Robust Vision

P. H. S. Torr, P. A. Beardsley and D. W. Murray
Robotics Research Group
Department of Engineering Science
Oxford University
Parks Road, Oxford, OX1 3PJ, UK
phst—pab—dwm @robots.ox.ac.uk

Abstract

The computation of structure and motion from an image sequence is a fundamental problem in vision. A fully automatic approach requires not only an understanding of geometry, on which a wide range of work has been carried out in the past, but also the ability to deal with incorrect data (such as mismatched features) which will inevitably arise in a real system. It is often assumed that a standard least squares framework is sufficient to deal with *outliers* (data that does not agree with a postulated model). However, outliers can so distort a fitting process that the final result is arbitrary. We eschew the non-robust approach as unworkable except for carefully controlled scenes, and we present a system—consisting of feature detector, feature matcher, estimation of the Fundamental Matrix, and estimation of structure—that emphasizes robustness to outliers at each stage. The system is fully automatic, and results are shown for the computation of structure-from-motion in a cluttered and unknown environment.

Spetsakis and Aloimonos [9] divided research into the problem of structure from motion into three epochs. The first was spent finding out whether the problem presented had a solution—is it possible to make three dimensional inferences from multiple distinct digitized images of an object? Once it was ascertained that there was indeed a solution, the next period had researchers devising constructive proofs of uniqueness of the solution, involving the minimum number of points e.g. [5]. Unfortunately initial algorithms were highly sensitive to noise, leading to an erroneous belief that recovery of structure was essentially an ill-posed problem and that only 'qualitative' solutions were possible. A third period of research was then directed towards using redundant information in an optimal manner to minimize the effects of noise; methods have been proposed that use more correspondences [11] and more images [9] than are necessary. The goal of defeating noise inexorably leads to methodologies that combine many observations, and this is usually done in a least squares frame work. A major drawback, however, is that outliers, which are inevitably included in the initial fit, can so distort the fitting process that the result can be arbitrary. This is especially pertinent when the true data is degenerate but the solution appears non-degenerate due to a handful

BMVC 1994 doi:10.5244/C.8.14

of rogues. Thus motion research needs to enter a fourth period, the search for efficient and robust algorithms that will provide as their output, not only the solution to the problem but a list of data that appears to be in disagreement with this solution. With this end in the view this paper presents a new Bayesian approach to outlier detection, defining a cost function and a suitable algorithm to locate the minimum of this cost function, founded on random sampling. In Section 1 we describe our robust method for parameter estimation. In Section 2 we show how the feature matching may be incorporate into the probabilistic framework. Finally we describe how the feature matching can be improved further by the use of 3D information, as implemented in the Structure from Motion system in [1].

1 Estimation of the Fundamental Matrix

A completely general model of rigid motion is provided by the Fundamental Matrix. This model does not require knowledge of the camera intrinsic parameters (focal length, aspect ratio, principal point). The case for algorithms that do not require calibration has been strongly made in [2]. Suppose that a set of points arise from an object which has undergone a rotation and non-zero translation. After the motion, the set of homogeneous image points $\{\underline{\mathbf{x}_i}\}, i = 1, \ldots N$, is transformed to the set $\{\underline{\mathbf{x}_i}'\}$ related by

$$\underline{\mathbf{x}_i}'^{\mathsf{T}}[\mathbf{F}]\underline{\mathbf{x}_i} = 0 \tag{1}$$

where $[\mathbf{F}]$ is the 3×3 Fundamental Matrix [2]. Given $n \geq 8$ we can solve for $[\mathbf{F}]$ by least squares. Unfortunately a few outliers are sufficient to skew the estimated epipolar geometry. Outliers typically arise from gross errors such as mis-matches or the occurrence of non-rigid movement inconsistent with the majority. The latter might be caused by features being on occluding contours, shadows or independently moving objects.

1.1 Robust Algorithm

In Torr [10] a comprehensive survey of robust estimators is reported, with comparisons of the results made on large scale synthetic tests as well as on real imagery. Methods were evaluated using several criteria: relative efficiency, breakdown point and computational complexity. The relative efficiency of a regression method is defined as the ratio between the lowest achievable variance for the estimated parameters (the Cramer-Rao bound) and the actual variance provided by the given method. The breakdown point of an estimator is the smallest proportion of outliers that may force the value of the estimate outside an arbitrary range. For a normal least squares estimator one outlier is sufficient to arbitrarily alter the result.

We discovered that no one method excels, but that a combination of robust methods can give a very good estimate of the epipolar geometry, even when the data is badly contaminated. We have found random sampling techniques, described in Section 1.2 give the best protection against outliers, with a high breakpoint at the expense of efficiency. Thus random sampling techniques can provide a good first guess at a solution, but the solution will need improvement and not all outliers will be detected. In order to overcome this deficiency we then apply a case deletion diagnostic [10] combined with an iteratively re-weighted least squares method to further improve the solution. We have found that the the case deletion diagnostic has a lower breakpoint but much greater efficiency. The iterative

method given above does not enforce the constraint that the determinant of the Fundamental Matrix must be zero. If the Fundamental Matrix has non-zero determinant then the epipolar lines do not all intersect and one cannot define a unique epipole. Furthermore Luong and Faugeras [6] have shown that the linear methods causes a bias of the epipole towards the image centre, and our experimental results bear this out. In order to overcome this deficiency we need a third stage: a non-linear minimization, using the parameterization they suggest. We have found that this tripartite approach gives very satisfactory results, but can yet be further improved by a conjunction of the estimation process with the matching process, as will be shown in Section 2. In Section 1.2 we outline the pertinent facts about random sampling. In Section 1.2.1 we describe the maximum likelihood approach.

1.2 Random Sampling Algorithms

An early example of a robust algorithm is the random sample consensus paradigm (RANSAC) [3] another is LMS—the least median of squares estimator [7]. We use the random sampling techniques to gain a first estimate of the solution and flush the bulk of outliers. Given that a large proportion of our data may be useless the approach is the opposite to conventional smoothing techniques. Rather than using as much data as is possible to obtain an initial solution and then attempting to identify outliers, we use as small a subset of the data as is feasible to estimate the parameters (e.g. two point subsets for a line, seven correspondences for a Fundamental Matrix). We repeat this process enough times to ensure that there is a 95% chance that one of the subsets will contain only good data points. The solution is chosen to be that which minimizes the median squared residual of all the data, in the case of LMS, or maximizes the number of inliers for RANSAC. We eschew these heuristics and propose a new probabilistic cost function given in Equation 1.2.1 and derived in Section 13. In the case of the Fundamental Matrix we have found that the best error measure to minimize is the distance of each point to its estimated epipolar line. Henceforth this shall be what we minimize as our residual.

To estimate the Fundamental Matrix we select seven points and form the data matrix:

$$[\mathbf{Z}] = \begin{bmatrix} x'_1x_1 & x'_1y_1 & x'_1 & y'_1x_1 & y'_1y_1 & y'_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ x'_7x_7 & x'_7y_7 & x'_7 & y'_7x_7 & y'_7y_7 & y'_7 & x_7 & y_7 & 1 \end{bmatrix} . \tag{2}$$

Examining the null space of this matrix we gain a one parameter family of solutions: $\alpha[\mathbf{F}]_1 + (1-\alpha)[\mathbf{F}]_2$. Introducing the constraint $|[\mathbf{F}]| = 0$ [2] allows us to obtain a cubic in α from which we obtain 1 or 3 solutions—all of which we try to see which gives the best result.

We now calculate how many subsamples we require [3, 7] propose slightly different means of calculation but the number of samples is roughly the same and here the method of calculation given in [7] is used. Ideally we would like to consider every possible subsample, but this is usually computationally infeasible, thus we wish to choose m, the number of samples, sufficiently high to give a probability in excess of say 95% that we have included a good subsample within

our selection. The expression for this probability Υ is

$$\Upsilon = 1 - (1 - (1 - \epsilon)^p)^m \tag{3}$$

where ϵ is the fraction of contaminated data. p is the number of points needed to make an estimate, in this case 7. Table 1 gives some sample values of the number m of subsamples required to ensure $\Upsilon \geq 0.95$ for given p and ϵ. It can be seen

Dimension	Fraction of Contaminated Data						
p	5%	10 %	20 %	25 %	30 %	40 %	50 %
7	3	5	13	21	35	106	382

Figure 1: *The number m of subsamples required to ensure $\Upsilon \geq 0.95$ for given p and ϵ, where Υ is the probability that all the data points we have selected in one of our subsamples are non-outliers.*

from this that far from be computationally prohibitive, the robust algorithm may require less repetitions than there are outliers, as it is not directly linked to the number of outliers only the proportion of outliers. If the fraction of data that is contaminated is unknown, as it is usual, an educated worst case estimate of the level of contamination must be made in order to determine the number of samples to be taken. We can reduce the number of samples needed as we proceed, by storing maximum percentage of inliers so far found, e.g. if one sample has 80% inliers we know that we have at most 20% contamination and need only take 13 samples, according to Figure 1.

To reduce the chance of selecting degenerate configurations we use spatial and velocity information, essaying the selection of correspondences widely separated spatially and in velocity. As points are selected at random for the sub-sample we create a probability that they might be rejected if they lie too close to points already selected within the sample.

1.2.1 A Maximum likelihood approach to random sampling

The disadvantages of RANSAC and LMS is that the cost criteria that they minimize are based upon heuristics. We propose a well principled maximum likelihood approach to mis-match detection. Consider a set of observed motion data, consisting of temporal matches of image features between two frames. We intend to find the most probable underlying interpretation Θ given the motion data D, according to the model we define, using the method of maximum likelihood. In this case an interpretation will divide the data into a set of outliers, κ_o and a set of inliers, κ_i, and provide a parameter estimate associated with the inliers. The most likely partition is obtained by maximising the joint likelihood function of all measurements over all possible partitions:

$$\max_{\Theta} \Pr[\Theta|D] \quad . \tag{4}$$

Now the conditional probability may be rewritten using Bayes' theorem,

$$\Pr[\Theta|D] = \frac{\Pr[D|\Theta]\Pr[\Theta]}{\Pr[D]} \quad . \tag{5}$$

and, as the prior $\Pr[D]$ does not depend on Θ, this is further simplified to finding

$$\max_{\Theta} \Pr[D|\Theta] \Pr[\Theta] \tag{6}$$

The likelihood function is composed of two components

$$\Pr[D|\Theta] = \Pr[\kappa_i|\Theta] \Pr[\kappa_o|\Theta] \quad , \tag{7}$$

where $\Pr[\kappa_i|\Theta]$ is the probability density function of the set of inliers, and is composed of two parts: one due to its deviation from its epipolar line, one due to the deviation in intensity between the two corners of the match. We shall assume that the variance in this distance to epipolar is Gaussian:

$$\Pr[d|\Theta] = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-d^2}{2\sigma^2}} \quad , \tag{8}$$

where σ is the standard deviation, which can be estimated from the median as will be described. The probability density function of the cluster is,

$$\Pr[\kappa_i|\Theta] = \prod_{i \in \kappa_i} \varpi_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-d_i^2}{2\sigma^2}} \quad , \tag{9}$$

the first term originates from our Gaussian assumption: d_i is the distance from epipolar line and σ its standard deviation. The second is the output of our corner matcher, giving the probability that two corners are the same from their intensities. If required we could assume a more exotic distribution for d_i but in many cases we find Gaussian assumptions suffice.

We shall assume that the distribution of distances of outliers from their estimated epipolar lines follows a uniform distribution: $\frac{1}{v}$, giving:

$$\Pr[\kappa_o|\Theta] = \left(\frac{1}{v}\right)^{n_o} \prod_{j \in \kappa_o} \varpi_j \quad , \tag{10}$$

where n_o is the number of mis-matches. Thus the total likelihood function for a given partition is:

$$\Pr[D|\Theta] = \left(\frac{1}{v}\right)^{n_o} \prod_{i \in \kappa_i} \left(\varpi_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-d_i^2}{2\sigma^2}} \right) \prod_{j \in \kappa_o} \varpi_j \tag{11}$$

If we have some knowledge about the rate of appearance of mis-matches we can make an educated guess to our prior $\Pr[\Theta]$. Let this rate be μ we can assume a Poisson process for the appearance of mis-matches:

$$\Pr[\Theta] = \frac{e^{-\mu} \mu^{n_o}}{n_o!} \quad . \tag{12}$$

Empirically we found that inclusion of the prior term gave around a 10% improvement in the standard deviation of the epipolar distance (distance of the projection

of a point to its epipolar line). Taking logarithms of the likelihood function we have the following expression for minimization:

$$-\ln \Pr[\Theta|D] = \sum_{i \in \kappa_i} \frac{d_i^2}{2\sigma^2} + n_o \ln\left(\frac{\upsilon}{\mu}\right) + \ln(n_o!) + n_i \ln\left(\sqrt{2\pi}\sigma\right) + \sum_{i \in \kappa_i \cup \kappa_o} w_i + \mu \quad (13)$$

where the weights $w_i = \ln \varpi_i$ are the likelihood of the matches given in Equation 19, and n_i is the number of inliers. We choose the random sample that minimizes this cost function and refer to this method as MLS—maximum likelihood sampling.

1.3 Standard Deviation and removal of outliers

Robust techniques to eliminate outliers are all founded upon some knowledge of the variance of the error in feature location. This variance is related to the characteristics of the image, feature detector and matcher and is not always available. If it is not, and the outliers are in the minority, a first estimate of the variance can be derived from the median squared error of the chosen parameter fit [7]:

$$\sigma = 1.4826 \left(1 + \frac{5}{n-p}\right) \sqrt{\text{med}_i d_i^2} \quad (14)$$

the factor $1.4826 = 1/\Phi^{-1}(0.75)$ was introduced because $\text{med}_i |d_i| / \Phi^{-1}(0.75)$ is a consistent estimator of σ when the d_i are distributed like $N(0, \sigma^2)$. However empirically it has been shown that the factor 1.4826 is not enough when $n \approx 2p$ and that a correction factor of $\left(1 + \frac{5}{n-p}\right)$ is needed to give a satisfactory solution. Given a solution, we must decide whether each match belongs to the inlier or outlier population, based upon its estimated distance from the epipolar line. Equation (13) defines the decision process, a point being allocated such that the contribution to this cost function is minimized. If a point is inlying its contribution is:

$$p_i = \ln\left(\sqrt{2\pi}\sigma\right) + \frac{d_i^2}{2\sigma^2} \quad . \quad (15)$$

Determination of the contribution of each outlier is more difficult as we need to know the number of outliers, n_o. So we shall consider what would be the addition to the cost function is a point that was considered an inlier is changed to an outlier, this is:

$$p_o = \ln\left(\frac{\upsilon}{\mu}\right) + \ln(n_o + 1) - p_i \quad , \quad (16)$$

some rearrangement leads us to the decision rule that we should consider a point as outlying if:

$$d_i^2 > 2\sigma^2 \ln\left(\frac{\upsilon(n_o + 1)}{\mu\sqrt{2\pi}\sigma}\right) \quad . \quad (17)$$

In order to apply this decision rule we consider matches in order of increasing d_i^2, assuming all the matches are inliers at first $n_o = 0$ and converting matches to outliers if they fulfill Equation 17, until there can be no more outliers. Thus we have established an adaptive threshold for our model considering the probability density functions of both the inlier and outlier sets.

We have rigorously tested the various methods presented on real and synthetic data. We ran the experiment randomly generating synthetic data in three space. The data was perturbed by Gaussian noise, standard deviation 1.0, and then quantized to the nearest pixel. We then introduced mismatched features to make a given percentage of the total, in this case 200 features, between 10 and 50 percent, (in increments of 5 percent). Each experiment was repeated 100 times to generate the graphs shown in Figure 2 comparing several categories of robust estimator. We measure the standard deviation of the distance of the *actual* noise free projections of the synthetic world points to their epipolar lines for several estimators. It can be seen that the ordinary least squares method (LS) is totally non-robust, giving a standard deviation of 4.7 when only 5% are introduced. The Huber M-estimator [4] rapidly breaks down after 35% outliers, but only provides inaccurate results below that. The case deletion [8] methods provide a more graceful degradation. The LMS methods provides good result but the synthesis of MLS and the case deletion diagnostic proves to be the best estimator.

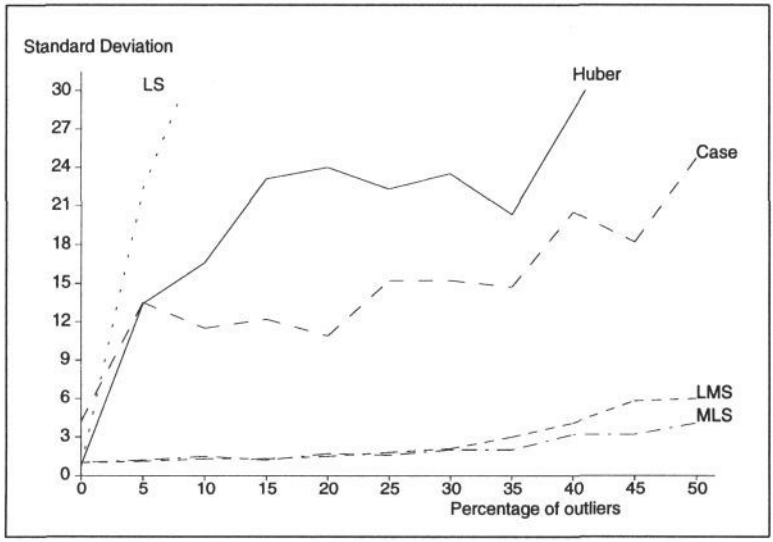

Figure 2: *Showing the standard deviation of the projections of the noise free points, to the estimated epipolar lines, over 100 tests on 200 points.*

2 Improvement of Matching

In the course of the matching process we are often presented with several candidate matches for each feature. Initially we select the one that is most similar in image intensities. The strength of match between is obtained by cross-correlation of image intensity over a 7×7 pixel patch:

$$C = \sum_{ij \in \text{patch}} (I_2(i,j) - I_1(i,j))^2 \tag{18}$$

where $I_n(i,j)$ is the image intensity at coordinate (i,j) in the nth image. We shall assume a Gaussian distribution for the variation in image intensities, thus

the probability of a match having arisen given image intensities is:

$$\varpi = \prod_{ij \in \text{patch}} \frac{1}{\sqrt{2\pi}\sigma_I} e^{\frac{-(I_2(i,j)-I_1(i,j))^2}{2\sigma_I^2}} \tag{19}$$

where σ_I is the standard deviation of the image intensities. As the match may be incorrect, it is desirable that, if in the course of the estimation process we discover that the feature is mismatched, we are able to alter this match. In order to achieve this we store for each feature not only its match, but all its candidate matches that have a similarity score over a user defined threshold. After each estimation of the epipolar geometry, in the iterative processes described above, points that are flagged as outliers are re-matched to their most likely candidate.

Feature matching proceeds as follows;

- Stage 1: for a corner $\mathbf{x}_1$ in image 1, the search area for its match $\mathbf{x}_2$ in image 2 is defined by the constraint that $\|\mathbf{x}_2 - \mathbf{x}_1\| < d_{max}$, where d_{max} is the maximum disparity of a point between image 1 and image 2.

- Stage 2: the Fundamental Matrix [**F**] is computed from the stage 1 matches, the matches are reassigned to minimize the term given in Equation 15. From these improved matches we can allow the computation of [**F**] to proceed for some further iterations thus effecting a conjunction of the matching and robust estimation processes.

- The third stage involves the use of structure once the system has created a map of the world and is described in more detail below.

2.1 Matching supported by 3D structure

The use of robust techniques in a Structure From Motion system which provides the basis for navigation of a robot arm is described in [1]. Robust methods have been found to significantly improve performance since even a few mismatches at each stage can have a large cumulative effect on recovered 3D structure over the course of an image sequence. Correspondence matching in the system is a three-stage process, the key idea being to successively reduce the search area for matching at each stage. In the first stage, the search area is constrained only by a threshold on the maximum disparity of a point between a pair of images. In the second stage, the Fundamental Matrix is computed from the stage 1 matches, and the search area for an unmatched corner's match is then constrained to lie on an epipolar line. The approach so far is similar in overview to that described previously, but there are differences - in particular, the only robust technique employed in the computation of [**F**] is iteratively reweighted least squares. This has proved sufficient since the only source of outliers is from mismatches (there is no non-rigid motion of the scene) and the number of outliers is typically below 10-15% in the indoor settings used for the work.

There follows a third stage of matching which utilises the ongoing estimate of the 3D structure of the scene computed by the system. The matches obtained in the first and second stages provide a correspondence between the 3D structure and the latest image. These correspondences are used to compute the perspective

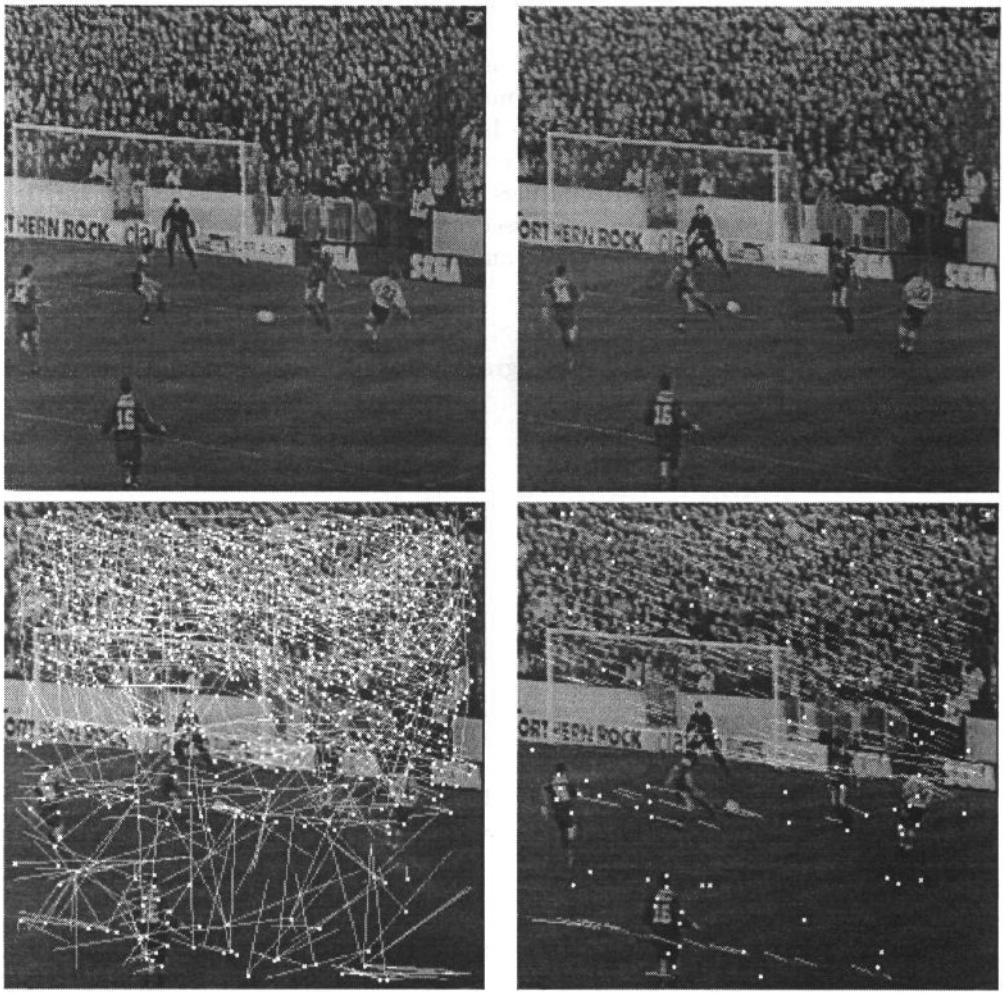

Figure 3: *Football matches.*

projection matrix $[\mathbf{P}]$ which describes the projection of the 3D structure $\mathbf{X}$ to the corners $\mathbf{x}$ in the latest image:

$$\mathbf{x} = [\mathbf{P}]\mathbf{X} \tag{20}$$

where $\mathbf{x} = (x, y, 1)^\top$, $\mathbf{X} = (X, Y, Z, 1)^\top$ are homogeneous vectors, and $[\mathbf{P}]$ is a 3×4 matrix. Each correspondence $\mathbf{x} \leftrightarrow \mathbf{X}$ provides two linearly independent equations in $[\mathbf{P}]$, so six correspondences are sufficient to uniquely determine $[\mathbf{P}]$. The computation utilises both the random sampling algorithm and iteratively reweighted least-squares. Once $[\mathbf{P}]$ is known, matching can be resumed for any unmatched corner $\mathbf{x}$ which has an associated 3D point $\mathbf{X}$, with a search area defined by the projected position $[\mathbf{P}]\mathbf{X}$ and the projected uncertainty ellipsoid for $\mathbf{X}$.

3 Conclusions

We have presented a thoroughly robust system that integrates the estimation of epipolar geometry, structure and matching under the common framework of maximum likelihood estimation. We have shown that a combination of robust methods allows us to attain the twin goals of efficiency and high breakpoint, and that negligence with regard to outliers will give a result of significantly inferior quality. Although the problem addressed in this paper has been a specific one the methodologies are completely general and can be extended to most types of robust estimation.

Acknowledgements

This work was supported by EPSRC grant No GR/H77668. PHST is in receipt of a SERC studentship.

References

[1] P. A. Beardsley, A. Zisserman, and D. W. Murray. Navigation using affine structure and motion. In *Proceedings of 3rd European Conference on Computer Vision*, pages 85–96. Springer-Verlag, 1994.

[2] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proceedings of 2nd European Conference on Computer Vision*, pages 563–578, 1992.

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Assoc. Comp. Mach.*, vol. 24:381–95, 1981.

[4] P. J. Huber. *Robust Statistics.* John Willey and Sons, 1981.

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

[6] Q. T. Luong, R. Deriche, O. D. Faugeras, and T. Papadopoulo. On determining the fundamental matrix: analysis of different methods and experimental results. Technical Report 1894, INRIA (Sophia Antipolis), 1993.

[7] P. J. Rousseeuw. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[8] 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.

[9] M. Spetsakis and J. Aloimonos. A multi-frame approach to visual motion perception. *International Journal of Computer Vision*, 6:245–255, 1991.

[10] P. H. S. Torr. *Outlier Detection and Motion Segmentation.* PhD thesis, University of Oxford, 1994. In preparation.

[11] J. Weng, T.S. Huang, and N. Ahuja. Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:451–476, 1989.