Motion From Point Matches Using Affine Epipolar Geometry

Larry S. Shapiro, Andrew Zisserman and Michael Brady

Robotics Research Group, Department of Engineering Science,
Oxford University, Parks Road, Oxford, OX1 3PJ.

Abstract. Algorithms to perform point-based motion estimation under orthographic and scaled orthographic projection abound in the literature. A key limitation of many existing algorithms is that they rely on the selection of a minimal point set to define a "local coordinate frame". This approach is extremely sensitive to errors and noise, and forfeits the advantages of using the full data set. Furthermore, attention is seldom paid to the statistical performance of the algorithms. We present a new framework that caters for errors and noise, and allows *all* available features to be used, without the need to select a frame explicitly. This theory is derived in the context of the *affine camera,* which generalises the orthographic, scaled orthographic and para-perspective models. We define the affine epipolar geometry for two such cameras, giving the fundamental matrix in this case and discussing its noise resistant computation. The two-view rigid motion parameters (the scale factor between views, projection of the 3D axis of rotation and cyclotorsion angle) are then determined *directly* from the epipolar geometry. Optimal estimates are obtained over time by means of a linear Kalman filter, and results are presented on real data.

1 Introduction

Orthographic and scaled orthographic projection are widely used in computer vision to model the imaging process [1, 3, 5, 7, 9, 10, 21, 22, 23]. They provide a good approximation to the perspective projection model when the field of view is small and the variation in depth of the scene along the line of sight is small compared to its average distance from the camera [20]. More importantly, they expose the ambiguities that arise when perspective effects diminish. In such cases, it is not only *advantageous* to use these simplified models but also *advisable* to do so, for by explicitly incorporating these ambiguities into the algorithm, one avoids computing parameters that are inherently ill-conditioned [7]. This paper investigates the motion estimation problem in the context of the *affine camera*, which generalises the orthographic, scaled orthographic and para-perspective models (see [18]).

Many existing point-based motion algorithms are of limited practical use because the inevitable presence of noise is often ignored [10, 12], unreasonable demands are often made on prior processing (e.g. a suitable perceptual frame must first be selected) [10], special case motions are often assumed (e.g. no rotation about a fixed axis) [8, 9], and some algorithms require batch processing rather than the more natural sequential processing [21]. The tool we employ to redress these

Lecture Notes in Computer Science, Vol. 801
Jan-Olof Eklundh (Ed.)
Computer Vision – ECCV '94

shortcomings is *affine epipolar geometry*. The epipolar constraint is well-known in the stereo literature, and has also been used in motion applications under perspective and projective viewing to establish motion correspondence, recover the translation direction and compute rigid motion [6, 12]. By contrast, *affine* epipolar geometry has seldom been used for motion estimation (though see [9, 10]).

Section 2 defines the epipolar geometry of the affine camera and derives its special fundamental matrix; no camera calibration is needed at this juncture. To obtain a reliable solution for these parameters, we evaluate three least squares algorithms based on image distances, and determine that a 4D linear method performs best. The utilisation of *all* available points (rather than just a minimum set) not only improves the accuracy of the solution (by providing immunity to noise and enabling detection of outliers), but also obviates the need to *select* a minimal point set. Section 3 relates the affine epipolar geometry to the rigid motion parameters, and formalises Koenderink and van Doorn's novel motion representation [10]. Using two views, we compute scale, cyclotorsion and the projected axis *directly* from the epipolar geometry, requiring only the aspect ratio. Our n-point framework subsumes the results for minimum configurations. For the multiple view case, we define a linear Kalman filter to determine optimal two-view estimates. Unlike some previous point-based structure and motion schemes (e.g. [4]), we do not assign an individual Kalman filter to each 3D feature; this liberates us from having to track individual 3D points through multiple views, so points can appear and disappear at will.

2 Affine epipolar geometry

2.1 Affine and weak perspective cameras

A camera projects a 3D world point $\mathbf{X} = (X, Y, Z)^\top$ into a 2D image point $\mathbf{x} = (x, y)^\top$. The *weak perspective* (or *scaled orthographic*) camera has the form

$$\mathbf{x} = \frac{f}{Z^c_{ave}} \begin{bmatrix} \xi \mathbf{R}_1^\top \\ \mathbf{R}_2^\top \end{bmatrix} \mathbf{X} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} = \mathbf{M}_{wp}\mathbf{X} + \mathbf{t}_{wp}, \tag{1}$$

where $\mathbf{M}_{wp}$ is a 2×3 matrix whose rows are the scaled rows of a rotation matrix $\mathbf{R} = [R_{ij}]$, and $\mathbf{t}_{wp} = (t_x, t_y)^\top$ is a 2-vector (the projection of the origin of the world coordinate frame, $\mathbf{X} = \mathbf{0}$). This equation is derived by approximating the depth Z^c_i of each individual point i (measured along the line of sight in the camera frame) by the average distance of the object from the camera, Z^c_{ave}. The camera is "calibrated" when its intrinsic parameters are known, namely the camera aspect ratio ξ and focal length f.

The *affine camera* has the same form as Equation (1) but has no constraints on the matrix elements. It is written as

$$\mathbf{x} = \mathbf{M}\mathbf{X} + \mathbf{t}, \tag{2}$$

where $\mathbf{M}$ is a general 2×3 matrix and $\mathbf{t}$ a general 2-vector. The affine camera has eight degrees of freedom and corresponds to a projective camera with its optical centre on the plane at infinity [14]. Consequently, *all projection rays are* *parallel*, and lines that are parallel in the world remain parallel in the image. The affine camera covers: (i) a 3D *affine* transformation between world and camera coordinate systems; (ii) parallel projection onto the image plane; and (iii) a 2D affine transformation of the image. It therefore generalises the weak perspective model in two ways: *non–rigid* deformation of the object is permitted (due to the 3D *affine* transformation) and calibration is unnecessary (unlike in Equation (1)).

Consider an affine stereo pair. A 3D world point $\mathbf{X}_i$ is projected by an affine camera $\{\mathbf{M}, \mathbf{t}\}$ to an image point $\mathbf{x}_i = \mathbf{M}\mathbf{X}_i + \mathbf{t}$, and the scene moves according to $\mathbf{X}'_i = \mathbf{A}\mathbf{X}_i + \mathbf{D}$, where $\mathbf{X}'_i$ is the new world position, $\mathbf{A}$ a 3×3 matrix and $\mathbf{D}$ a 3–vector. This *motion transformation* encodes relative motion between the camera and the world as a 3D affine transformation (12 degrees of freedom). The new world point projects to

$$\mathbf{x}'_i = \mathbf{M}\mathbf{X}'_i + \mathbf{t} = \mathbf{M}\left(\mathbf{A}\mathbf{X}_i + \mathbf{D}\right) + \mathbf{t} = \mathbf{M}\mathbf{A}\mathbf{X}_i + \left(\mathbf{M}\mathbf{D} + \mathbf{t}\right) = \mathbf{M}'\,\mathbf{X}_i + \mathbf{t}', \quad (3)$$

which can be interpreted as a second affine camera $\{\mathbf{M}', \mathbf{t}'\}$ observing the original scene, where $\{\mathbf{M}', \mathbf{t}'\}$ accounts for changes in both the extrinsic and intrinsic camera parameters.

2.2 The affine epipolar line and fundamental matrix

The concept of an epipolar line is well known in the stereo and motion literature. For an affine camera, the epipolar lines are all parallel, since the projection rays are parallel and the affine camera preserves parallelism. Thus, the epipoles lie at infinity in the image planes. An implicit form of the epipolar line is derived by eliminating the world coordinates (X_i, Y_i, Z_i) from Equations (2) and (3), giving a single equation in the *image measurables*:

$$\boxed{a\,x'_i + b\,y'_i + c\,x_i + d\,y_i + e = 0} \quad (4)$$

This *affine epipolar constraint equation* [24] is a *linear* equation in the unknown constants $a \ldots e$, which depend only on the camera and motion parameters, not structure. Only the ratios of $a \ldots e$ can be computed, so Equation (4) has only four independent degrees of freedom. Solving this equation does not require a calibrated camera, since an affine camera model has been used throughout. This equation may also be expressed in the form of a *fundamental matrix* $\mathbf{F}_A$,

$$\mathbf{p}'^{\top}\,\mathbf{F}_A\,\mathbf{p} = \begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0, \quad (5)$$

where $\mathbf{p}' = (x', y', 1)^{\top}$ and $\mathbf{p} = (x, y, 1)^{\top}$ are homogeneous image vectors. The matrix $\mathbf{F}_A$ has maximum rank two. The epipolar lines corresponding to $\mathbf{p}$ and $\mathbf{p}'$ are $\mathbf{u}' = \mathbf{F}_A\mathbf{p}$ and $\mathbf{u} = \mathbf{F}_A^{\top}\mathbf{p}'$ respectively, where $\mathbf{u} = (u_1, u_2, u_3)^{\top}$ represents the line $u_1 x + u_2 y + u_3 = 0$. The form of $\mathbf{F}_A$ in Equation (5) is a special case of the general 3×3 fundamental matrix $\mathbf{F}$ used in stereo and motion algorithms (e.g. [13]). Equation (4) can also be written as $\mathbf{r}_i^{\top}\mathbf{n} + e = 0$, where $\mathbf{r}_i = (x'_i, y'_i, x_i, y_i)^{\top}$ and $\mathbf{n} = (a, b, c, d)^{\top}$. Here, $\mathbf{n}$ is the normal to a 4D hyperplane and when $\mathbf{r}_i$ is noisy,

$|\,\mathbf{r}_i{}^\mathsf{T}\mathbf{n} + e\,|\;/\;|\,\mathbf{n}\,|$ is the 4D perpendicular distance from $\mathbf{r}_i$ to this hyperplane. For the following, $\bar{\mathbf{r}}$ will denote the centroid of the 4–vectors $\{\mathbf{r}_i\}$ and $\mathbf{v}_i$ the centred points $\mathbf{v}_i = \mathbf{r}_i - \bar{\mathbf{r}}$. Note that $\mathbf{n}_1 = (c, d)^\mathsf{T}$ and $\mathbf{n}_2 = (a, b)^\mathsf{T}$ are the 2D normals to the epipolars in I_1 and I_2 respectively (Figure 1).

2.3 Solving the epipolar equation

Equation (4) is defined up to a scale factor, so only four point correspondences are needed to solve for the four independent unknowns (conditions for *existence* of a solution are discussed in [19]). When n correspondences are available ($n > 4$), it is advantageous to use *all* n points, since this improves the accuracy of the solution, allows detection of (and hence provides immunity to) outliers, and obviates the need to select a minimal point set. The presence of "noise" (i.e. corner localisation/measurement error) in the overdetermined system means that the points won't lie exactly on their epipolar lines (Figure 1), and an appropriate minimisation is required. The perpendicular distance D'_i between $\mathbf{x}'_i$ and its associated epipolar line in I_2 is $D'_i = (\mathbf{r}_i^\mathsf{T}\mathbf{n} + e)/\sqrt{a^2 + b^2}$; the counterpart distance in I_1 is $D_i = (\mathbf{r}_i^\mathsf{T}\mathbf{n} + e)/\sqrt{c^2 + d^2}$.

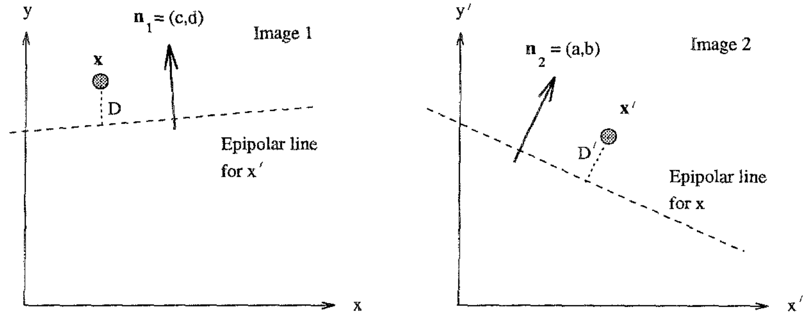


Fig. 1. *The normals to the epipolar lines are* $\mathbf{n}_1$ *and* $\mathbf{n}_2$. *Noise displaces a point* $\mathbf{x}'$ *in* I_2 *from the epipolar line associated with its counterpart* $\mathbf{x}$ *by perpendicular distance* D'. *A similar displacement by* D *occurs in* I_1.

We examine the following three minimum variance cost functions which involve the epipolar parameters, and differ in the image distances minimised:

$$E_1(\mathbf{n}, e) = \left(\frac{1}{a^2 + b^2} + \frac{1}{c^2 + d^2}\right) \sum_{i=0}^{n-1} (ax'_i + by'_i + cx_i + dy_i + e)^2 \tag{6}$$

$$E_2(\mathbf{n}, e) = \frac{1}{a^2 + b^2} \sum_{i=0}^{n-1} (ax'_i + by'_i + cx_i + dy_i + e)^2 \tag{7}$$

$$E_3(\mathbf{n}, e) = \frac{1}{a^2 + b^2 + c^2 + d^2} \sum_{i=0}^{n-1} (ax'_i + by'_i + cx_i + dy_i + e)^2 \tag{8}$$

All three functions minimise the sum of squares of a perpendicular distance measure, all are *scale–invariant* (i.e. if $\{\mathbf{n}, e\}$ is a solution, then so is $\{k\mathbf{n}, ke\}$ where k

is a non–zero scalar), and all can be minimised over e directly (giving $e = -\mathbf{n}^\top \bar{\mathbf{r}}$).

Discussion The three abovementioned functions all involve image distances; this is important since the observations are made in the image and the system noise originates there [7]. We assess these cost functions in terms of accuracy and complexity, and show that E_3 is superior to E_1 and E_2 in several respects.

Cost function E_1 sums the squared perpendicular image distances over I_1 and I_2, i.e. $E_1 = \sum_{i=0}^{n-1} D_i^2 + (D_i')^2$. The solution satisfies a system of non–linear simultaneous equations and requires non–linear minimisation. Cost function E_2 sums the squared perpendicular distances in a single image, e.g. $E_2 = \sum_i (D_i')^2$ (for I_2). The solution involves a 2D eigenvector equation. Cost function E_3 sums the squared 4D perpendicular distances between the concatenated image points and the 4D fitted hyperplane, i.e. $E_3 = \sum_i (\mathbf{r}_i \cdot \mathbf{n} + e)^2 / \,|\,\mathbf{n}\,|^2$. This is classic linear least squares, or *orthogonal regression.* The solution satisfies the eigenvector equation $\mathbf{W}\,\mathbf{n} = \lambda_1\,\mathbf{n}$, where $\mathbf{W} = \sum \mathbf{v}_i\,\mathbf{v}_i^\top$ and $\mathbf{n}$ is the unit eigenvector corresponding to the minimum eigenvalue λ_1.

Faugeras et al. [13] evaluated candidate cost functions for computing the fundamental matrix $\mathbf{F}$ of a *projective* camera; E_1 is the affine analogue of their favoured non–linear criterion (using distances to epipolar lines[1]) and E_3 is the analogue of their linear criterion (using the eigenvector method). They criticised the linear approach for failing to impose the rank constraint on $\mathbf{F}$ and for introducing a bias into the computation by shifting the epipole towards the image centre. In the *affine* case, however, $\mathbf{F}_A$ is *guaranteed* to have a maximum rank of two (cf. Equation (5)) and the epipole lies at infinity, removing these two objections against the linear method. Furthermore, although E_3 may be interpreted as a 4D algebraic distance measure, it is equivalent to an *image* distance measure based on point–to–point (rather than point–to–line) distances. It measures the distance between the observed image location and the location predicted by projecting the computed affine structure using the computed affine cameras (cf. Equations (2) and (3)), that is,

$$E_{TK} = \sum_{i=0}^{n-1} |\,\mathbf{x}_i - \mathbf{M}\mathbf{X}_i - \mathbf{t}\,|^2 + \sum_{i=0}^{n-1} |\,\mathbf{x}_i' - \mathbf{M}'\mathbf{X}_i - \mathbf{t}'\,|^2 \,. \tag{9}$$

Reid [16] showed Equation (9) to be the cost function minimised by Tomasi and Kanade [21]. We have shown further [19] that after differentiating E_{TK} with respect to $\mathbf{t}$, $\mathbf{t}'$ and $\mathbf{X}_i$ and resubstituting, E_3 obtains. It is sensible to minimise E_{TK} since it involves the exact number of degrees of freedom in the system, namely $\mathbf{t}$, $\mathbf{t}'$, $\mathbf{M}$, $\mathbf{M}'$ and $\mathbf{X}_i$. Thus, E_3 is optimal with respect to both the structure $\mathbf{X}_i$ and the camera parameters $\{\mathbf{M}, \mathbf{t}\}$ and $\{\mathbf{M}', \mathbf{t}'\}$.

It can be shown that E_2 is the affine version of the expression minimised by Harris [7]. This approach has several drawbacks, the most important being that by only minimising the noise in *one* image, the errors are unevenly distributed between I_1 and I_2: a set of epipolars which fits one image well, may not do likewise

[1] They also weighted each point by its inverse distance to the epipole; for the affine case, the epipole lies at infinity so all points are weighted equally.

in the other image, leading to discrepancies in the epipolar geometry [13]. The E_2 method is therefore unattractive.

Noise model The noise characteristics of linear least squares solutions (such as E_3) were analysed in [17]. Suppose each data point $\mathbf{r}_i$ is perturbed by independent, isotropic, additive, Gaussian noise $\delta\mathbf{r}_i$. The noise has zero mean ($E\{\delta\mathbf{r}_i\} = \mathbf{0}$) with variance σ^2, so $E\{\delta\mathbf{r}_i \delta\mathbf{r}_j^\top\} = \delta_{ij}\sigma^2\mathbf{I}_4$, where δ_{ij} is the Kronecker delta function and $\mathbf{I}_4$ the 4×4 identity matrix. The noise in $\mathbf{r}_i$ induces an error $\delta\mathbf{v}_i$ in the centred data point $\mathbf{v}_i$, which propagates through to the solution $\mathbf{n}$. The eigenvalues of $\mathbf{W}$, $\{\lambda_1, \ldots, \lambda_4\}$, are arranged in increasing order with corresponding eigenvectors $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$. The eigenvector corresponding to the minimum eigenvalue, $\mathbf{u}_1$, gives the solution vector $\mathbf{n}$. The covariance matrix for $\mathbf{n}$ is [17]

$$\Lambda_{\mathbf{n}} = [\Lambda_{ij}] = E\{\delta\mathbf{n}\,\delta\mathbf{n}^\top\} = \sigma^2 \sum_{k=2}^{4} (\mathbf{u}_k\,\mathbf{u}_k^\top)/\lambda_k. \tag{10}$$

This matrix provides a confidence measure in the parameters of the epipolar fit. Furthermore, it facilitates the rejection of *outliers*, "rogue observations" which plague data analysis techniques such as linear least squares regression. Removing these outliers is crucial since an analysis based on the contaminated data set distorts the underlying parameters. This is another reason for using all available points, since outliers cannot be identified using minimal point sets. We employ the eigenvalue–based regression diagnostic of Shapiro and Brady [17].

Results Figure 2 shows two sequences, one with a camera moving in a static world and the other with an object moving relative to a stationary camera. Corner features were extracted and tracked over time (using the scheme in [19]), and outliers removed. Figure 2 shows the computed epipolar lines. The mean perpendicular distances between each corner and its epipolar line are 0.76 and 0.49 for the two sequences respectively; the epipolar lines are thus typically within pixel accuracy (on 256 × 256 images) and so provide effective constraints for correspondence.

Figure 2(e) illustrates the advantage of using *all* available points when computing epipolar geometry. A synthetic scene with 63 points (no outliers) had its 256 × 256 images corrupted by independent, isotropic, Gaussian noise ($\sigma = 0.6$ pixels). Subsets of the data comprising p points (where p varied from 4 to 63) were randomly selected and a fit $\{\mathbf{n}, e\}$ computed using this subset. The E_1 distance was then calculated for the whole point set, summing the squared perpendicular image distances from each point to its computed epipolar line. For each value of p, 500 experiments were performed. The median distance and the standard deviation of the distances are shown for each value of p. Both decrease as p increases, showing that the use of more points leads not only to better fits but also to more consistent ones.

3 Rigid motion: two views

It is well–known that two distinct views of four non–coplanar, rigid points generate a one–parameter family of structure and motion solutions under parallel projection [3, 9, 10]. This section shows how to compute the partial two–view motion solution *directly* from the affine epipolar geometry.

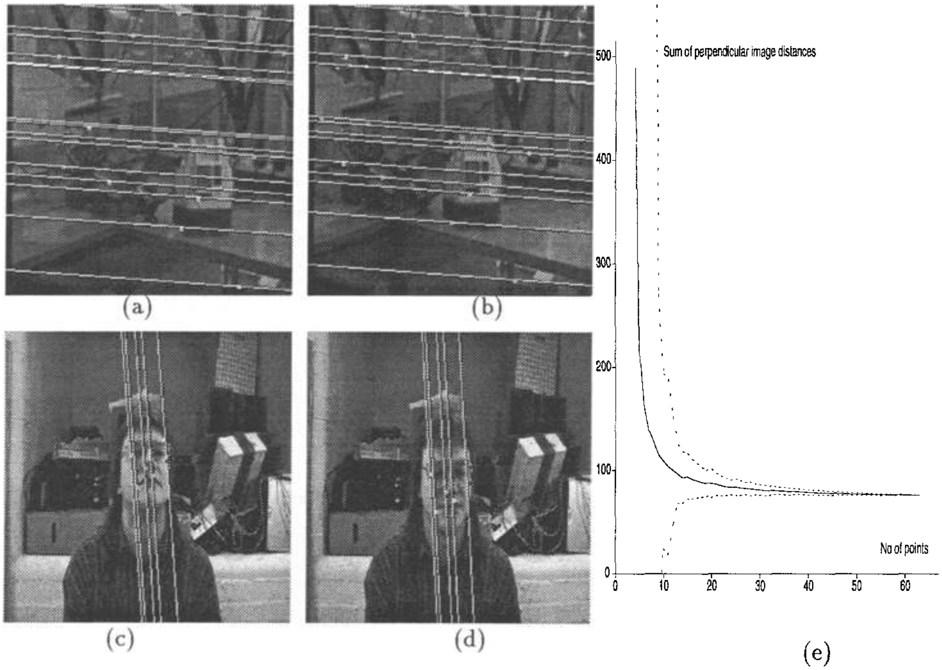


Fig. 2. Corner points with associated affine epipolar lines: (a)(b) The camera moves (every 10^{th} line shown); (c)(d) The object moves (every 2^{nd} line shown); (e) Improvement in the epipolar geometry as the number of points increases. The solid line shows median perpendicular distance between points and their epipolar lines and the dotted line shows the standard deviation (1σ level).

3.1 Previous work

Harris [7] used a weak perspective camera and the Euler angle representation to solve for rotation angles over two frames. The weak perspective form of E_2, whose shortcomings were outlined in Section 2.3, was minimised and shown to be *independent* of the turn angle out of the plane, illustrating the bas–relief ambiguity. No confidence estimates in the solution were provided, and only the projected axis was interpreted (not the cyclotorsion angle or scale). Koenderink and van Doorn [10] solved for the scale factor and the projections of the axes of rotation by observing a chosen local coordinate frame comprising 4 non–coplanar world points. Our scheme retains the underlying principles of their approach, but uses *all* available points and obviates the need to first define an affine basis. Lee and Huang [11] independently described the same technique as that of Koenderink and van Doorn.

Huang and Lee [9] assumed orthographic projection and proposed a linear algorithm to solve the equation $R_{23}\Delta x' - R_{13}\Delta y' + R_{32}\Delta x - R_{31}\Delta y = 0$ (a special case of the form given later in Equation (11)). Hu and Ahuja [8] criticised this approach, noting that the equation has only *two* independent unknowns, since $R_{13}^2 + R_{23}^2 = R_{31}^2 + R_{32}^2 = 1 - R_{33}^2$. Our formulation has *three* independent unknowns since we also cater for the scale factor s, making a linear solution valid.

None of the above authors [8, 9] noted that the projections of the axis of rotation could be found directly from R_{13}, R_{23}, R_{31} and R_{32}. Huang and Lee [9] deduced that two views yield a one–parameter family of motion (and structure) solutions, since R_{13}, R_{23}, R_{31} and R_{32} could only be recovered up to a scale factor.

3.2 Weak perspective epipolar geometry

Rigidity is imposed on the world motion parameters $\{\mathbf{A}, \mathbf{D}\}$ by requiring $\mathbf{A}$ to be a rotation matrix $\mathbf{R}$. This reduces the degrees of freedom in the motion parameters from 12 to 6. The use of *relative image coordinates* (or "difference vectors") cancels out translation effects, where the Δ notation denotes registration with respect to a designated reference point.

Three rotational degrees of freedom then remain. Since solving for $\mathbf{R}$ requires the measurement of angles (which are not affine invariants), it is necessary to use *weak perspective* cameras, $\mathbf{M}_{wp}$ and $\mathbf{M}'_{wp}$ (cf. Equation (1)). We introduce the scale factor $s = Z^c_{ave}/Z^{c'}_{ave}$ ($s > 1$ for a "looming" object) and define scaled depth $\Delta z_i = f \Delta Z^c_i / Z^c_{ave}$. The aspect ratios ξ and ξ' must be known in order to compute angles, and the ratio of focal lengths f/f' must be known (or unity if unknown) in order to determine scale. No other calibration parameters are needed. The rigid motion, difference–vector form of the affine epipolar constraint equation (Equation (4)) is then

$$\boxed{R_{23}\Delta x' - R_{13}\Delta y' + sR_{32}\Delta x - sR_{31}\Delta y = 0} \tag{11}$$

This equation generalises the pure orthographic forms ($s = 1$) derived by Huang and Lee [9] and used in [8]. There are only three independent degrees of freedom in Equation (11), since only the ratios of the coefficients may be computed; we show these to be the scale factor s and two rotation angles.

There are various ways to parameterise rotation angles, the most popular being Euler angles and the angle–axis form. Koenderink and van Doorn [10] introduced a novel rotation representation (which we term *KvD* and show in [19] to be a variant of Euler angles), and presented a geometric analysis of it. We formalise their representation algebraically to illustrate its advantages. In *KvD*, a rotation matrix $\mathbf{R}$ is decomposed into two parts, $\mathbf{R} = \mathbf{R}_\rho \, \mathbf{R}_\theta$. First, there is a rotation $\mathbf{R}_\theta$ in the image plane through angle θ (i.e., about the line of sight). This is followed by a rotation $\mathbf{R}_\rho$ through an angle ρ about a unit axis Φ lying in a plane parallel to the image plane and angled at ϕ to the positive X axis, i.e., a pure rotation *out of* the image plane. We write $\Phi = (\cos\phi, \sin\phi)^{\mathsf{T}}$.

The *KvD* representation has three main advantages. First, rotation about the optic axis provides no new information about structure, and it therefore makes sense to first remove this "useless" component. Second, it explicitly captures the depth–turn (or *bas–relief*) ambiguity in a way that the more popular angle–axis form doesn't – an advantage of Euler forms in general [7]. Third, it is elegant in that two views enable us to completely solve for two rotation angles (ϕ and θ), with the third (ρ) parameterising the remaining family of solutions. This contrasts with the angle–axis form, for which only one angle is obtained from two views, the two remaining angles satisfying a non–linear constraint equation [3]. The disadvantage of *KvD* is that the physical interpretation of rotation occurring about a single 3D axis is lost.

3.3 Solving for s, ϕ and θ

We now solve for the scale factor (s), the projection of the axis of rotation (ϕ) and the cyclotorsion angle (θ) directly from the affine epipolar geometry. Substituting the KvD expressions for R_{ij} into the epipolar constraint of Equation (11) gives

$$\boxed{\sin\rho\ [\cos\phi\ \Delta x'_i + \sin\phi\ \Delta y'_i - s\cos(\phi-\theta)\ \Delta x_i - s\sin(\phi-\theta)\ \Delta y_i] = 0} \qquad (12)$$

It is evident from Equation (12) that s, θ and ϕ can be computed directly from the affine epipolar geometry, because the difference vector form of Equation (4) is

$$a\Delta x'_i + b\Delta y'_i + c\Delta x_i + d\Delta y_i = 0,$$

and a direct comparison with Equation (12) yields

$$\tan\phi = b/a, \quad \tan(\phi-\theta) = d/c \ \text{ and } \ s^2 = (c^2+d^2)/(a^2+b^2), \qquad (13)$$

with $s > 0$ (by definition). This illustrates, for instance, that *the projection of the axis of rotation Φ is perpendicular to the epipolar lines.* (Recall from Figure 1, for instance, that $\mathbf{n}_2 = (a,b)^{\mathsf{T}}$ is the normal to the epipolar line in I_2.) Equation (12) also shows immediately that Equation (11) has only *two* independent rotation parameters, θ and ϕ, because the angle ρ cancels out (provided it is non–zero). If $\rho = 0°$, there is no rotation *out of* the image plane and Φ is obviously undefined, so this technique cannot be used. Equation (12) is therefore more informative than Equation (11) since it identifies explicitly what quantities can be computed, and under what circumstances.

Error model and Kalman filter We now compute noise models for s, ϕ and θ, each of which is a non–linear function of $\mathbf{n}$. Given the covariance matrix $\Lambda_{\mathbf{n}}$ from Equation (10), the task is to compute the means and variances of s, ϕ and θ. Let the true (i.e. noise–free) value of $\mathbf{n}$ be $\tilde{\mathbf{n}}$, with $\mathbf{n} = (n_1, n_2, n_3, n_4)^{\mathsf{T}}$. The noise perturbation of $\tilde{\mathbf{n}}$ is $\delta\mathbf{n}$, so $\mathbf{n} = \tilde{\mathbf{n}} + \delta\mathbf{n}$. The diagonal elements of $\Lambda_{\mathbf{n}}$ define the variances of δn_i while the off–diagonal elements define the covariances. The Taylor series for a function $q(\tilde{\mathbf{n}})$ expanded about $\mathbf{n}$ is

$$q(\tilde{\mathbf{n}}) = q(\mathbf{n} - \delta\mathbf{n}) = q(\mathbf{n}) - \sum_{i=1}^{4} \frac{\partial q(\mathbf{n})}{\partial \tilde{n}_i}\, \delta n_i + \frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4} \frac{\partial^2 q(\mathbf{n})}{\partial \tilde{n}_i\, \partial \tilde{n}_j}\, \delta n_i\, \delta n_j - \cdots$$

We ignore terms above second order, assume that $\partial^2 q/\partial n_i\, \partial n_j = \partial^2 q/\partial n_j\, \partial n_i$, and note that $E\{\delta\mathbf{n}\} = \mathbf{0}$ and $E\{\tilde{\mathbf{n}}\} = \tilde{\mathbf{n}}$. The estimate of q is in general biased, since $E\{q(\mathbf{n})\} = q(\tilde{\mathbf{n}}) - B$, with the bias term $B = \frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4} \frac{\partial^2 q(\mathbf{n})}{\partial \tilde{n}_i\, \partial \tilde{n}_j}\, \Lambda_{ij}$. Expressions for the variance and covariances of q can then be derived, and these provide confidence regions for the two–frame motion parameters.

Physical objects have inertia and it is sensible to exploit this temporal continuity to improve the motion estimates. We achieve this by means of a linear discrete–time Kalman filter [2], a popular framework for weighting observations and predictions. We estimate s, ϕ and θ, employing a constant position model ($\dot{s} = \dot{\phi} = \dot{\theta} = 0$). The state vector is $(s, \phi, \theta)^{\mathsf{T}}$ with state transition matrix $\mathbf{I}_3$. We observe s, ϕ and $\phi - \theta$, giving the observation vector $(s + B_s, \phi + B_\phi, \phi - \theta + B_{\phi-\theta})^{\mathsf{T}}$, where B_i are the relevant bias terms.

Results Figure 3 shows a subject shaking his head. The true axis is unknown, but it is approximately vertical and the results are qualitatively correct. Figure 4 shows the algorithm running on the images and corner data of Harris [7], where a car rotates on a turn–table about a known fixed axis. There is no scale change between views, and the fiducial axis is 10° off the vertical. Figure 4(a) graphs the successive two–frame estimates of the projected axis angle together with the computed errors, which serve as the filter input. Our unfiltered solution (using E_3) is identical to the Harris values (obtained using E_2); this will always be true when the scale s is unity (see Shapiro et al. [18]). The error estimates correctly bound the true parameter values (which lie within the computed 95% error bounds). The filtered output is shown in Figure 4(b) with the Kalman filter's 95% confidence intervals. The solution is clearly smoother (and more reliable) after filtering.

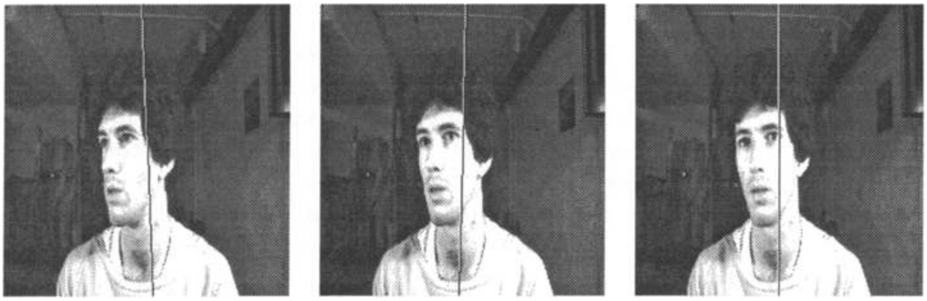


Fig. 3. *A shaking head, where the true axis is roughly vertical. The computed axis is drawn through the image centre in both black and white to enhance contrast.*

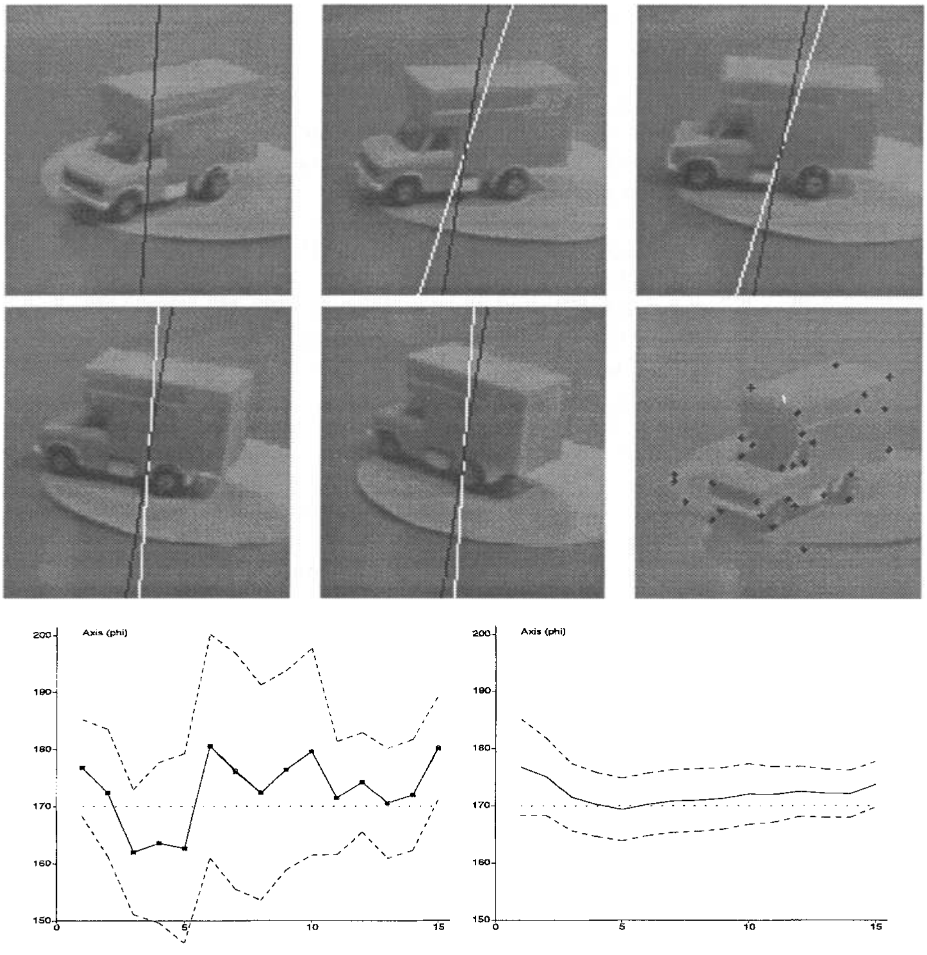
4 Conclusions

We have proposed a new framework, based on the affine camera and its epipolar geometry, for computing motion from point features viewed under parallel projection. This framework accounts for the major theoretical results pertaining to this problem [3, 7, 9, 11, 21, 22], including partial solutions, ambiguities and degeneracies [18]. The affine camera enables the identification of necessary camera calibration parameters, and the facility to use all available points both ensures robustness to noise and obviates the need to choose a local coordinate frame. Noise models provide confidence estimates in the computed parameters, and the processing of successive frame–pairs permits straightforward extension to long sequences in sequential mode.

Acknowledgements. We are grateful for financial support from the ORS Award, the Foundation for Research Development (RSA) and Esprit BRA VIVA. We have had fruitful discussions with Paul Beardsley, Andrew Blake, Phil McLauchlan, Ian Reid and Phil Torr (RRG) along with Richard Hartley and Joe Mundy (GE). Chris Harris kindly supplied his images and corner points.

References

1. J.Y. Aloimonos, "Perspective approximations", *Image and Vision Computing,* Vol. 8, No. 3, August 1990, pp. 179–192.



(a) (b)

Fig. 4. The Harris truck sequence [7] (every second image). The unfiltered (white) and filtered (black) axes are superimposed. The final image shows a typical set of corner points. The graphs plot the solutions over the whole sequence: solid lines show computed values, dotted lines show "true" values, and dashed lines show 95% confidence intervals: (a) Our two-frame solution (circles) coincides with that of Harris (crosses); (b) The improved filtered values (dashed lines show 95% confidence levels).

2. Y. Bar–Shalom and T.E. Fortmann, *Tracking and data association,* Academic Press Inc., USA, 1988.

3. B.M. Bennett, D.D. Hoffman, J.E. Nicola and C. Prakash, "Structure from two orthographic views of rigid motion", *Journal of Optical Society of America,* Vol. 6, No. 7, July 1989, pp. 1052–1069.

4. D. Charnley, C. Harris, M. Pike, E. Sparks and M. Stephens, "The DROID 3D vision system: algorithms for geometric integration", Plessey Research, Roke Manor, Technical Note 72/88/N488U, Dec. 1988.

5. R. Cipolla, Y. Okamoto and Y. Kuno, "Robust structure from motion using motion parallax", *International Conference on Computer Vision (ICCV'4),* Berlin, May 1993, pp. 374–382.

6. O.D. Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig?" in G. Sandini (ed.), *Proceedings European Conference on Computer Vision* (ECCV–92), 1992, pp. 563–578.
7. C. Harris, "Structure–from–motion under orthographic projection", *First European Conference on Computer Vision* (ECCV–90), 1990, pp. 118–123.
8. X. Hu and N. Ahuja, "Motion estimation under orthographic projection", *IEEE Transactions on Robotics and Automation,* Vol. 7, No. 6, pp. 848–853, 1991.
9. T.S. Huang and C.H. Lee, "Motion and structure from orthographic projections", *IEEE Trans. Pattern Anal. Machine Intell.,* Vol. PAMI–11, No. 5, pp. 536–40, 1989.
10. J.J. Koenderink and A.J. van Doorn, "Affine structure from motion", *Journal of Optical Society of America,* Vol. 8, No. 2, Feb 1991, pp. 377-385.
11. C. Lee and T. Huang, "Finding point correspondences and determining motion of a rigid object from two weak perspective views", *Computer Vision, Graphics and Image Processing,* Vol. 52, 1990, pp. 309–327.
12. H.C. Longuet–Higgins, "A computer algorithm for reconstructing a scene from two projections", *Nature,* Vol. 293, 1981, pp. 133–135.
13. Q–T. Luong, R. Deriche, O. Faugeras and T. Papadopoulo, "On determining the fundamental matrix: analysis of different methods and experimental results", Tech. Report 1894, INRIA (Sophia Antipolis), April 1993.
14. J.L. Mundy and A. Zisserman (eds), *Geometric Invariance in Computer Vision,* MIT Press, USA, 1992.
15. L. Quan and R. Mohr, "Towards structure from motion for linear features through reference points", *IEEE Workshop on Visual Motion,* New Jersey, 1991.
16. I.D. Reid, "The SVD minimizes image distance", Oxford University Robotics Research Group Internal Memo, Sept. 1993.
17. L.S. Shapiro and J.M. Brady, "Rejecting outliers and estimating errors in an orthogonal regression framework", to appear in *Philosophical Transactions of the Royal Society,* 1994.
18. L.S. Shapiro, *Affine Analysis of Image Sequences,* PhD thesis, Dept. Engineering Science, Oxford University, 1993.
19. L.S. Shapiro, A. Zisserman and M. Brady, "Motion from point matches using affine epipolar geometry", to appear in *International Journal of Computer Vision.*
20. D.W. Thompson and J.L. Mundy, "Three dimensional model matching from an unconstrained viewpoint" in *IEEE Conference on Robotics and Automation,* Raleigh, NC, 1987, pp. 208–220.
21. C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method", *International Journal of Computer Vision,* Vol. 9, No. 2, Nov 1992, pp. 137–154.
22. S. Ullman, *The Interpretation of Visual Motion,* MIT Press, USA, 1979.
23. G. Xu, E. Nishimura and S. Tsuji, "Image correspondence and segmentation by epipolar lines: Theory, Algorithm and Applications", Technical Report, Dept. Systems Engineering, Osaka University, July 1993.
24. A. Zisserman, *Notes on geometric invariance in vision: BMVC'92 tutorial,* Leeds, Sept 1992.