Robust Egomotion Estimation

from Affine Motion Parallax

Jonathan M. Lawn and Roberto Cipolla

CUED/F-INFENG/TR 160

January, 1994

A condensed version of this paper
will be presented at ECCV'94

Department of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
England

Email: jml@eng.cam.ac.uk

Robust Egomotion Estimation from Affine Motion Parallax

Jonathan M. Lawn and Roberto Cipolla
Department of Engineering,
University of Cambridge,
Cambridge CB2 1PZ, England.

Abstract

A method of determining the motion of a camera from its image velocities is described that is insensitive to noise and the intrinsic camera parameters and requires no special assumptions. This algorithm is based on a novel extension of motion parallax which does not require the instantaneous alignment of features, but uses sparse visual motion estimates to extract the direction of translation of the camera directly, after which determination of the camera rotation and the depths of the image features follows easily. A method for calculating the expected uncertainty in the estimates is also described which allows optimal estimation and can also detect and reject independent motion and false correspondences. Experiments using small perturbation analysis show a favourable comparison with existing methods, and specifically the Fundamental Matrix method.

Keywords: Structure from Motion, Parallax, Epipole, Egomotion, Uncertainty

1 Introduction

The visual motion of a scene from a camera provides a cue for the determination of the scene structure and the viewer motion (*egomotion*). If the camera motion is known to a reasonable accuracy, only the *depths* in the scene need be computed (the distances from the camera centre to the points in the scene). However, many systems would benefit from using a freely moving camera with no external motion sensors, which requires that the visual motion is decomposed to give the camera motion and the scene structure.

An important criterion for the judging of algorithms is *robustness*: insensitivity to, and graceful failure from, image motion measurement noise, independent motion and erroneous measurements, sparse image motion data and narrow field of view. There are many methods for estimating the camera motion from an image pair or sequence. Some methods use higher derivatives of the image motion than just the velocities [1, 34], but they are therefore susceptible to measurement noise [28] and are not considered further here. Many try to minimise a function for all five egomotion variables[1] [28], including the Essential Matrix method [22, 11, 36], or even the egomotion and depth variables [18, 27, 32, 33]. However, as well as being computationally expensive, searches over high dimensional spaces also appear to produce more false minima. Others find only approximate solutions for one egomotion component, either the rotation [29] or the translation [2, 17], by assuming that it is dominant. This is much more direct, but the solutions tend to be heavily biased when the other component is significant.

The *Fundamental Matrix* method [10] attempts to find the *epipole*, the intersection of the direction of motion with the imaging surface, independent of intrinsic camera parameters, by finding the linear function that produces the epipolar lines that fit the observed point matches best in both images. (The method of forming the matrix is identical to the Essential Matrix method, though the interpretation is less constrained.) This involves estimating a 3×3 singular homogeneous matrix (7 independent variables) – further description follows in section 4.2. Heeger

[1] The egomotion is described by the camera rotation and translational direction. The absolute magnitude of the camera translation cannot be found, because of the speed-scale ambiguity, though it can be expressed in terms of the scene depths.

and Jeepson's Subspace method [12] also attempts to find the epipole initially, by a two-dimensional search of a non-linear function.

Motion parallax methods [23, 30, 14] try to find the epipole (2 variables) independent of rotation and *intrinsic camera calibration* (knowledge of the projection plane centre, the aspect ratio and the focal length, assuming no other distortions are significant) by using the fact that the relative visual motions of points instantaneously coincident in the image can be used to cancel the rotational part of the visual motion. This pure translation component will be towards or away from the epipole and therefore the epipole can be determined from two or more of these *epipolar line* constraints. Once the epipole is known, the camera rotation can be found from the component of the visual motion orthogonal to the epipole. Unfortunately motion parallax has many practical shortcomings, particularly the need for dense velocity field measurements at sudden depth changes [14, 30].

This paper describes a method that will provide the parallax constraints, independent of the camera rotation and point depths, without the need for the instantaneous alignment of features or dense image velocity measurements (section 2). A method of calculating the uncertainty of the constraints and of optimally combining them is also presented (section 3). An empirical comparison is then made with other significant methods (section 4).

2 Theoretical Framework

This algorithm measures motion parallax without requiring the instantaneous alignment of features by using affine (linear) approximations to projection. Below is given an introduction to the notation, a description of motion parallax and affine image motion, and then the algorithm that combines them. It is then shown how this can be used to find the epipole, independent of the intrinsic camera parameters and from only sparse image velocity measurements, and then infer the complete egomotion.

2.1 Image motion

Stationary *feature* points in space are given in the camera coordinate system by vectors $\mathbf{X}$ measured from the camera (projection) centre. However only the directions of these vectors can be measured by a camera from a single position, and so the image positions of the points are represented here by the vectors, $\mathbf{Q}$, the intersections of $\mathbf{X}$ with the unit sphere. An alternative is to use a projection plane, which is less elegant because it involves the introduction of an extra variable, the plane normal, in later equations. *However all techniques used here are equally applicable to image hemispheres or planes, and to the motion being represented by discrete displacements or velocities.*

$$\mathbf{Q} = \frac{\mathbf{X}}{|\mathbf{X}|} \tag{1}$$

As the camera moves with translational velocity $\mathbf{U}$ and angular velocity $\mathbf{\Omega}$ (also in camera coordinates), the relative positions of the points in space change

$$\dot{\mathbf{X}} = -\mathbf{\Omega} \times \mathbf{X} - \mathbf{U} \tag{2}$$

($\times$ is the symbol for the vector product), and this induces image motion. Differentiation of (1) and substitution into (2) gives [5, 26]

$$\dot{\mathbf{Q}} = -\mathbf{\Omega} \times \mathbf{Q} - (\mathbf{Q} \times \frac{1}{r}\mathbf{U}) \times \mathbf{Q} \tag{3}$$

where $\dot{\mathbf{Q}}$ is the image motion, and $r = |\mathbf{X}|$, the feature depth. This clearly shows that the visual motion is made up of two components, one depending on the translational velocity of the camera and the scene structure (depths), and one depending on the rotational velocity and only the image positions [3]. It is this separation of image velocity measurements into components that is difficult, although numerous methods have been proposed.

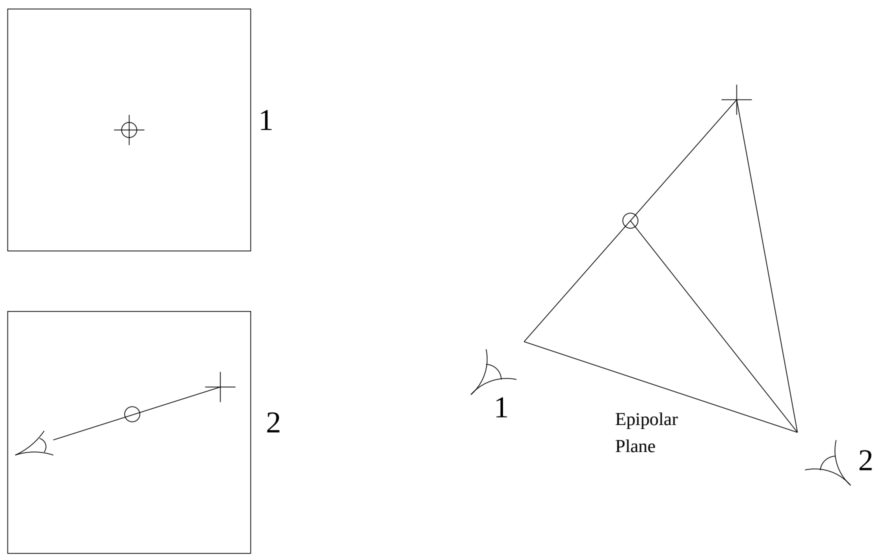


Figure 1: Motion Parallax
If two points in space, the circle and the cross, are coincident in one image, this means that the camera position is collinear with them in space. Therefore if all three are viewed from a second camera position, they will appear collinear, and even if the image of the first camera position (which gives the direction of travel of the camera, the *epipole*) is not known, it is constrained to lie on the *epipolar line* given by the images of the two points.

2.2 Motion Parallax

If two points in space, $\mathbf{X}_a$ and $\mathbf{X}_b$, project to the same point on the imaging sphere momentarily (ie. $\mathbf{Q}_a = \mathbf{Q}_b = \mathbf{Q}$), then their image velocities will have identical rotational components (from 3), and the difference of their image velocities, $\dot{\Delta}_{\mathbf{Q}}$, will depend only on the translational velocity of the projection centre and on their image position and depths. This is called *motion parallax* [13, 23] (see Figure 1).

$$\dot{\Delta}_{\mathbf{Q}} \equiv (\dot{\mathbf{Q}}_a - \dot{\mathbf{Q}}_b) \text{ where } \mathbf{Q}_a = \mathbf{Q}_b = \mathbf{Q} \tag{4}$$

$$= (\mathbf{Q} \times \mathbf{U}) \times \mathbf{Q}\left(\frac{1}{r_b} - \frac{1}{r_a}\right) \tag{5}$$

The *epipole*, $\mathbf{Q}_e$ is defined by the direction of the camera velocity vector $\mathbf{U}$: it follows that $\mathbf{Q}_e$ and $\mathbf{U}$ are parallel. Substituting this into Equation 5 gives:

$$(\dot{\Delta}_{\mathbf{Q}} \times \mathbf{Q}).\mathbf{Q}_e = \mathbf{n}_{\mathbf{Q}}.\mathbf{Q}_e = 0 \tag{6}$$

This implies that the epipole is constrained by each measurement of $\dot{\Delta}_{\mathbf{Q}}$ to lie on an *epipolar plane* in space[2] passing through the projection centre and having normal $\mathbf{n}_{\mathbf{Q}} = \dot{\Delta}_{\mathbf{Q}} \times \mathbf{Q}$. Therefore two motion parallax measurements in different parts of the visual field can determine the epipole. Unfortunately implementation requires sudden depth changes and the measurement of the visual motion of instantaneously aligned image points. The solution will be incorrect or ill-conditioned if:

- the points used to calculate the parallax are not close to coincident,
- there is no significant depth change between the points, giving only a small velocity difference, or

[2] which produces a great circle of the unit projection sphere or an *epipolar line* on the projection plane

• the field of view is too small to allow good triangulation on the epipole.

How can we compute motion parallax when the features are not aligned? We propose a solution that exploits the linear motion of features in a small neighbourhood.

2.3 Affine Image Motion

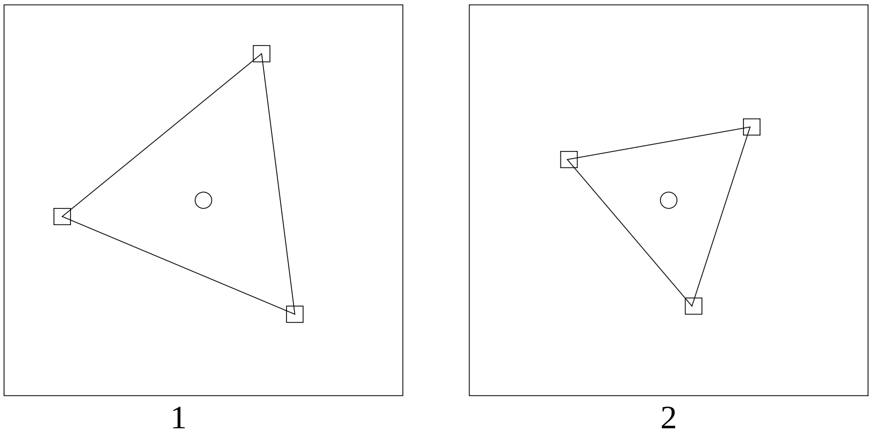


Figure 2: Affine Image Motion
If the image of a rigid plane in space is deforming linearly, then that deformation can be determined from the motions of three *basis* points in that plane (the squares). The image motion of any forth point in the plane (the circle) due to this deformation (shear, rotation and expansion) can be predicted from the motion of the three basis points.

Weak perspective [31] makes the assumption that there is a linear transformation between the positions of the points in 3D and their projections. From small perturbation analysis of (3), it can be seen that the image velocity is in general a non-linear function of both the image position, $\mathbf{Q}$, and the scene depth, r, and therefore, for the visual motion to vary linearly (*affinely*), the visual region considered must be small and contain small depth changes.

Under these conditions, a point in a certain plane in space will have an image velocity that is a linear function of its image position only. This *affine transformation* can be determined from, for instance, the visual motion of a minimum of three points in the plane (see Figure 2), or the deformation of a closed curve in the plane [4].

2.4 Affine Motion Parallax

Conventional motion parallax uses two points instantaneously aligned, but *affine motion parallax* [6, 19] uses four nearby points, not necessarily on the same object or surface (see Figure 3). The motion parallax comes from one real *parallax* point, and one *virtual* point, taken to be momentarily aligned with the parallax point in the image, but at a different depth. The movement of this virtual point is calculated by assuming it is on a plane defined by a *basis* of three other nearby points. *We have computed motion parallax even though the features were not aligned by using the weak perspective approximation.* Weak perspective is valid as long as the region it is applied to is small in comparison with the distance from the camera [6].[3] The algorithm allows much sparser features to be used than conventional motion parallax methods, and it is valid equally for discrete displacements or velocities. This implementation assumes that plenty of useful parallax measurements will be produced. However once depth estimates for the points are available (by using information from previous frames), then bases can be chosen that not only span the image well, but also do

[3] There is a less intuitive but stronger homogeneous formulation presented in Appendix B, which shows that, in this case, a sufficient condition is that the region is small in the image.

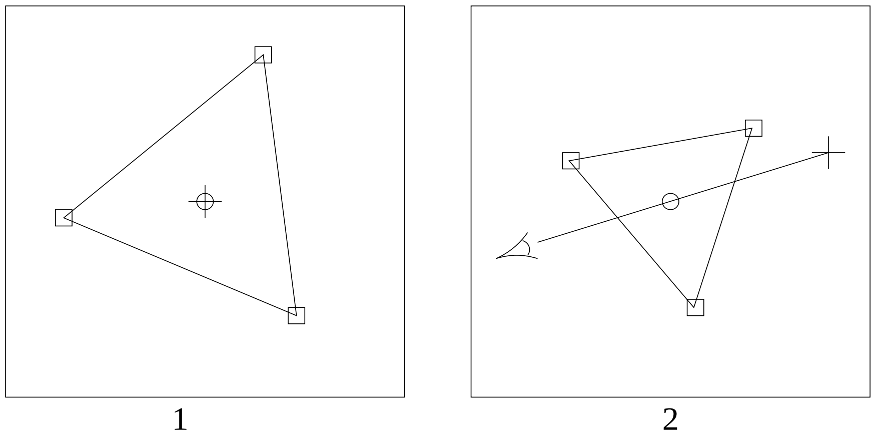


Figure 3: The Affine Motion Parallax Algorithm
In view one, there are four real feature points, the *parallax* point (the cross), and the three *basis* points (the squares). The circle is an imaginary *virtual* point that is defined to be coincident in the image with the parallax point, but on the plane in space defined by the basis points. In view two, the motion of the parallax point is observed, and that of the virtual point is deduced from the basis points, assuming the plane that they define is deforming affinely in the image. We have therefore recovered exact motion parallax, the relative motion of the virtual point and the parallax point, from four sparse points [6, 19].

not contain excessive depth changes that might jeopardise the affinity. In addition, parallax points that provide significant parallax velocities can be chosen.

2.5 Epipole Determination

Epipole extraction methods that attempt to extract the epipole that rely on small global perspective effects tend to suffer badly with noisy data that masks them. Longuet-Higgins and Prazdny [23] showed, as above, that two or more local measurements of motion parallax can determine the epipole but true parallax is difficult to measure. Using the weak perspective assumption globally will result in parallel motion parallax vectors [19], as it is only valid for small fields of view. However weak perspective assumptions can be used in a number of small regions of a wide view.

Here we propose a method that extends the use of affine motion parallax to a number of small neighbourhoods in a full perspective image, so that the relative velocities can be used to determine the epipole. The epipole can be obtained entirely from geometric construction on the image plane, though here it is mapped onto the sphere to avoid infinities. Each neighbourhood of at least four nearby points generates a motion parallax vector which provides a great circle constraint on the imaging sphere on which the epipole must lie (6), using only sparse visual motion estimates. Two or more separated cases can determine the epipole exactly at the intersection of their constraints (see Figure 4).

3 Uncertainty and Errors

For real, noisy data the epipole constraints will not intersect. A method is needed that will optimally integrate the constraints and reject the outliers due to independent motion and incorrect visual motion measurements .

3.1 Estimating Uncertainty

The feature measurement errors on the image plane determine the image velocity measurement uncertainty, which then in turn affects the rest of the calculation through to the motion parallax vectors and then the epipole estimate. The uncertainty on each parallax constraint will depend

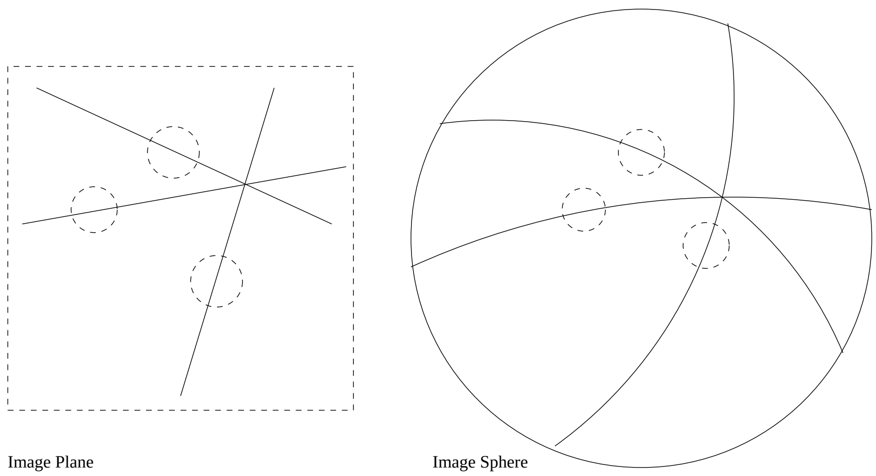


Figure 4: Finding the Epipole
Illustrations of the epipole being given by the intersection of three cases of (affine) motion parallax, on the image plane and the image hemisphere. In reality, errors will mean that the (affine) motion parallax constraints do not quite intersect (see Section 3).

not only on the image point location noise, but also on the magnitude of the motion parallax and the geometry of the affine basis (collinear basis points will not define the affine deformation well). If all the image positions are represented as normalized 3D vectors to the projection sphere, then all the calculations can be expressed in standard vector arithmetic. Assuming that the uncertainty on each vector (and scalar) in the calculations is small, additive and independent, then it can be represented by a covariance matrix [15, 20], calculated at the same time as the estimated value. The significance of the motion parallax vectors can now be defined in terms of their uncertainty estimates. For instance, if any point is found to have no significant affine motion parallax, it may well be in the plane of the three points providing its virtual pair, and a new basis can be chosen if necessary.

The arithmetic calculations assume the independence of all estimates, and this is generally true if some care is taken. However where estimates or constraints coming from shared original measurements are combined, independence is lost. Here, the weights calculated for the constraints assume their independence, and therefore in reality will be only approximately optimal. Also the scale of the uncertainty may be wrong, but this can be compensated for by looking at the degree of overuse of the measurements, though often many constraints are rejected as excessively noisy so the independence assumption is a good approximation.

3.2 Optimal Constraint Combination

A number of different methods for combining constraint planes of differing uncertainties to find the epipole (ie. for finding a common normal for the constraint plane normals) have been proposed. Some (eg. Hildreth [14]) have used Hough space methods, and have found the maximum of a discrete histogram of the constraints on the sphere, with each 'bin' scoring for each constraint plane that approximately intersects it, but achieving high accuracy and producing an uncertainty estimate would require a very dense array of bins and therefore excessive computation[4]. Other methods have been investigated here.

1. **Closed Form Solution** One method uses the weighted average of the normalised vector products of every possible pair of constraint plane normals, $\mathbf{n}_i \times \mathbf{n}_j$.[5] If the trace of the co-

[4]Considerable effort has been expended attempting to improve the properties and speed of Hough Space Techniques [21].

[5]This estimate has also been made in other work in the context of finding vanishing points [25].

variance matrix of the epipole estimate (ie. the sum of the eigenvalues) is chosen as the scalar measure of uncertainty to be minimised, then the scalar uncertainty of the epipole estimate is just the sum of the scalar uncertainties of the component vector products, and the reciprocal of this scalar uncertainty is its optimal weight for each component (see Appendix A).

2. **Iterative Solution** A better solution is the estimate $\mathbf{e}$ of the epipole $\mathbf{Q}_e$ that gives the least sum of the square scalar products between itself and the constraint plane unit normals, $\mathbf{n}_i$ [15, 7], given that $\mathbf{e}^T\mathbf{e} = 1$. This is equivalent to finding the eigenvector associated with the smallest eigenvalue of a matrix, N, formed by the sum of the outer products of the constraint normals, $\mathbf{n}_i.\mathbf{n}_i^T$.

$$\min \sum_i w_i(\mathbf{e}.\mathbf{n}_i)^2 \quad = \quad \min\ \mathbf{e}^T \left(\sum_i w_i \mathbf{n}_i.\mathbf{n}_i^T \right) \mathbf{e} = \min\ \mathbf{e}^T N \mathbf{e} \tag{7}$$

   However, to minimise the uncertainty of the estimate by weighting the sum, an *a priori* estimate of the epipole is needed so that the scalar weight can be determined only by the significant component of the uncertainty (parallel to the epipole). Again the weighting should be proportional to the inverse of the scalar variance. Since the constraint uncertainties are often highly anisotropic, many iterations are often necessary.

   An attractive feature of this form of epipole estimation is that the weighted outer product sum encodes the estimate, $\mathbf{e}$, and the variance (uncertainty) of the estimate, $V[\mathbf{e}]$.

$$C \quad = \quad \sum_i C_i = \sum_i \frac{1}{\sigma_i^2} \mathbf{n}_i\, \mathbf{n}_i^T \tag{8}$$

   where $\sigma_i^2 = \mathbf{e}^T V[\mathbf{n}_i]\mathbf{e}$ is the variance of the constraint normal in the (estimated) direction of the epipole, and $\mathbf{e}^T\mathbf{e} = 1$. The estimate of the epipole, $\mathbf{e}$, minimises $\mathbf{e}^T C\mathbf{e}$ (ie. $\mathbf{e}$ is the unit eigenvector associated with the minimum eigenvalue of C) and $V[\mathbf{e}] = C^{-1}$.

3. **Unbiased Solution** Bias in the epipole estimate towards the centre of the field of view has been noted by a number of researchers using a variety of methods. Daniilidis and Nagel [8] showed that the linear Fundamental method was intrinsically biased, but bias has not been noted for the non-linear Fundamental matrix formulations, as implemented as a comparison here (see Section 4).

   Methods using the combination of motion parallax constraints have also been found to have an inward bias [14], and Kanatani [16] has shown that this can be accounted for as a statistical bias in the estimator.

$$N \quad = \quad \sum_i w_i \mathbf{n}_i.\mathbf{n}_i^T \tag{9}$$

$$E[\Delta N] \quad = \quad E\left[\sum_i w_i \left(\Delta\mathbf{n}_i.\mathbf{n}_i^T + \mathbf{n}_i.\Delta\mathbf{n}_i^T + \Delta\mathbf{n}_i.\Delta\mathbf{n}_i^T \right) \right] \tag{10}$$

$$\quad = \quad \sum_i w_i V[\mathbf{n}_i] \neq 0 \tag{11}$$

   where $E[\Delta\mathbf{n}_i] = 0$. He proposes the use of another unbiased estimator, $\hat{N}$ (*renormalisation*).

$$\hat{N} \quad = \quad \sum_i w_i \left(\mathbf{n}_i.\mathbf{n}_i^T - V[\mathbf{n}_i] \right) \tag{12}$$

$$E[\Delta\hat{N}] \quad = \quad 0 \tag{13}$$

Section 4.2 describes the findings of our tests of these methods.

3.3 Robust Rejection

Independent motion is when a point in the scene is not stationary in 3D space, and therefore has an image motion that is not due entirely to its position in space and the camera motion. Incorrect visual motion measurements (caused by bad matching or tracking of features, or by independently moving objects or spurious features) can be removed at the constraint fusion stage. One method is to *build up* the estimate by starting from the best estimate available from two constraints and combining only those constraints that *agree*. This agreement can be stated in terms of the likelihood known from the uncertainties calculated for each constraint using a $\mathcal{X}^2$ test and the Mahalanobis distance. If too few constraints agree then a new initial estimate is needed.

A different method is currently used, which disassembles the estimate using *take-one-out cross-validation* [35]. The initial estimate is formed using all constraints, and then each constraint is removed in turn and is checked against an estimate formed by the others. The constraint that agrees least is removed and the process repeated, until all constraints agree sufficiently.

All those features that were not part of any included (affine motion parallax) constraint can be rejected as corrupted and, if necessary, the calculation can be repeated to improve its accuracy (by producing more constraints that are acceptable) or the label can be passed on to improve the calculation accuracy in future frames. Other ways of recognizing a bad feature include checking that its calculated depth is positive and changing smoothly.

3.4 Egomotion and Feature Depths

So far all calculations have been independent of the intrinsic camera parameters (despite the representation of the image on a unit sphere [25]). To find the camera rotation and the feature depths these values should be known. There will be inconsistencies if incorrect intrinsic parameters are used (and visual calibration methods exist) but they are likely to be small compared to the noise. However, even if the intrinsic camera parameters are not known, the scene structure can still be extracted to within a 3D projective transformation once the epipole is found [9].

Assuming we have knowledge of the intrinsic parameters, the camera rotation can be found by a least-squares estimate from the component of visual motion perpendicular to the direction of camera translation [12]. This requires a minimum of three points. From Equation (3),

$$\begin{aligned} \text{Define } \mathbf{m} &= \mathbf{Q}_e \times \mathbf{Q} && (14) \\ \mathbf{m}^T \dot{\mathbf{Q}} &= -\mathbf{m}^T (\boldsymbol{\Omega} \times \mathbf{Q}) && (15) \\ &= (\mathbf{m} \times \mathbf{Q})^T \boldsymbol{\Omega} && (16) \end{aligned}$$

Then the scene depths can be determined from the component of visual motion towards the epipole not accounted for by the camera rotation.

The scene structure can be stored from frame to frame to help track features in the image, choose affine regions, detect independent motion and estimate the depths accurately. Egomotion can also be predicted, though rotation in particular is likely to have a very high noise component. Depth estimates, as well as being fed back to improve the affine basis selection procedure, could also be used to verify the intrinsic calibration by comparison with the rotation and depth calculations.

4 Experimental Comparisons

In this section, we test the algorithm described above and compare it with an existing technique, the Fundamental Matrix method. These experiments are to test attempts at interpreting visual motion, and therefore we shall assume that a number of 'corner' features can be tracked between frames.

4.1 Conditions

The features used are hand-picked corners with significant amounts of noise (including integer quantization) taken from low resolution video sequences from a hand-held camera (600×500 pixels).

These show outdoor scenes with a consistent camera translation direction but high random camera rotation. No detectable independently moving points are included.

4.2 Implementations

Affine Motion Parallax Method

Each detected corner is taken as the parallax point for a constraint vector, using its three nearest neighbours as the affine basis points. The calculations are formulated using the image hemisphere and velocities rather than discrete displacements. Constraints with excessive errors are rejected.

The three constraint combination methods mentioned in section 3.2 were implemented and tested. The 'closed form' method was implemented first, but it was found that the optimisation criterion tends to produce undesirable results: the statistical dependence proved significant and the algorithm produced an epipole estimate with highly anisotropic uncertainty. The 'unbiased' method was also tested. However, this has a number of disadvantages. Firstly, the variance of the constraints must be known quantitatively, and not just relative to one another. (These variances can be found within the same iterative scheme that adjusts the weights, w_i [16].) A second, more important problem is that though the estimate becomes less biased, its variance will increase, because the variances of the constraints, $V[\mathbf{n}_i]$, will be less certain that the constraint components, $\mathbf{n}_i\mathbf{n}_i^T$. Experimentally, renormalisation has been shown to be detrimental to the results of our algorithm, and therefore the 'iterative' technique was used.

Fundamental Matrix Method

This implementation scheme uses a non-linear method (DIST-L [24]) which aims to minimise the distance from the points to their epipolar lines in both images of the pair, given that the transformation is singular, by varying eight homogeneous coefficients. First order perturbation analysis is used to determine the uncertainties. A good initial estimate of the correct Fundamental Matrix is required if the iterative minimisation is to converge to a good solution.

4.3 Results

The 'Kings College Chapel North Gate' Sequence

This sequence of frames shows the case of translation in approximately the direction of the camera axis (towards the chapel gates) and large arbitrary rotations. As the sequence progresses, the number of features giving the image motion estimates decreases. Figure 5 shows a specific example of a sparse set of points between two frames (the third and forth) and Figure 6 shows the sequence and the epipoles found.

Despite the fact that the point set becomes very sparse from the third frame, the affine motion parallax method consistently finds the epipole with high accuracy and reasonable certainty (Figure 5). Despite the lack of intrinsic camera calibration, the translational image motion components calculated, which are the last stage of the decomposition of the image motion, intersect at the epipole well and have qualitatively correct magnitudes (see equation 3). The predicted uncertainty seems to regularly overestimate the truth (as found by small perturbation analysis), which may be due to the dependence of the constraints being less significant than anticipated.

It is interesting to note that all the independently moving objects in this sequence (including those without selected features) are undetectable. This is because they are all moving in their epipolar plane. Other constraints, such as that the objects rest on the ground, must be applied to interpret the scene correctly.

The Fundamental Matrix method is accurate in this sequence, when there are a fair number of features, as long as it is initialised well. However it becomes very unstable as the feature set becomes sparse (Figure 9).

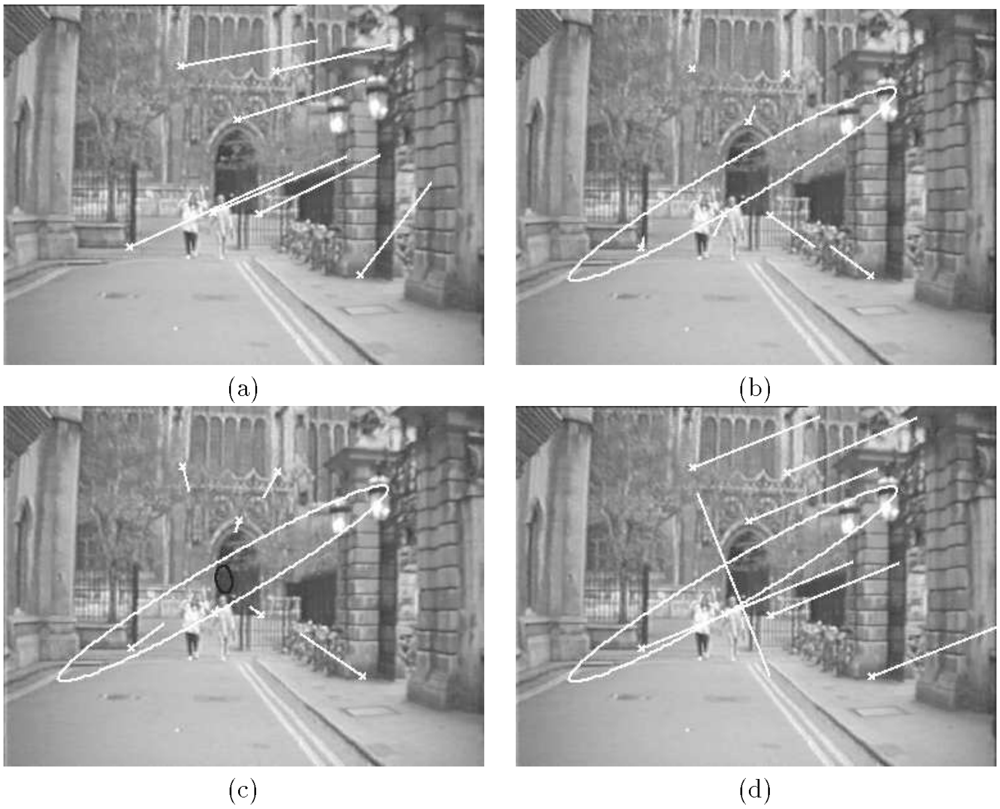


Figure 5: The approach to Kings College Chapel North Gate, frames (3-)4 using Affine Motion Parallax. Even given very few image motion measurements (a), the algorithm extracts five good constraints (b) (length indicating certainty), and makes an accurate estimate of the epipole (c). Also shown in (c) are the 1 standard deviation uncertainty ellipses (assuming $\sqrt{2}$ pixel standard deviation image measurement noise) plotted around the found epipole (white for the prediction, black for that found by small perturbation analysis), and the translational component of the image velocities, calculated by subtracting the rotational components found assuming approximate intrinsic camera parameters. With accurate intrinsic calibration, the translational components will all point at the epipole and have a magnitude approximately inversely proportional to their depth and proportional to their distance from the epipole (equation 3), and qualitatively this can be seen here. The ellipses are both based on small perturbations and do not attempt to model the non-linearity of large errors, but still the predicted uncertainty is rarely very accurate. This is partially because the independence assumption, though good enough to produce near optimal weightings for the epipole estimate, is not sufficiently good to produce an accurate estimate of uncertainty. These variance predictions have been expanded by a factor (the degree of overuse of each measurement) to compensate for the dependence of the variables and therefore tend to overestimate. (d) shows the projection of the axis of rotation and the rotational components which, though large, can still be modelled to a sufficient accuracy by the velocity model.

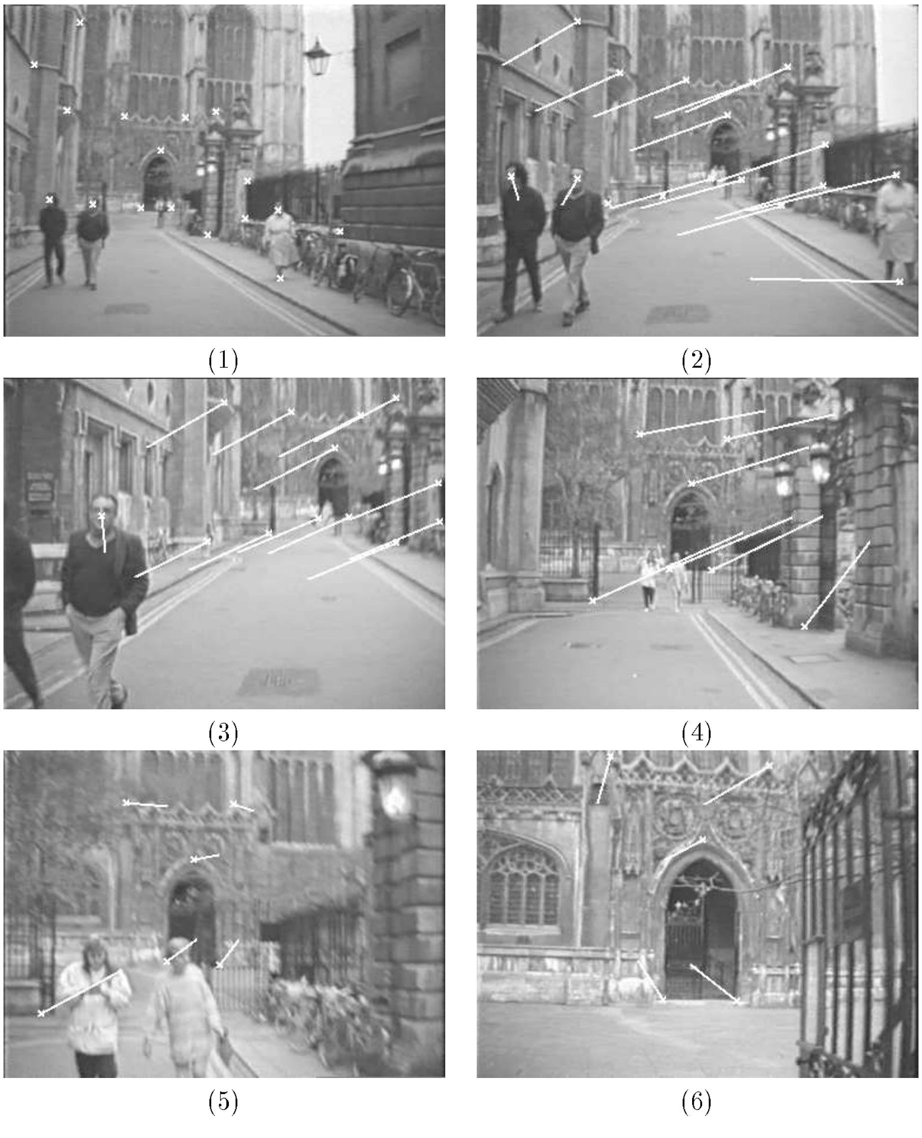


Figure 6: 'Kings College Chapel North Gate' (Frames 1–6). The full sequence of points, showing the feature movements from the previous frames due to motion towards the gates and arbitrary rotations.

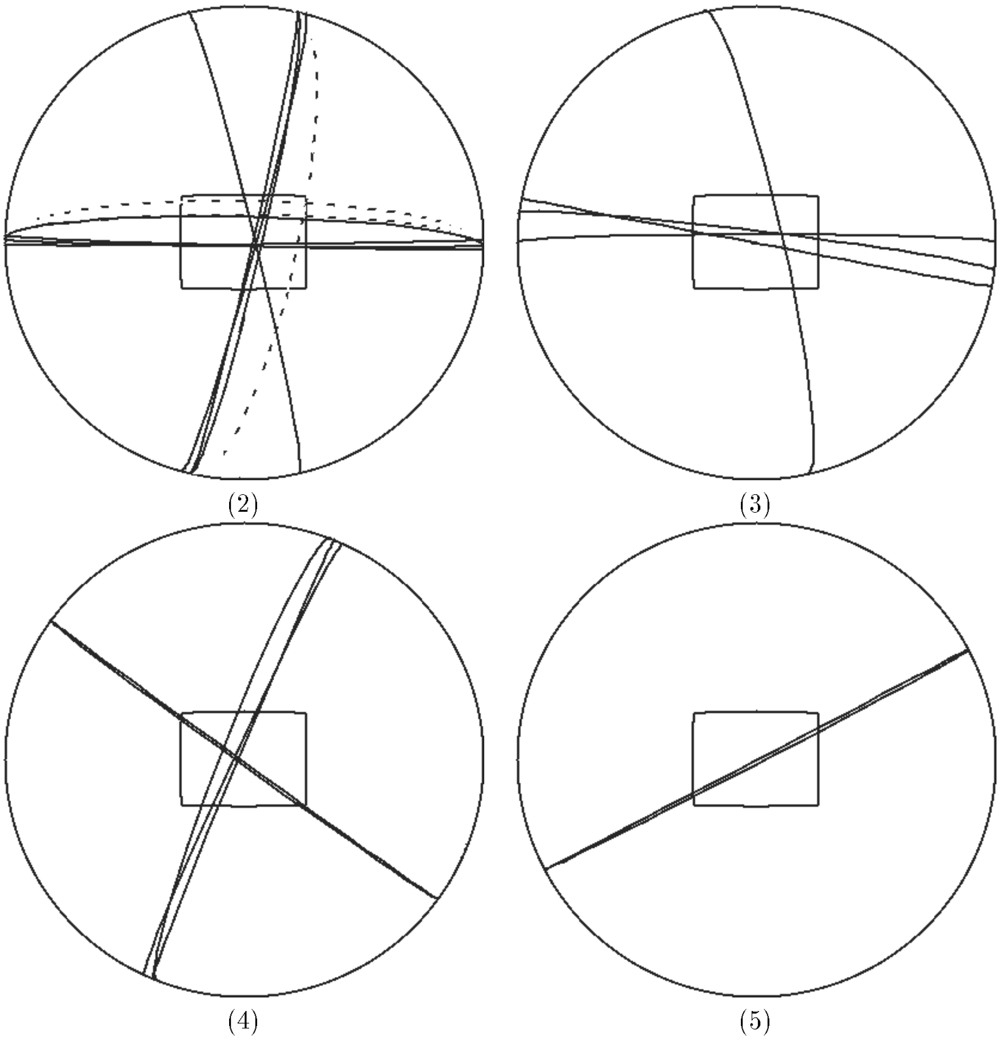


Figure 7: 'Kings College Chapel North Gate' (Frames 2–5) showing epipole constraints found plotted on a hemisphere. Those that are dashed were rejected as inconsistent. The number of constraints found decreases as the points tracked become more sparse. In (4-5) there is only one point out of the plane of the wall, and therefore only one constraint. In (5-6) there are none and therefore no constraints.

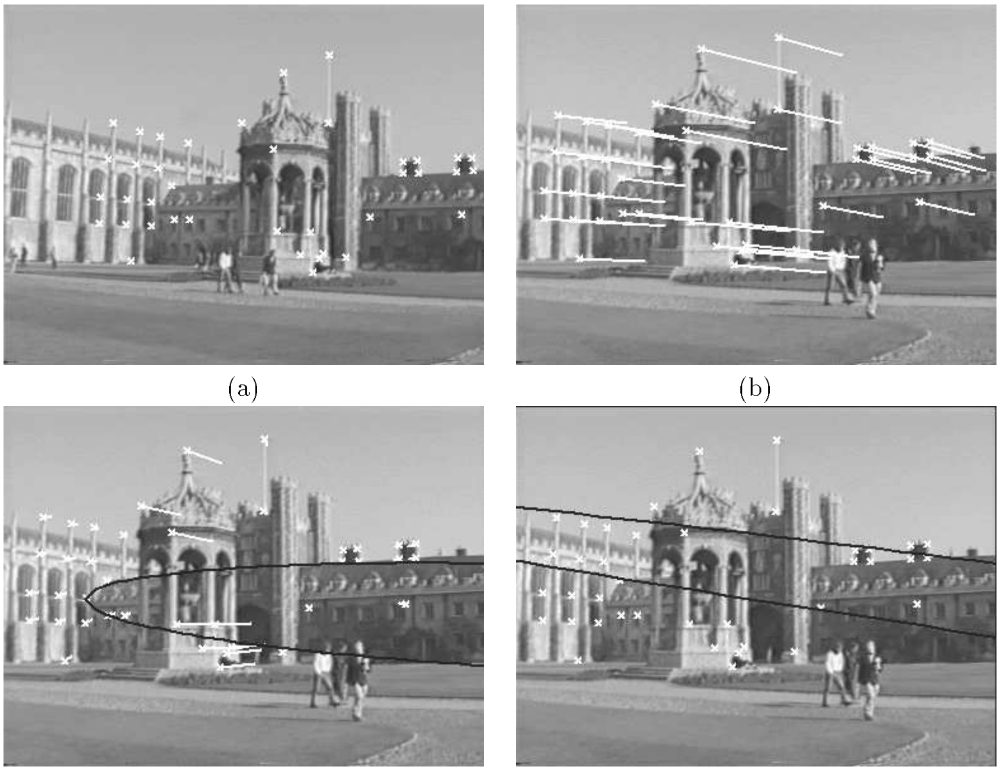
Figure 8: The approach to Kings College Chapel North Gate (Frames 1–4) using Affine Motion Parallax, showing the epipole and extracted translational components (a) and the extracted rotation axis and rotational components (b) (as in figure 5). Again the translational component magnitudes appear as expected (see Equation 3) and are all in the direction of the epipole. The people move in the image as stationary objects twice as close would, and therefore are not rejected as independent motion.

(2)

(3)

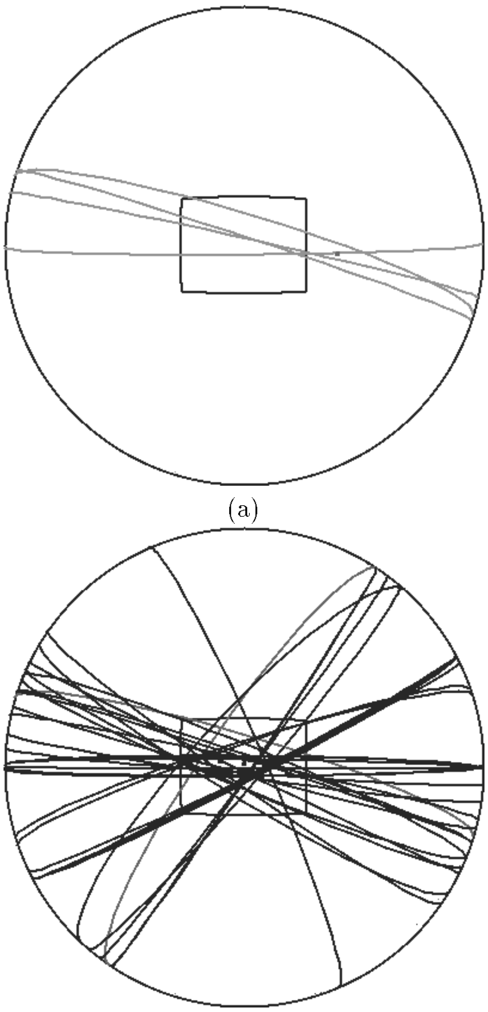
(4)

Figure 9: The approach to Kings College Chapel North Gate (Frames 1–4) showing the epipoles found by the Fundamental Matrix method (accurately initialised), and the 1 standard deviation uncertainty ellipses (assuming $\sqrt{2}$ pixel standard deviation image measurement noise) found by small perturbation analysis plotted around the found epipole (equivalent to the black ellipses in figure 8).



(a) (b) (c) (d)

Figure 10: 'Trinity College Great Court Fountain', frame (2-)3. The image features in frame 2 (a) are displaced in frame 3 due to camera translation along the path visible in the bottom right of the picture (putting the epipole on the right image boundary) and arbitrary rotation (b). The epipole found by Affine Motion Parallax (with translational image velocity components shown, as in figure 8) is reasonably accurate but somewhat uncertain (c), but that found by the Fundamental Matrix method (as in figure 9) is too far right and very sensitive to noise in one direction (d).



(a)

(b)

Figure 11: 'Trinity College Great Court Fountain', frame (2-)3 (see Figure 10). The epipolar plane constraints found using Affine Motion Parallax plotted onto a hemisphere: (a) rejecting all uncertain constraints, and (b) including all constraints. Constraints with significant error do not add linearly to the estimate and therefore improve it, but tend to bias it towards the centre of the field of view.

The 'Trinity College Great Court Fountain' Sequence

These frames are very interesting because they contain a reasonable number of features, though many lie on common planes and most are in the background, which would make them accessible to more conventional motion parallax algorithms such as Rieger and Lawton's [30]. Also the epipole is just out of view, which many methods find testing. Figure 10c shows the Affine Motion Parallax estimate, which is accurate, though high uncertainty is calculated.

The 'Trinity College GCF' frames show a problem that has appeared in many works on epipole detection: the epipole is estimated to be closer to the centre of view than it should be [14]. In the case of the combination of parallax constraints (as here), this occurs because, though small errors in the constraints will produce a linear sum of errors in the result (and therefore minimal bias (see Section 3), as implemented and shown in Figure 11a), large errors will usually constrain the epipole to lie in or near the field of view (as that is the one place that all constraints pass through, as is demonstrated when all the epipolar planes collected are displayed, as in Figure 11b).

The linear Fundamental Matrix method has been shown to be biased [8], but the non-linear method here appeared not to be (inwards at least). It is, however, very unstable, and the actual epipole predicted is too far off the field of view, though the uncertainty analysis does show that this is unlikely to be a consistent error (Figure 10d).

4.4 Comparison

The results above appear very conclusive, but it must be remembered that these examples all use relatively sparse point sets and high rotation and noise levels in comparison to most system specifications. However the robustness of the affine motion parallax algorithm, and the graceful degradation, are obvious. Though the rejection scheme does occasionally fail, this is probably inevitable, and hopefully using temporal filtering along a sequence of frames will provide sufficient information to correct this.

The Fundamental Matrix method, as implemented, regularly fails to locate the correct minima: if a good estimate of the egomotion is used to derive an initial F matrix from which to start the minimization, then the method can improve the accuracy of the estimates (better than the Affine Motion Parallax method), but without a good initialization, a large number of false minima exist and minimisation becomes very difficult. The instability may come from a lack of implicit constraints: though it is true that the affine motion parallax method finds the epipole independent of the projection plane centre, the focal length and the aspect ratio, it does assume that they are constant from frame to frame. The Fundamental Matrix method makes no such assumption and can therefore handle any two unmatched cameras [9], which is unnecessary here. There is little evidence that the matrix tends towards having two small eigenvectors, as was found for the Essential Matrix method [8], except when the image motion measured was very small.

Another obvious disadvantage of the Fundamental Matrix method is that no uncertainty measure is produced, whereas the data from the Affine Motion Parallax algorithm is easily fused with independent information.

5 Conclusion

5.1 Summary

We have presented here a method of extracting from a pair of frames first the epipole and then the camera rotation and feature depths. The method is extremely robust because it uses first derivatives of the image positions only and because the first stage involves a direct solution for only two variables (the epipole direction) and is independent of camera rotation and intrinsic parameters. The reduced dimension of the minimisation also considerably reduces the computational expense. A method is also presented for estimating the uncertainties of each estimate, which allows optimal combination of constraints. There are a number of approximations due to statistical dependence and the affine assumption, but results show the 'optimal' solution found to be very good. Results have been given from real scenes with sparse point sets that show that this method compares favourably with the Fundamental Matrix method, especially when the epipole is not in view or the point set is sparse.

5.2 Future Work

We are currently making a quantitative evaluation in comparison with the Fundamental Matrix method. The emphasis then will be on integrating the frame pair measurements into a frame sequence system, using Kalman filtering and the depth feedback mentioned above. Also a real-time corner tracker will be implemented and the system attached to a mobile robot.

References

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

[2] Y. Aloimonos and Z. Duric. Active egomotion estimation : A qualitative approach. In *Proc. 2rd European Conf. on Computer Vision*, pages 497–510, 1992.

[3] A.R. Bruss and B.K.P. Horn. Passive navigation. *Computer, Vision Graphics and Image Processing*, 21:3–21, 1983.

[4] R. Cipolla and A. Blake. Surface orientation and time to contact from image divergence and deformation. In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision*, pages 187–202. Springer–Verlag, 1992.

[5] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *Int. Journal of Computer Vision*, 9(2):83–112, 1992.

[6] R. Cipolla, Y. Okamoto, and Y. Kuno. Robust structure from motion using motion parallax. In *Proc. 4th Int. Conf. on Computer Vision*, pages 374–382, 1993.

[7] R.T. Collins and R.S. Weiss. Vanishing point calculation as a statistical inference on the unit sphere. In *Proc. 3rd Int. Conf. on Computer Vision*, 1990.

[8] K. Daniilidis and H-H. Nagel. Analytical results on error sensitivity of motion estimation from two views. *Image and Vision Computing*, 8(4):297–303, 1990.

[9] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision*, pages 563–578. Springer–Verlag, 1992.

[10] O.D. Faugeras, Q.-T. Luong, and S.J. Maybank. Camera self-calibration: Theory and experiments. In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision*, pages 321–334. Springer–Verlag, 1992.

[11] O.D. Faugeras, F. Lustman, and G. Toscani. Motion and structure from motion from point and line matches. In *Proc. 1st Int. Conf. on Computer Vision*, pages 25–34, 1987.

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

[13] H. von. Helmholtz. *Treatise on Physiological Optics*. Dover (New York), 1925.

[14] E.C. Hildreth. Recovering heading for visually guided navigation in the presence of self-moving objects. *Philosophical Transactions of the Royal Society of London, Series B*, 337:305–313, 1992.

[15] K. Kanatani. Computational projective geometry. *Computer Vision, Graphics and Image Processing (Image Understanding)*, 54/3:333–348, 1991.

[16] K. Kanatani. Renormalization for unbiased estimation. In *Proc. 4th Int. Conf. on Computer Vision*, pages 599–606, 1993.

[17] J.J. Koenderink. Optic flow. *Vision Research*, 26(1):161–179, 1986.

[18] J.J. Koenderink and A.J. van Doorn. Facts on optic flow. *Biol. Cybernetics*, 56:247–254, 1987.

[19] J.J. Koenderink and A.J. van Doorn. Affine structure from motion. *J. Opt. Soc. America*, 8:377–385, 1991.

[20] J.M. Lawn and R. Cipolla. Epipole estimation using affine motion-parallax. In *Proc. 4th British Machine Vision Conference*, pages 379–388, 1993.

[21] H. Li, M.A. Lavin, and R.J. Le Master. Fast hough transform: A hierachical approach. *Computer Vision, Graphics and Image Processing*, 36:139–161, 1986.

[22] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

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

[24] Q.-T. Luong, R. Deriche, O. Faugeras, and T. Papadopoulo. On determining the fundamental matrix: Analysis of different methods and experimental results. Technical Report 1894, INRIA, France, 1993.

[25] M.J. Magee and J.K. Aggarwal. Determining vanishing points in perspective images. *Computer Vision, Graphics and Image Processing*, 26:256–267, 1984.

[26] S.J. Maybank. The angular velocity associated with the optical flow field arising from motion through a rigid environment. *Proc. Royal Society, London*, A401:317–326, 1985.

[27] S.J. Maybank. Algorithm for analysing optical flow based on least squares method. *Image and Vision Computing*, 4:38–42, 1986.

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

[29] J.A. Perrone. Model for the computation of self-motion of biological systems. *J. Opt. Soc. America*, A9(2), February 1992.

[30] J.H. Rieger and D.L. Lawton. Processing differential image motion. *J. Opt. Soc. America*, A2(2):354–360, 1985.

[31] L.G. Roberts. Machine perception of three - dimensional solids. In J.T. Tippet, editor, *Optical and Electro-optical Information Processing*. MIT Press, 1965.

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

[33] C. Tomasi and J. Shi. Direction of heading from image deformation. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 422–427, 1993.

[34] A.M. Waxman, B. Kagmar-Parsi, and M. Subbatao. Closed-form solutions to image flow equations for 3D structure and motion. *Int. Journal of Computer Vision*, 1:239–258, 1987.

[35] D.H. Wolpert. Combining generalizers using partitions of the learning set. In L.Nadel and D.Stein, editors, *1992 Lectures in Complex Systems*. Addison-Wesley, 1993.

[36] X. Zhuang, T.S. Huang, and R. M. Haralick. A simplification to linear two-view motion algorithms. *Computer Vision, Graphics and Image Processing*, 46:175–178, 1989.

A The Optimal Sum of Estimates

Let $\tilde{\mathbf{x}}$ be the optimal estimate of $\mathbf{x}$. $\tilde{\mathbf{x}}$ is found from a number of estimates $\mathbf{x}_i$ by taking the weighted average of them, $\tilde{\mathbf{x}} = \sum w_i \mathbf{x}_i$ (where $\sum w_i = 1$), that minimises the expected standard deviation σ. The estimates are assumed to be unbiased, so that their expected value $E[\mathbf{x}_i] = \mathbf{x}$, and to have only independent small additive gaussian noise, $\mathbf{x}_i = \mathbf{x} + \Delta\mathbf{x}_i$. This is represented by known covariance matrix, $V[\mathbf{x}_i] = E[\Delta\mathbf{x}_i \Delta\mathbf{x}_i^T]$. σ^2 is defined as $\sigma_x^2 + \sigma_y^2 + \sigma_z^2$, ie. the trace of $V[\tilde{\mathbf{x}}]$, and σ_i^2 is the trace of $V[\mathbf{x}_i]$.

$$\tilde{\mathbf{x}} = \sum w_i \mathbf{x}_i \tag{17}$$

$$E[\tilde{\mathbf{x}}] = E\left[\sum w_i \mathbf{x}_i\right] \tag{18}$$

$$\sigma^2[\tilde{\mathbf{x}}] = E\left[\left(\sum w_i \Delta\mathbf{x}_i\right)^2\right] = E\left[\sum (w_i \Delta\mathbf{x}_i)^2\right] = \sum w_i^2 \sigma_i^2 \tag{19}$$

We want to minimise $\sigma^2[\tilde{\mathbf{x}}] = \sum w_i^2 \sigma_i^2$ given that $\sum w_i = 1$. We therefore minimise $C = \sum (w_i^2 \sigma_i^2) - \lambda(\sum w_i - 1)$ for each w_i (Lagrange) :

$$\frac{dC}{dw_i} = 2 w_i \sigma_i^2 - \lambda = 0 \tag{20}$$

$$\sum w_i = \sum \frac{\lambda}{2\sigma_i^2} = \lambda \sum \frac{1}{2\sigma_i^2} = 1 \tag{21}$$

$$\lambda = \left(\sum \frac{1}{2\sigma_i^2}\right)^{-1} \text{ and } w_i = \frac{1}{\sigma_i^2 \sum \frac{1}{\sigma_i^2}} \tag{22}$$

The weight given to each estimate is proportional to the inverse of its variance, $\frac{1}{\sigma_i^2}$.

B Affine Motion Parallax is independent of rotations

B.1 Using an Image Sphere

Define a point $\mathbf{p}$ as our parallax point, and $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ as our affine basis defining virtual point $\mathbf{q}$. The affine motion parallax vector, $\dot{\Delta}$, is in the direction of the difference between the image velocities of $\mathbf{p}$ and $\mathbf{q}$:

$$\mathbf{q} = \alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c} \tag{23}$$

$$\dot{\mathbf{q}} = \alpha\dot{\mathbf{a}} + \beta\dot{\mathbf{b}} + \gamma\dot{\mathbf{c}} \tag{24}$$

$$\dot{\Delta} \propto \dot{\mathbf{p}} - \dot{\mathbf{q}} \tag{25}$$

$$\propto \dot{\mathbf{p}} - \alpha\dot{\mathbf{a}} - \beta\dot{\mathbf{b}} - \gamma\dot{\mathbf{c}} \tag{26}$$

$$\alpha = \frac{(\mathbf{p}, \mathbf{b}, \mathbf{c})}{(\mathbf{a}, \mathbf{b}, \mathbf{c})} \tag{27}$$

where $(*, *, *)$ is the vector triple product. Similarly for β and γ. It is neater to write the equation for $\dot{\Delta}$ in an homogeneous form.

$$\dot{\Delta} \propto (\mathbf{p}, \mathbf{c}, \mathbf{b})\dot{\mathbf{a}} + (\mathbf{p}, \mathbf{a}, \mathbf{c})\dot{\mathbf{b}} + (\mathbf{p}, \mathbf{b}, \mathbf{a})\dot{\mathbf{c}} + (\mathbf{a}, \mathbf{b}, \mathbf{c})\dot{\mathbf{p}} \tag{28}$$

$$\propto \sum_{a,b,c,p} (\mathbf{a}, \mathbf{b}, \mathbf{c})\dot{\mathbf{p}} \tag{29}$$

The image velocities can now be expanded into the camera motion and scene structure terms that caused them (Equation 3).

$$\dot{\Delta} \propto \sum_{a,b,c,p} -\mathbf{\Omega} \times \mathbf{p}(\mathbf{a}, \mathbf{b}, \mathbf{c}) - (I - \mathbf{p}\mathbf{p}^T)\mathbf{U}(\mathbf{a}, \mathbf{b}, \mathbf{c})\frac{1}{r_p} \tag{30}$$

$\sum_{a,b,c,p} \mathbf{p}(\mathbf{a},\mathbf{b},\mathbf{c})$ is zero for any vectors.

$$\begin{aligned}(\mathbf{p},\mathbf{c},\mathbf{b})\mathbf{a}+(\mathbf{p},\mathbf{a},\mathbf{c})\mathbf{b} &+ (\mathbf{p},\mathbf{b},\mathbf{a})\mathbf{c}+(\mathbf{a},\mathbf{b},\mathbf{c})\mathbf{p} \\ =(\mathbf{p}\times\mathbf{c})\times(\mathbf{a}\times\mathbf{b}) &+ (\mathbf{b}\times\mathbf{a})\times(\mathbf{c}\times\mathbf{p}) \\ &= 0\end{aligned} \tag{31}$$

The rotation term therefore cancels, leaving an equation relating the measured motion parallax to the cameras translational velocity, and if the image points are close we can recover a projection of the direction of motion,

$$\dot{\Delta} \propto \left(\sum_{a,b,c,p} -(I-\mathbf{p}\mathbf{p}^T)(\mathbf{a},\mathbf{b},\mathbf{c})\frac{1}{r_p}\right)\mathbf{U} \tag{32}$$

$$\approx K(I-\mathbf{m}\mathbf{m}^T)\mathbf{U} \tag{33}$$

where K is an arbitrary scalar and $\mathbf{m}$ is an average image position. If three points lie at similar depths, then $\mathbf{m}$ will be equivalent to the forth. Otherwise $\mathbf{m}$ must be approximated, and therefore its accuracy will be determined by the nearness of the points. This uncertainty in the position of $\mathbf{m}$ can be included in the covariance matrix of the constraint, giving a truer representation of the constraint uncertainty. This also improves the performance of the constraint combination methods by making the covariance matrices more isotropic.

B.2 Using an Image Plane

If the image plane, with normal $\mathbf{n}$, is used instead, the equation is no longer linear

$$\dot{\Delta} = \sum_{a,b,c,p} -(I-\mathbf{p}\mathbf{n}^T)(\mathbf{\Omega}\times\mathbf{p}+\frac{1}{d}\mathbf{U})(\mathbf{a},\mathbf{b},\mathbf{c}) \tag{34}$$

$$= \sum_{a,b,c,p} -\mathbf{\Omega}\times\mathbf{p}(\mathbf{a},\mathbf{b},\mathbf{c}) - (\mathbf{p},\mathbf{n},\mathbf{\Omega})\mathbf{p}(\mathbf{a},\mathbf{b},\mathbf{c}) - (I-\mathbf{p}\mathbf{n}^T)\frac{1}{d}\mathbf{U}(\mathbf{a},\mathbf{b},\mathbf{c}) \tag{35}$$

but the rotational terms will still cancel if $(\mathbf{p},\mathbf{n},\mathbf{\Omega})$ is approximately constant or small compared to Ω or the translational term, eg. when $\mathbf{p}\approx\mathbf{n}$.