Reactive Control of Zoom while Fixating Using Perspective and Affine Cameras

Ben Tordoff and David Murray, Member, IEEE

Abstract—This paper describes reactive visual methods of controlling the zoom setting of the lens of an active camera while fixating upon an object. The first method assumes a perspective projection and adjusts zoom to preserve the ratio of focal length to scene depth. The active camera is constrained solely to rotate, permitting self-calibration from the image motion of points on the static background. A planar structure from motion algorithm is used to recover the depth of the foreground. The foreground-background segmentation exploits the properties of the two different interimage homographies which are observed. The fixation point is updated by transfer via the observed planar structure. The planar method is shown to work on real imagery, but results from simulated data suggest that its extension to general 3D structure is problematical under realistic viewing and noise regimes. The second method assumes an affine projection. It requires no self-calibration and the zooming camera may move generally. Fixation is again updated using transfer, but now via the affine structure recovered by factorization. Analysis of the projection matrices allows the relative scale of the affine bases in different views to be found in a number of ways and, hence, controlled to unity. The various ways are compared and the best used on real imagery captured from an active camera fitted with a controllable zoom lens in both look-move and continuous operation.

Index Terms—Active vision, zoom control, fixation, tracking, self-calibration, perspective projection, affine projection.

1 INTRODUCTION AND MOTIVATION

The motions of a camera and lens combination used to follow objects under surveillance, or to keep up with the action in televised broadcasts such as sporting events, are typified by substantial rotations, small or zero translations, and zooming in and out. By comparison with the considerable attention lavished on automated fixation of the scene motion, automatic zoom control is a rather unexplored area. It is one, however, which progress over the last few years in the theory of structure from motion and self-calibration of cameras lays open to practical investigation. All will have watched broadcasts where skilled camera-work enhances the information flow to the viewer, and a similar autonomous capability should be of benefit to a computer vision system.

Under the control of a human operator, zooming is initiated in the two broad ways suggested by Fig. 1. The first one might call purposeful zooming, where some higher-level process indicates that it would be valuable either to zoom in to collect more object detail, say for recognition, or to zoom out to obtain contextual information. Aspects of high-level camera management have been studied by, for example, Bobick and Pinhanez [1], and the possibilities for fusion of vision and graphics in automated video production were more recently surveyed by Gleicher et al. [2]. The other way zooming is used is more reactive. As suggested in Fig. 1b, the response of concern in this paper is the recovery of scale to preserve the “image size” of a target as it translates and twists in front of a fixating camera.

Although scale is frequently mentioned in the context of recovering structure from motion, rather little attention has been paid to its use in active sensor management. Though there are some examples of using other reactive responses. As one of several modules in a gaze control system, Viéville et al. [3] proposed balancing a pressure to reduce the image size of the target with a pressure to decrease the average retinal disparity between features in the stereo images. With too small an image, many features would be unmatched, increasing the average disparity. Unfortunately, their paper does not show results from this binocular approach. As discussed below, Fayman et al. [4] have suggested using zoom to remove looming motion. More recently, the present authors [5] have discussed zoom control based on kinematic uncertainty, zooming in if the scene motion is well-predicted and out if not.

Here, though, we explore methods of achieving reactive size-preserving zoom control under two imaging regimes: perspective and affine. The paper reports results from look-move and video-rate implementations of each, but also explores their limitations using synthetic data.

First, the meaning of image size must be considered more carefully. Were a camera observing a disc spinning about a rotation axis fixed at some constant distance from the camera, it would be quite inappropriate for its zoom lens to oscillate in and out. Evidently, using a naïve metric like image area is inappropriate, and invariance to rotation about axes fixed relative to the camera is desirable. Referring to Fig. 2a, under perspective projection, it is appropriate to preserve the ratio ρ of image area to scene area projected along the ray direction. The solid angles
The work of Fayman et al. [4] mentioned above also suggested preserving the ratio $f/Z$. They arrived at the constraint in a different and more specialized way, using zoom to null the divergent components of optic flow from a fronto-parallel plane translating with velocity $V$. Paraphrasing their argument, if image points $x$ are related to scene points $X$ as $x = (f/Z)X$ then, noting that $X = V$ when there is no rotation, the projected image motion is $x = \frac{f}{Z} X + (f/Z)V$. If $V$ is constant, zooming can cancel the radial field and indeed make the flow field constant when $f/Z$ is constant.

The two methods presented here explore more general geometry and consider calibration and fixation. The first method assumes a perspective camera undergoing pure rotation. The foreground scene motion relative to the camera is unconstrained, and the rotation requirement is imposed by the chosen method of camera self-calibration, that of de Agapito et al. [6], [7], which uses correspondences between points imaged from static background parts of the scene. The depth of the foreground is recovered using the planar structure from motion algorithm of Faugeras and Lustman [8]. This respects the local geometry of Fig. 2a, but also makes the segmentation of foreground from background a simpler problem of classification between two homographies. Fixation is based not on an individual feature, any of which is unlikely to survive over an extended sequence, but rather on the collective representation of many features via the planar structure.

As implementation and experiment will show, the perspective method has two drawbacks. First, the need for continuous self-calibration is onerous and, second, the small perspective distortion in most surveillance and broadcast imagery prevents its extension from planar to 3D structure.

Most such imagery can be well modeled by the weak perspective projection. The quantity $f/Z$ is again the appropriate measure to preserve, with $Z$ interpreted as the depth of the fronto-parallel plane onto which rays are projected orthogonally before being projected perspectively (Fig. 2b). However, the second method developed here is for the more general affine camera, where preserving $f/Z$ generalizes in order to preserve the scaling of the affine bases. The burden of calibration is lifted and the zooming camera may move quite freely while viewing an arbitrary scene. Affine transfer is used to maintain fixation [9], a process that is fundamentally invariant to zoom and which tolerates features appearing at and disappearing from the edge of the image as a wider or narrow view is taken [10]. Here, by finding the transformation which upgrades the affine structure to Euclidean [11], the ratio $f/Z$ is recovered up to an overall scale. Laying a zoom-variant process over a zoom-invariant fixation competence seems attractive from an architectural standpoint.
Zoom Control and Fixation for Perspective Planar Structure

1. Rotate camera to obtain initial self-calibration.
2. while Planar object detected do
5. Use SVD to find planar structure and motion.
6. Compute new gaze position.
7. Compute zoom scaling.
8. Send demands to camera axis and lens motor.
9. If background homography useful, improve calibration.
10. end while

Fig. 3. Algorithm for zoom control of a perspective camera.

2 ZOOM CONTROL FOR A ROTATING PERSPECTIVE CAMERA

The method for zoom control under perspective projection, summarized in Fig. 3, relies on the evaluation and analysis of two homographies, one relating matched features from the fixed foreground plane, and the other relating matches from the static background, which need not be planar.

2.1 The Foreground Homography and Planar Structure

In an arbitrary projective frame, the projection into the image of scene structure \(X\) is \(x = PX\), where \(P\) is the \(3 \times 4\) projection matrix and where equality is established only up to a nonzero scale factor. To recover depth, metric structure \(X_E\) is required, which is related to the projective structure \(X\) by a nonsingular \(4 \times 4\) transformation \(X = H_P X_E\) recovered by calibrating the camera. Then, \(x = P_X X_E\) where \(P_X = PH_P\). The Euclidean projection matrix can be decomposed as \(P_X = KH_X\), where \(K\) aligns the plane’s frame with the camera frame, \(t\) is the origin of the plane’s system in the camera system, and \(X\) is the matrix of intrinsic camera parameters (focal lengths \(f_u, f_v\), skew \(s\) and principal point \(u_0, v_0\))

\[
K = \begin{pmatrix} f_u & s & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{pmatrix}
\]

Let us first assume that the matrix \(K\) of varying intrinsic parameters in each frame is known and use it to recover the disposition of foreground scene points, assuming them to lie on a planar surface. With no loss of generality, the planar points have \(Z_E = 0\) and, for any view, \(x = KH(X_E, Y_E, 1)^T\) with \(H\) taking the form

\[
H = \begin{pmatrix} R_{11} & R_{12} & t_1 \\ R_{21} & R_{22} & t_2 \\ R_{31} & R_{32} & t_3 \end{pmatrix}
\]

When the plane moves between positions \(i\) and \(j\), corresponding foreground points are related by

\[
x_j = A_{ij} x_i = K_j H_j H_i^{-1} K_i^{-1} x_i.
\]

After determining the interimage homography \(A_{ij}\) from point correspondences in the image sequence and recovering the \(K\) matrices by self-calibration, one needs to decompose \(Q_{ij} = H_j H_i^{-1}\) to recover the disposition of the plane. Although the form of each \(H\) is simple, that of \(Q_{ij}\) is less so, but, as shown by Faugeras and Lustman [8], it can be decomposed by considering the rotation in the camera frame.

Let \(X_i = (X, Y, Z)_i^T\) be the inhomogeneous coordinate in the camera frame of a point on the plane \(P_i\), so that \(X_i \cdot n_i = d_i\) where \(d_i\) is the perpendicular distance from the plane to the origin, and \(n_i\) is the normal to the plane pointing into the surface. If the plane undergoes a rigid motion \(R_{ij}, t_{ij}\) into \(P_j\), the transformation \(X_j = R_{ij} X_i + t_{ij}\) can also be written

\[
X_j = \left( R_{ij} + t_{ij} n_i^T/d_i \right) X_i.
\]

In the camera frame \(x_i = \kappa_i X_i\) and \(x_j = \kappa_j X_j\) and, so, recalling (1),

\[
Q_{ij} = \kappa_j H_j H_i^{-1} = R_{ij} + t_{ij} n_i^T/d_i.
\]

Faugeras and Lustman [8] applied the Singular Value Decomposition (SVD) \(USV^T \rightarrow Q_{ij}\) to find \(R_{ij}, T_{ij} = (t_{ij}/d_i)\), and \(n_i\) from the orthonormal matrices \(U\) and \(V\) and the matrix of ordered singular values \(\Sigma\).

Depending on the multiplicity of the ordered singular values, different types of motion can be deduced [8], [12]. For general motion, there remains an ambiguity between motion and structure, exactly that elucidated by Longuet-Higgins in earlier work on planar structure from motion [13]. Faugeras and Lustman suggest a number ways in which the ambiguity can be removed. Here, as in [14], a third image is used to choose the plane normal which is consistent with the rotational motion over time.

2.2 Generating Fixation and Zoom Demands

In the first frame, the depth/translation scaling ambiguity is broken by setting \(d_1\) to unity, so that \(t_{12} = T_{12}\). The desired fixation point on the plane is initialized manually for the entire sequence by choosing its projection \(g_i\), a point which need not be associated with an actual image feature. The back-projected point on the scene plane is then

\[
G_1 = g_i/(g_i \cdot n_1)^T
\]

and, noting that \(d_1 = 1\), the gaze point in the next frame is found as

\[
G_2 = R_{12} G_1 + t_{12}.
\]

Because the motion of the foreground plane is determined within the camera frame, these equations are quite independent of rotations of the camera for fixation, and do not require the camera to be pointing in exactly the desired gaze direction.

For fixation and zooming, axis and lens demands are then generated to set

\[
\cos^{-1}\left(\frac{G_2 \cdot \tilde{z}}{G_1 \cdot \tilde{z}}\right) = 0, \quad \text{and} \quad f_2 = f_1 \frac{G_2 \cdot \tilde{z}}{G_1 \cdot \tilde{z}},
\]

where \(\tilde{z} = (0, 0, 1)^T\) is the optic axis. There is no imperative for the demands to be fully met before the next image is captured.

The demand for fixation involves a unit vector in the gaze direction and is independent of scale. The demand for
zoom is independent of the scale chosen for any pair of images. Thus, there is no need to maintain an overall scale. Let us use image pairs 1, 2, and 2, 3 to illustrate scale-free operation. \( G_2 \) is found from \( Q_{12} \), as just described. However, its 3D value is not carried forward when considering images 2 and 3. Instead, only its projection \( g_2 \) survives. Then, \( d_2 \) is set to unity, a new \( G_2 \) found from \( g_2/G_2 \cdot n_2 \) and images 2 and 3 treated exactly as 1 and 2, and so on.

What is lost here is a trace of the 3D motion of the gaze point, consistent to some overall scale. However, it is possible to recover this. Dropping subscripts and denoting the pairwise scale as \( S \), the recovered \( Q \) matrix is actually \( S Q \). Because the matrices \( U \) and \( V \) must be orthogonal, the SVD of \( S Q \) places the scaling into the singular values and, hence, into the recovery of translation. That is, \( R' \), \( S(t/d) \), and \( n \) are recovered. Now,

\[
S Q = S R' + S(t/d) n^T
\]

from which \( S R' \) can be found. But, \( R' \) is known, so \( S \) is found, and so too is \( t/d \).

Again, consider frame pairs 1, 2, and 2, 3 as an example. Using \( d_1 = 1 \), find \( G_2 \) from the first pair as before and use it to find \( d_2 \) for use in the next pair from

\[
d_2 = G_2 \cdot n_2 = G_2 \cdot R'_{ij} n_1.
\]

The overall scale established by the choice of \( d_1 = 1 \) is now preserved throughout the sequence.

In the case of pure rotation, \( H_{ij} = 0 \), and the plane normal is not recovered from Faugeras-Lustman. However, if the normal was known earlier, then it is a straightforward matter to update it using the rotation. In the unlikely event of continuous pure rotation by the foreground, the normal is never recovered, but its value is actually irrelevant. For example, setting \( n_1 \) such that \( G_1 = g_1 \) generates a succession of gaze points \( G_i \) sufficient to derive the demands specified above. Put another way, pure rotation allows the foreground points to be noncoplanar, in the same way that background points are noncoplanar.

### 2.3 Self-Calibration

The discussion so far has assumed knowledge of \( K \) for each frame \( i \). Because the focal length is continually changing, a self-calibration routine for a rotating camera is run along-side the structure from motion recovery. Maybank and Faugeras [15] and Pollefeys et al. [16] considered the self-calibration problem for general motion with fixed and varying intrinsic parameters, respectively. However, it is now well-appreciated that general motion methods do not perform gracefully in special motion cases. Hartley considered a rotating camera with fixed intrinsics in [17], and de Agapito et al. [6] and Seo and Hong [18] devised methods for handling varying intrinsics. These latter methods are based on the observation that for a rotating camera, the projections of static structure in successive images are related by a \( 3 \times 3 \) homography, even when the camera’s intrinsic parameters change. The same observation was made by Luong and Viéville [19], but not exploited for calibration. The self-calibration routine can be specialized to incorporate known and fixed quantities. Here, it is assumed that the skew is zero, that the aspect ratio is fixed and predetermined, and that through registration the principal point is at the origin.

The assumption that the camera motion is a pure rotation about the optic center is of course only an approximation. The axes of a pan-tilt mechanism are unlikely to intersect and, even if they were to, the optic center of a zoom lens moves along the optic axis as the lens is adjusted. Expressions for the error in both the discrete and instantaneous motion cases are derived in [20] and the influence on the self-calibration of rotating cameras shown to be small and certainly better treated as a perturbation to the rotating case than by a self-calibration routine for general motion. In our case, the correction is of order 1 percent and is neglected.

Defining the static structure in a Euclidean frame as \( X_S \), its projections in images \( i \) and \( j \) are \( x_i = P_E X_S \) and \( x_j = P_E X_S \). As the camera only rotates about its optic center, then background points in the two views are \( x_i = K_1 i R_1 i / j X_S \) and \( x_j = K_1 i R_1 i / j X_S \) and, hence,

\[
x_j = B_j x_i = K_1 i R_1 i / j K_1 j x_i.
\]

Since \( R_{ij} = R_{ij} \) is a rotation matrix it has the property that \( R_{ij} = R_{ij}^{-1} \) and, so \( K_1 i K_1 j T_j \) is a statement of the infinite homography constraint [17], [19]. The constraint may also be expressed using the image of the absolute conic

\[
\omega_j = (K_1 j K_1 j T_j)^{-1} = B_j^{-1} \omega_i B_j^{-1}
\]

where, at the last step, the reference image (subscripted 1) is referred to by multiplication through the chain of homographies, \( B_{ij} = B_{ij} \cdot B_{ij} \cdot B_{ij} \). In [6], de Agapito et al. used a nonlinear method to recover the intrinsic parameters, which proved too computationally expensive for real-time implementation. However, in [7] they, with Hartley, presented a linear solution. The infinite homography constraint shows that, at any frame, the six functions of the internal parameters defined by the symmetric \( \omega_j \) are linearly related to the same functions in the reference frame. For skew \( s = 0 \),

\[
\omega_j = \begin{bmatrix}
    a_{j1} & a_{j2} & a_{j3} \\
    a_{j2} & a_{j3} & a_{j1} \\
    a_{j3} & a_{j1} & a_{j2}
\end{bmatrix}
\begin{bmatrix}
    1/ f_u^2 & 0 & -u_0/ f_u^2 \\
    0 & 1/ f_v^2 & -v_0/ f_v^2 \\
    -u_0/ f_u^2 & -v_0/ f_v^2 & 1 + u_0/ f_u^2 + v_0/ f_v^2
\end{bmatrix}
\]

giving the constraint

\[
a_{j2} = (B_j^{-1} \omega_i B_j^{-1})_{12} = 0,
\]

which in turn contributes a row to the measurement equation for the functions of the internal parameters in the reference frame.
\[
(\begin{array}{cccccc}
M_{j1} & M_{j2} & M_{j3} & M_{j4} & M_{j5} & M_{j6} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\end{array})
(\begin{array}{c}
\alpha_{11} \\
\alpha_{12} \\
\alpha_{16} \\
\end{array}) = 0.
\]

The \(M_s\) here are functions of the measured elements of the \(B_s\). Stability is much improved by adding in knowledge of the principal point and aspect ratio. Registering the image sets the principal point at \((u_0, v_0) = (0, 0)\), from which
\[
(B_{j1}^{-1}B_{j1}^{-1})_{13} = 0 \quad \text{and} \quad (B_{j1}^{-1}B_{j1}^{-1})_{23} = 0,
\]
and setting the aspect ratio \(\alpha = f_u/f_v\) to a known fixed value gives
\[
\alpha^2(B_{j1}^{-1}B_{j1}^{-1})_{11} - (B_{j1}^{-1}B_{j1}^{-1})_{22} = 0.
\]

These add three further rows per frame to the design matrix in (2), from which the least-squares solution is found by eigen-decomposition.

3 IMPLEMENTATION OF THE PERSPECTIVE APPROACH

The approach has been implemented as both look-move and video-rate processes using an active robotic head whose cameras are fitted with computer-controlled zoom lenses from EIA. The algorithm given in Fig. 3 has as its first step an initial self-calibration, after which a loop is entered which computes information for fixation and zooming from the foreground, and allows updating of the self-calibration from the background. Comments on the implementation follow.

3.1 Initial Calibration

Both segmentation and fixation depend on having estimates of the focal length from the outset, and an initial calibration must be performed. Indeed, it proves useful to build a lookup table of measured focal length against zoom setting to use as a fallback if the background homography calculation fails, perhaps because too little of the background is visible. Precalibration is also useful for removing radial distortion, which would otherwise bias the online calibration results [21].

Some care has been taken to explore how best to coordinate rotation of the camera and zooming of the lens to improve the calibration’s robustness to image noise and minimize its failure rate [12]. The principal recommendation is that the movement be closed so that initial and final frames can be matched, and error distributed uniformly amongst each in the chain of homographies.

3.2 Segmentation and Homography Computation

On both foreground and background, image corners are located using Harris’ detector [22] and matched using a disparity approach. In the absence of prior knowledge, segmentation would be performed by applying RANSAC [23], or a refinement such as MLESAC [24], [25], incorporating linear homography estimation [26] to all the matched pairs. However, a short cut is available. As the active head’s encoders allow measurement of the camera’s pure rotation and as estimates of the camera intrinsics are available, image motion from the static background can be predicted. It is then a simple matter to identify background matches, leaving only foreground matches and outliers—possibly noise, possibly from other moving objects—to be separated using RANSAC. Offline comparison tests have shown that the segmentations obtained 1) using RANSAC alone and 2) using the odometry to peel away the background have no systematic differences and, so, the time-saving approach is adopted. Part of Fig. 7 below shows typical segmentation results obtained in real time.

4 EXPERIMENTAL RESULTS FOR THE PERSPECTIVE METHOD

4.1 Rotation at Fixed Depth

A book was placed in front of the camera and rotated about a vertical fixed axis. Fig. 4 compares the performance of the perspective approach with two image-based approaches, one which zoomed to preserve the target’s image area and one which attempted to preserve the object’s maximum dimension. While the perspective method kept the zoom fixed, both the image methods caused zoom to change. The area method performs worst: In the extreme, the area drops to zero as the plane turns side-on.

4.2 General Motion

A look-move implementation of the algorithm was used to generate the results of Fig. 5. A box was tracked as it was moved in a horizontal circle in front of the camera, while being twisted arbitrarily by some ±45 degrees about a vertical axis. The change of depth of the object is compensated for by zooming, and the twisting is ignored. The graphs show the recovered normalized distance (ratio of focal length to its initial value of 30mm) compared with ground truth (ratio of physical distance from the object to the camera to its initial value of 1.5m) and the recovered rotation angle again compared with ground truth. Their combination gives a representation of the trajectory followed, showing that the change in focal length follows the change in depth.
4.3 Looming Motion

From the stills of the general motion example in Fig. 5, it is difficult to see that the background is changing size. The toy robot sequence in Fig. 6 makes this far more obvious by compensating looming motion by zooming out. The object is not planar, but the depth-variation on the front side is sufficiently small that the motion is reasonably modeled using a planar homography. The motion was repeated 10 times, and the error bars show one standard deviation on either side of the mean for each position.

4.4 Video-Rate Results

Fig. 7 shows results from a video-rate (25Hz) implementation which processed 192 × 144 imagery using one 400MHz PIII processor. The stills cut from the video show the segmented matches from static background and foreground (a waving hand), with the foreground’s position stabilized by fixation and its size preserved by zooming.

The first two stills are from the period labeled “Motion I” on the read-outs of camera pan and tilt angles and lens focal length. The hand is being waved in an arc at constant depth as the variation in angles and fixed focal length reflect. The last three stills are from the “Motion III” period. Here, the hand is being waved back and forward, again apparent in the read-outs of constant angles and oscillating zoom. The constant image size of the hand is in marked contrast to the variation in image size of the monitor in the background.

The activity chart shows several stages when the calibration was updated during fixation. Updates occurred whenever the background homography was lost, but more than 20 sequential homographies were available.

5 DISCUSSION

The perspective method contains many elements of a successful zoom control system. However, it relies on a
computationally onerous self-calibration which, in turn, constrains the camera motion. Another irksome aspect is that realistic imagery contains too little perspective information to make the method stable for 3D rather than planar structure recovery. This conclusion was drawn from computer experiments in which 3D scene points were generated in a cube in front of a synthetic camera, modeled to have realistic field of view. Noise was added to the points after projection into the image. Matched triples of points were used to recover the trifocal tensor relating three views $ijk$. The algorithm went on to recover scale in the five steps given in Fig. 8.

Step 4 involves the use of a match to disambiguate the four possibilities for the motion [27], a step which fails, along with the control of zoom, in a variety of mildly adverse conditions, such as image noise, scene planarity, and small object relief compared to depth. The key observations are that the 3D method is very sensitive to planarity, and the solution becomes a matter of chance if the thickness is less than 0.1 of the width (75 percent failure, as shown in Fig. 9a). Similar degradation occurs when the object distance from the camera exceeds $\sim 10$ times the depth (as shown in Fig. 9b: but why Step 4 fails only in around 50 percent of cases remains puzzling).
Steps in 3D perspective method
1. Recover the fundamental matrices $F_{ji}$, $F_{ki}$, from the tensor.
2. Use the known calibrations $K_i$, $X_j$, $X_k$ to convert the fundamental matrices into essential matrices $E_{ji}$ and $E_{ki}$.
3. Decompose the essential matrices to find the rotations and translations up to scale.
4. Compose the camera matrices $P_i$, $P_j$, $P_k$.
5. Find the two reconstructions of all object points (including the fixation point) in view $i$ using the view pairs $ij$ and $ik$. Find the scale change that aligns the two reconstructions, and hence the scale of the translations.

6 ZOOM CONTROL FOR AN AFFINE CAMERA

Almost all surveillance and broadcast imagery contains little perspective distortion, and is quite adequately modeled by an affine projection. The recovery of scale to control zoom from such projections is the aim of the second method. The subcases for discernible and no discernible 3D information are considered. The method is again reliant on obtaining point matches, but now, after segmentation, only foreground matches are utilized in the recovery of affine structure and motion. The need for calibration of the camera is removed and the camera can move generally.

An unregistered image point $p_i$ in frame $i$, $x_{ip}$, is projected from the scene point $X_p$ (in inhomogenous coordinates) as

$$x_{ip} = M_i X_p + m_i,$$

and registered points formed by $m_i = \sum_{p=1}^{P} x_{ip}/P$ and $x_{ip} = x_{ip} - m_i$. From $P$ registered point correspondences established over $I$ frames, Tomasi and Kanade [28] recover affine structure and motion in batch mode by SVD of the $2I \times C^2$ registered measurement matrix

$$U \Sigma V^T \leftarrow W = \begin{pmatrix}
  x_{1,1} & \ldots & x_{1,I} \\
  \vdots & & \vdots \\
  x_{I,1} & \ldots & x_{I,I}
\end{pmatrix}$$

and use its rank-3 property to find the optimal affine projection matrices and structure from the ordered columns of $U$ and $V$

$$
\begin{pmatrix}
  M_1 \\
  M_I
\end{pmatrix} = (\sigma_1 u_1 \sigma_2 u_2 \sigma_3 u_3); \begin{pmatrix}
  X_1^T \\
  \vdots \\
  X_I^T
\end{pmatrix} = (v_1 v_2 v_3).
$$

6.1 Generating a Directional Demand for Fixation

Affine transfer [9], [29] uses the registered gaze points $g_i$ in the first $i = 1, \ldots, I - 1$ frames to find the 3D position of the gaze point $G$ from

$$G = \begin{pmatrix}
  M_1 \\
  \vdots \\
  M_{I-1}
\end{pmatrix}^+ \begin{pmatrix}
  g_1 \\
  \vdots \\
  g_{I-1}
\end{pmatrix},$$

where $+$ denotes the pseudoinverse. The gaze point is transferred to the latest frame $I$ using

$$g_I = M_I G.$$

6.2 Generating a Scale Demand for Zoom Control

The only requirement on the image points in a row of matrix $W$ in (3) is that they form an affine projection and, so, fixation using affine transfer is fundamentally invariant to zoom, whatever the movement of the camera [10]. However, it is possible to recover the relative scale of the affine basis for the control of zoom. Two ways of achieving this using 3D affine structure and two ways of approximating it using 2D measures have been explored.

6.2.1 3D Method I. Scale from Euclidean Constraints in Three Views

In inhomogeneous coordinates using registered image points and assuming zero pixel skew, one sound affine projection of a Euclidean structure is $x = M_E X_E$ with

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

where $\alpha$ is the aspect ratio and the $R$s are elements of the $3 \times 3$ rotation matrix between the world frame and the
camera frame. For a weak perspective camera, \( S = f/\tilde{Z} \), where \( \tilde{Z} \) is the mean depth of the points in the camera frame, but the more general affine case is pursued.

Structure \( X \) recovered from factorization must be related to \( X_E \) by an unknown affine transformation, \( HX_E = X \), and similarly each \( \Pi \) is related to its Euclidean counterpart by \( \Pi = \Pi_E \). The form of \( \Pi_E \) ensures that for the \( \ell \)-th frame [28]

\[
\Pi \Pi^T \Pi^T = \begin{pmatrix}
    p & q & r \\
    s & t & u \\
    h_1 & h_2 & h_3 \\
    h_2 & h_4 & h_5 \\
    h_3 & h_5 & h_6 \\
    p & s & t \\
    q & t & u \\
    r & u & v
\end{pmatrix}
= S^2 \begin{pmatrix}
    1 & 0 & 0 \\
    0 & 1/\alpha^2 & 0 \\
    0 & 0 & 1/\alpha^2
\end{pmatrix}.
\]

Matrix \( \Pi \Pi^T \) is symmetric and, hence, has 6 degrees of freedom. Assuming the aspect ratio is known, 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 fixation. Note too, that again only the projection matrices \( \Pi \) are used and, so, eigen-decomposition rather than SVD suffices to recover scale.

To recover both relative scales and the transformation \( \Pi \Pi^T \), set \( S_1 = 1 \) and rearrange (4) to read

\[
D \begin{pmatrix}
    h_1 \\
    S_2 \\
    S_3
\end{pmatrix} = \begin{pmatrix}
    1 & 0 & 1/\alpha^2 \\
    0 & 1/\alpha^2 & 0
\end{pmatrix},
\]

where the known matrix \( D \) involves terms in \( 1/\alpha^2 \) and \( p_i^2 \), \( 2p_iq_i \), and \( (p_i t_i + q_i s_i) \), for \( i = 1, 2, 3 \). The system can be solved for \( S_{2,3} \) and \( h = (h_1, \ldots, h_6) \) using SVD. Although unnecessary for zoom control, the transformation \( \Pi \) can be recovered from \( \Pi \Pi^T \) by Cholesky decomposition.

### 6.2.2 3D Method II. Scale from Epipolar Geometry and Two Views

Three-dimensional affine transfer requires three camera matrices—which fits well enough with the three-view scale method above. However, it is possible to recover structure and scale from just two [30], [31]. Here, matrices \( \Pi_2 \) and \( \Pi_3 \) are the obvious choices.

From the elements of the affine fundamental matrix expressing the epipolar constraint

\[
x_3^T F_{23} x_2 = x_1^T \begin{pmatrix}
    0 & 0 & a \\
    0 & 0 & b \\
    c & d & e
\end{pmatrix} x_2 = 0,
\]

where \( e = 0 \) when the points are registered, the relative scale is recovered as

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

where \( \alpha \) is the aspect ratio. The elements of the registered fundamental matrix can be found directly from the epipoles \( e_{23} = (d, -c, 0)^T \) and \( e_{32} = (b, -a, 0)^T \), which are in turn

\[1\] This expression differs from (36) of Shapiro et al. [31] because, in that section of analysis, they tacitly assumed unity aspect ratio.

Fig. 10. A scatter plot of the recovered scale factor versus actual scale factor for the 3D Euclidean and Epipolar methods and the 2D Determinant and 2-norm methods. Positional noise of 1 percent was added to image features. All behave similarly.

Given by the determinants of four \( 3 \times 3 \) minors involving rows three to six of the joint projection matrix which appears as the left column of the system [32], [33], [34]

\[
\begin{pmatrix}
    M_1 \\
    M_2 \\
    M_3
\end{pmatrix} = \begin{pmatrix}
    x_1 \\
    x_2 \\
    x_3
\end{pmatrix}, \quad \begin{pmatrix}
    X
\end{pmatrix} = \lambda \begin{pmatrix}
    I
\end{pmatrix}.
\]

Writing the \( j \)-th row of the \( I \)-th projection matrix \( M_I \) as \( m_i^j \) the components of the epipoles are [34]

\[
e_{iI} = \det \begin{pmatrix}
    m_i^j \\
    m_j^i
\end{pmatrix}, \quad j \in \{1, 2\}; \quad e_{iI}^3 = 0.
\]

Fig. 10 compares the recovered scale factors \( S_3/S_1 \) from Methods I and II (plotted as circles and crosses labeled “Euclidean” and “Epipolar,” respectively) as a function of the inverse ratio of mean depths \( Z_1/Z_3 \) in an experiment where 20 scene points were 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 1 percent of the spread of points in the image (thereby simulating a constant uncertainty in pixels were the camera to be rezoomed in between frames). Points on the graphs should lie on the unit slope line, and the two methods appear to perform equally well over a large range of relative scale changes.

### 6.2.3 2D Methods III and IV for Scale

Although we have earlier argued against image-based methods, it is nonetheless worthwhile to see how slightly more sophisticated image-based approximations perform; and so also plotted on Fig. 10 are the results from two such.

The first, labeled “Determinant,” supposes that as the 3D affine transformation \( \Pi \) 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 (4)
The determinant of $M_i M_i^\top$ gives the square of the area spanned by the rows of $M_i$. The approximation is equivalent to scene-based scale recovery when $H$ is close to identity or $M_i$ and $M_j$ are related only by an image rotation about the optic axis.

As the introductory example of the spinning disc suggested and the experiment in Fig. 4 showed, in situations where the target rotates, using relative area to control zoom is a poor choice and so therefore should be the ratio of determinants. It is better to approximate the change in a single dimension of the object, ideally that along the projection of the rotation axis in the image. In the method labeled “2-norm” the individual singular values of $M_i M_i^\top$ 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

$$\left(\frac{S_i}{S_j}\right)_{\text{area}} \approx \frac{\det(M_i H H^\top M_i^\top)}{\det(M_j H H^\top M_j^\top)} \approx \frac{\det(M_i M_i^\top)}{\det(M_j M_j^\top)}.$$  

The determinant of $M_i M_i^\top$ gives the square of the area spanned by the rows of $M_i$. The approximation is equivalent to scene-based scale recovery when $H$ is close to identity or $M_i$ and $M_j$ are related only by an image rotation about the optic axis.

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.

**6.3 Further Comparisons**

The relative robustness of the 3D and 2D methods as the level of noise is increased is illustrated in the two parts of Fig. 11a. The ratio of recovered scale to actual scale was determined for each of a 1,000 trials at every noise value and the mean $\mu$ of the errors and standard deviation $\sigma$ of the errors determined.

If the interframe rotation becomes small, or the object has little depth variation, then the rank of the measurement matrix $W$ in (3) drops to two (at least in zero noise). Fig. 11b shows the case of the object rotation decreasing to zero with image noise fixed. Both scene and image-based methods work equally well across the full range. Note, however, that, for the image-based measures, the errors shown are between the recovered dimension change and the actual dimension change, and the small error does not indicate a small difference between scene and image methods.

The final test of Fig. 11c 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 percent of the other two. For thicknesses below 10 percent, the scene-based methods degrade rapidly giving unpredictable ($\sigma$ large) and biased ($\mu \neq 0$) results. As expected, the 2-norm method is little affected as only one object dimension is required. This
situation in examined in more detail by considering purely planar objects.

### 6.4 Affine Views of a Planar Scene

#### 6.4.1 The Failure of Scene-Based Methods

Without loss of generality, the scene points are assumed to lie on a plane $X_E = (X, Y, 0)$. Recovering planar affine motion [9] is analogous to the 3D method, but the analogy for scale recovery cannot be completed. One reaches

$$M_i M_i^T = S_i^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1/\alpha \\ 0 & 0 & 1/\alpha \end{pmatrix} M_i^T \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1/\alpha \\ 0 & 0 & 1/\alpha \end{pmatrix}$$

but, unlike the 3D case, matrix $N_i$ is not the product of a rotation with its transpose and hence not an identity matrix.

One can analyze $N_i$ directly using the elements of the $3 \times 3$ rotation matrix, but an alternative is to parameterize the rotation as a succession of rotations by angles $\gamma$ around the $z$-axis, $\beta$ around the $y$-axis, and $\delta$ around the $x$-axis. Then, the matrix is

$$N_i = \begin{pmatrix} \cos^2 \beta & -\sin \gamma \sin \beta \cos \beta \\ -\sin \gamma \sin \beta \cos \beta & \sin^2 \gamma \cos^2 \delta + \cos^2 \beta \end{pmatrix}. $$

Actually independent of $\gamma$. Although there are range constraints on the various elements, the only hard constraint is that the matrix’s larger eigenvalue $\lambda_i^\beta(N_i)$ is unity. It is no longer possible to recover the transformation if one fares no better with the epipolar method, as the $3 \times 3$ 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 [31].

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 $z$-axis plus an opposite scaling. Unless either object structure or motion is known one must resort to image-based measures.

#### 6.4.2 Image-Based Methods

For a truly planar object, if the object turns side-on to the camera its area drops to zero and the determinant of $HH^T$ is an unsuitable measure, as was seen in Fig. 11c. The 2-norm method however does remain valid and stable because, when the projection of a 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 (rotation about the optic axis plus scaling, $HH^T$ equal to the identity matrix or maximum dimension coincident with the rotation axis).

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

---

**Zoom Control and Fixation using Affine Structure**

1: Wait for detection of independent motion  
2: Initialise fixpoint at centroid  
3: while Object detected do  
4: Detect and match corner features  
5: Segment back/foreground using guided MLESAC  
6: Factorise and recover $N_i$  
7: Transfer fixation  
8: if Third singular value significant then  
9: Perform Euclidean scale recovery (Method I)  
10: else  
11: if Translating then  
12: Perform determinant scale recovery (Method III)  
13: else  
14: Perform 2-norm scale recovery (Method IV)  
15: end if  
16: end if  
17: if Live capture then  
18: Demand pan and tilt to recentre fixation  
19: Demand zoom to compensate scale  
20: else  
21: Digitally zoom image  
22: end if  
23: end while

---

Fig. 12. Algorithm for zoom control of an affine camera.

### 7 Implementation of and Results from the Affine Method

Fig. 12 shows the algorithm implemented for frame-rate processing of video—either stored or actively captured. Corner features localized using Harris’ detector [22] and tracked individually over time again form the basic input. Segmentation between foreground and background matches is made using guided MLESAC algorithm [25], [24]. The foreground matches are passed to the 3D transfer method using eigen-decomposition of $WW^T$ as we are not explicitly interested in structure. Because transfer involves reprojection, its performance does not degrade either when there is insufficient rotation to recover 3D structure, or when the viewed object is planar [9].

If the three singular values (squares of the eigenvalues) are broadly similar, the 3D scene-based techniques are used. Of these, full Euclidean reconstruction (Method I) is favored over the epipolar constraint as the earlier trials showed it somewhat more resilient to planarity and similarly robust to noise.

However, in typical imagery, if the third singular value drops compared with the other two it usually indicates that the object is translating. Method I (or II) will often continue to function satisfactorily, but Methods III and IV are more robust to noise and are used instead. It is not unusual for one surface of an object to dominate feature generation in a particular view and for the points to be close to coplanar. However, for this to continue over time usually means that the object is translating, a case dealt with by either Methods III or IV. As soon as rotation occurs, new surfaces usually become visible, breaking coplanarity, and Method I is used. In the unlikely event of the object rotating but the point set remaining coplanar, Method IV, taking 2-norms is the only preferred option.
For live capture demands are sent to the active platform in order to recenter the fixation point and to the zoom lens to compensate scale changes, but for processing of stored video the images are scaled digitally.

7.1 Truck Sequence
The sequence shown at the top of Fig. 13 follows a truck as it approaches and leaves a road junction. It is initially head-on to the surveillance position, giving small image area and dimension. As the truck turns the corner—approximately 90 degrees—the maximum dimension and area increase significantly. The truck then translates into the distance.

The fixation point was initialized in the center of the visible features that were segmented as foreground, placing it just behind the cab. Rather than switch methods, the scales recovered by all four are shown in Fig. 15a. The plot

Fig. 13. (a) Stills from the “scaffolding truck” surveillance sequence during which the actual zoom setting was fixed. Corner features were detected, tracked, and segmented, and foreground points fed to the 3D affine transfer and scale recovery algorithm. The cross marks the fixation point. (b) The images digitally zoomed to compensate for scale changes using the determinant method (Method III).
clearly shows the increase in target dimension and area, and the corresponding decrease in zoom demand for the 2-norm and determinant methods, respectively. The ratio of singular values in Fig. 15b indicates that the motion is effectively planar throughout.

Fig. 13b shows the images digitally zoomed to the scale recovered from the determinant method. While there is a distinct size change between the first image and the rest, the scaled images clearly show the improvement in object stability.

Fig. 14. (a) 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. (b) The corresponding images after digital zooming using the scale factor recovered from Method I.
7.2 White Van Sequence
The second traffic sequence in Fig. 14 involves a white van approaching the same road junction from the opposite direction. The scales recovered using all four methods are shown in Fig. 15c. Again, the singular values of Fig. 15d indicate effectively planar motion throughout the sequence and, so, to demonstrate the validity of the 3D methods, we force use of the full Euclidean method (Method I) throughout.

Fig. 14a shows stills from the original sequence and Fig. 14b shows the digitally zoomed sequence using the recovered scale. The preservation of size in this automatically zoomed sequence appears to be of similar quality to that exhibited in the sequences captured by skilled human operators, such as that in Fig. 1b in the introduction.

8 Discussion and Conclusions
In this paper, we have suggested that automatic zooming may be as valuable as automatic fixation in enhancing the visual information content of video supplied to a computer vision system and have justified the geometric constraint which compensates for depth motion by zooming, a constraint which is applicable to both perspective and affine projections.

Control of zoom has been implemented under perspective projection by constraining the camera motion to pure rotation. This permits a linear method of self-calibration to provide both an initial calibration which is stored in a look-up table and a continuous update of the calibration during fixation and zooming.

Experiments using synthetic data show that full 3D reconstruction is not viable and, so, scene depth is recovered using a planar structure from motion algorithm. The structure is used to transfer the gaze position and, hence, to compute the relative depths of new and old gaze positions to change the focal length in the same ratio. The method has been demonstrated in a number of offline look-move experiments and in an online video-rate implementation using an active head fitted with zoom lenses.

The perspective approach is interesting, but challenging to apply. Most imagery simply contains too little information, either for reconstruction or for maintaining the
calibration. The latter requires many matches across several views, whereas often the background is blank, the restricted depth of focus smears out features, or the fixated object fills most of the field of view. Care must be taken that a calibration update is only performed when the background shows a large number of matches and is tracked across a large number of views.

The second part of the paper turned to consider affine methods, where the need for self-calibration is removed. The paper outlined two 3D methods and two 2D methods for recovering scale from affine sequences and explored their performance under a range of noise conditions and under decreasing rotation and increasing coplanarity both of which cause the 3D affine structure recovery to become degenerate. 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.

Video-rate implementation of the affine process has been demonstrated using a laboratory bound active stereo rig, and applied to surveillance footage using post processing. The invariance of the affine transfer method of fixation to scale changes means that the tracking and zooming competences can be treated as independent, even though the control of zoom uses information from the control of fixation.

Table 1 summarizes our conclusions as to the advantages and disadvantages of the two principal methods and their possible variants.

The fixation and zooming system described here has not included a motion model for the foreground object. Recent work by the authors has investigated this area. The difficult issue turns out not to be that of modeling scene motion, but that of modeling the opto-mechanical performance of zoom lenses which typically run open-loop and slowly.

ACKNOWLEDGMENTS

This work was supported by Grants GR/L58668 and GR/N03266 from the UK Engineering and Physical Science Research Council. B. Tordoff was supported by an EPSRC Research Studentship and Research Assistantship.

REFERENCES
