactive zoom control while tracking. Technical Report OUEL2228/00, Department of Engineering Science, Oxford University, 2000.

[13] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78:138–156, 2000.