On Accurate and Robust Estimation of Fundamental Matrix

Mirosław Bober, Nikos Georgis and Josef Kittler
Department of Electronic and Electrical Engineering,
University of Surrey, Guildford GU2 5XH, United Kingdom

ees{3mb,1ng,1jk}@ee.surrey.ac.uk

Keywords: *Optical Flow Segmentation and Estimation, General Motion Estimation.*

Abstract

This paper is concerned with the accurate and robust estimation of the fundamental matrix. We show that, given a certain conditions, a basic linear algorithm can yield excellent accuracy, in cases two orders of magnitude better than sophisticated algorithms. The key element of the success is the accuracy and the statistical distribution of the errors of displacement estimates used as input. We propose a low–level, gradient–based (as opposed to feature–based) algorithm based on the Hough Transform to extract the low–level measurements. We show that it is much more efficient, both in terms of computational expense and accuracy of the final estimate, to remove the errors in the intermediate representation (optic–flow) than to attempt to improve the final estimate by complicated non–linear algorithms. Experimental results are also included.

1 Introduction

The correspondence analysis of the images of two views of a scene is a problem which arises in a number of applications in computer vision and image communication. It involves two different aspects. The first is concerned with the task of establishing which point in the second frame is physically identical to a point in the first frame. By physical identity we mean that the pair of points, each taken from a different frame, correspond to the same point in the scene. The second aspect is concerned with the measurement of the relative position of the corresponding points in the two frames.

Among the tasks for which the correspondence analysis constitutes an essential prerequisites are *depth from stereo*, *camera calibration*, *structure from motion*, *egomotion* and *image coding*. The complexity of the correspondence analysis problem depends greatly on both the scene content and the two views (the relationship of the camera to the scene and the camera transformation between the two views). For instance, the analysis of two stereo frames obtained with similarly oriented

BMVC 1996 doi:10.5244/C.10.3

cameras having a small base line separation will constrain the corresponding points to lie within a narrow horizontal band in the two images. By the same token, in the case of two images obtained with the same camera within a short period of time, the corresponding point in the second frame will be constrained to lie within a small neighbourhood of the pixel position of its physical equivalent in the first frame, as the motion due to either the dynamics of the scene or the camera will result in a limited relative displacement of corresponding objects in the two views. The difficulty of the correspondence analysis problem increases dramatically when the two views are very different, or even obtained with two different cameras with widely differing extrinsic parameters, or when the two frames are obtained at two instances of time the separation of which is significant in relation to the scene and/or camera dynamics.

Once the correspondence is established, the position of the corresponding points is used for the estimation of the relevant parameters. Depending on the application these may be the stereo camera parameters, depth from stereo or motion, camera parameters from motion, the transformation matrix between the two frames, etc. Invariably, the parameters being estimated encapsulate the 3D nature of the world captured in the two frames. As their estimation is based on 2D image measurements, the parameters can be over-sensitive to the accuracy of the 2D measurements [5].

The effect of inaccuracies in measurements is more subtle in the case of the frame to frame mapping and fundamental matrix estimation. In order to combat their undesirable influence, sophisticated nonlinear estimation procedures have been proposed [6, 10, 8, 9].

However, this widely–used approach is computationally very intensive and does not always guarantee that a correct solution to the estimation problem will be found. Moreover, non–linear algorithms are iterative and therefore they cannot be guaranteed to be implemented in real time. As an alternative, Hartley [7] recently demonstrated that the effect of measurement errors can be contained to some degree by means of a suitable normalisation after which the fundamental matrix can be estimated using a linear algorithm. However, the normalisation process does not attempt to remove the errors in measurement. It only conditions the problem in such a way that their effect on the accuracy of the final estimate is reduced. In situations when the images are very noisy the normalisation may not be sufficient.

In this paper we argue that, rather than focusing the attention on the development of very sophisticated estimation procedures, more emphasis should be placed on improving the quality of data used for estimation. Of course, the need to use as accurate image data as possible has been recognised for some time now. Accordingly, a series of enhancements has been suggested in the literature to ameliorate the inaccuracy of measurements extracted from the image. These include the use of increased image resolution, but this normally can be achieved only at the expense of narrowing the field of view and the consequential loss of the global understanding of the scene structure. The conventional reliance on feature detectors such as corner detectors to define points of interest in the two frames [11] has recently been abandoned in favour of a correlation function which makes a better use of the local grey–level information. This, therefore, overcomes the poor localisation properties

of corner detectors [4]. An additional advantage of a correlation–based matcher is that it offers a subpixel precision of the feature position measurement. However, it should be noted that an improved precision does not necessarily guarantee a better accuracy! If the measurement is inherently biased, any measurement error will eventually be dominated by the bias, and improvements in the measurement precision will not be reflected in more reliable parameter estimates. Unfortunately, correlation matching is notorious for being biased in the case of:

- a potential mismatch between the actual local transformation from a point of interest in one frame to the corresponding point in the other frame and its assumed model,
- a change in illumination,
- the presence of outliers (due to partial occlusion, for instance)

In order to overcome these problems we propose a novel method of correspondence matching which can be used to extract the relative displacement of the corresponding points to a much better accuracy. The method is based on a robust Hough transform with a smooth kernel voting which guards against outliers and contamination, while guaranteeing high statistical efficiency. The method has been developed by adapting the successful motion estimation method presented in [1]. The proposed method employs an affine feature-to-feature transformation model and it is invariant to illumination changes, in addition to being statistically robust. In this way all the major causes of bias have been eliminated to obtain accurate image-to-image measurements. Moreover, the method incorporates a scale estimator which facilitates adaptation to changing noise conditions. In principle the method can provide dense estimates of correspondence which makes the recovery of structure from motion considerably easier. A crucial advantage of the method is its internal measure of reliability for the hypothesised match which can be used as a measure of confidence in the acquired relative feature displacement measurement.

The paper is organised as follows. In section 2 we describe the low–level process of establishing correspondence and the construction of the estimate confidence measures. Experimental results investigating the accuracy and robustness of the proposed algorithm are presented in section 3. We finally conclude in section 4.

2 Robust low-level motion estimation

Let us consider the problem of a robust and accurate estimation of the correspondences between images. For the problem at hand, only a small number of correct matchings is required, say in the order of several hundred! Consequently, we are at liberty to select locations that are well suited for estimation. It seems intuitively that feature points, such as edges or corners are the best choice. Many authors use the following strategy. Firstly, feature points corresponding to high curvature are extracted from each image and then correlation is used to establish matching pairs between images. The major problem with such an approach stems from the fact that feature detectors suffer from poor localisation properties.

One of the theses argued in this paper is that a much more accurate estimate of displacement, as compared to the one based on feature correspondences, may be extracted from well–textured regions in an image. There are however two problems with using regions for motion estimation: there may be multiple motions present within a region, and the motion of large regions cannot always be approximated by a pure translational model. We have previously developed a technique for simultaneous motion estimation and segmentation based on the Hough transform and robust statistics (RHT) [3]. The technique can cope with complex motions by supporting several parametric motion models: translational, four–parameter one (quasi-affine) and affine. In addition it provides parallel motion segmentation, so that large regions can be used for estimation. We have shown the RHT be extremely robust to noise and illumination changes; it was therefore a good candidate for the low-level displacement estimation.

For the purpose of this paper, we have constructed a new confidence measure based on properties of the region and the value of support. The following subsections describe the principle behind the RHT technique and the construction of confidence measure.

2.1 Robust Hough technique

Let us define the transformed pixel difference as:

$$\epsilon(\vec{a}, p) = I_0(p) - I_1(p') \tag{1}$$

where $I_0(p)$ and $I_1(p')$ are the grey-level intensities at pixel location p and p' in the reference I_0 and consecutive frame I_1 respectively. Displacements of pixels within the region, and consequently their positions p and p' are constrained by the parametric motion model $T_{\vec{a}}$: $p' = T_{\vec{a}}(p)$ where $T_{\vec{a}}(p)$ is a geometric transformation with parameter vector $\vec{a}$. For each pixel p (except for uncovered or occluded regions) one can find a displacement vector $\vec{d_p} = (d_x, d_y)$ such that $p' = p + \vec{d_p}$. Displacement $\vec{d_p}$ is defined for translational, quasi-affine and affine models as follows:

$$\vec{d_p} = (a_1, a_2) \tag{2}$$

$$\vec{d_p} = (a_1 x - a_2 y + a_3, a_2 x + a_1 y + a_4) \tag{3}$$

$$\vec{d_p} = (a_1 x + a_2 y + a_3, a_4 x + a_5 y + a_6) \tag{4}$$

In the cases of the four parameter motion model (eq. 3) and the affine motion model (eq. 4) displacement is a function of the position of pixel in the image.

In the Robust Hough Transform the support h from any pixel p for a motion vector $\vec{a}$ is defined by a kernel function $\rho(\cdot)$:

$$h(\vec{a}, p) = \rho(\epsilon(\vec{a}, p)) \tag{5}$$

$$H(\Re, \vec{a}) = \sum_{p \in \Re} \rho(\epsilon(p, \vec{a})) \tag{6}$$

Eq. (6) expresses the total amount of support $H(\Re, \vec{a})$ received by the motion vector $\vec{a}$ from the region $\Re$. The estimate of the block motion(s) is recovered from

the position of the maximum of function H defined in the multidimensional motion parameter space. Since it is convenient to define the problem as minimisation, we redefine the support function (for example by multiplying by factor -1) so that strong support corresponds to small values of H.

The Hough Space may have up to six dimensions (for the affine motion model) and therefore the exhaustive search for a minimum has to be ruled out as being too computationally costly. However the support function (Hough space) H is well-behaved in the vicinity of the minimum and, consequently, one of the gradient based methods may be applied. We use the steepest descent search on a multiresolution discrete grid (in Hough space) [2, 3].

2.2 Confidence measures

A well–behaved confidence measure should reflect the quality of the local motion estimate. Firstly, it should reject grossly erroneous estimates which could strongly bias the final estimate. Secondly, if we assume that the random errors have unimodal distribution, the confidence measure should reflect the spread of such distribution. Finally, we would like to have a warning about the motion model failure. To achieve the above objectives, the final confidence measure is based on several factors.

The first confidence factor C_t reflects how much grey-level texture the region exhibits. It is computed before the estimation process, based on a region of interest in the reference frame. This confidence factor is based on the assumption that the accuracy of the estimate depends on how sharp and high the peak of the support function is. In practice, for n-dimensional motion model, the confidence measure C_x is an $n \times n$ dimensional matrix with the elements c_{ij} defined as:

$$c_{ij} = \frac{\partial H}{\partial a_i}\frac{\partial H}{\partial a_j} \tag{7}$$

In order to save computational time, we only proceed with the estimation for the top 10% of regions.

The second confidence factor C_t is used to reject gross errors. The total support for the region of interest H_{ROI}^{min} is compared to the average support for all regions in the image $\hat{H}^{min}$. Note, that for region with no texture, there would be a high level of support (low minimum values of H) for all motion parameters.

Finally, we examine the proportion of outliers in the region, to detect cases where the region overlaps the motion boundary (there could be some moving objects in the scene) or depth discontinuity (which would result in motion boundary in 2D optic flow field). Information about outliers is provided by the RHT module.

2.3 Adaptation of the RHT method for Fundamental Matrix estimation

We have selected the four parameter motion model, with the Tukey biweight kernel [3]. It seems that the four parameters motion model offers a good trade-off between complexity and performance. We noticed that using more complex motion models (affine) does not improve the accuracy of the final estimate and makes

(a) Image 1 (location of matches from the feature–based technique)

(b) Image 2 (location of matches from the feature–based technique)

Figure 1: The *car* sequence

it more susceptible to noise. This is hardly surprising – we give the algorithm two extra degrees of freedom, that are not really needed.

Two resolutions in the image space and four resolutions in the Hough space are used. For the experimentation, rectangular regions of 15×15 pixels in size were used.

3 Experimental Results

For the purpose of establishing the accuracy of the estimated fundamental matrix, we calculated the average point–epipolar line distance in pixels as also used by Hartley [7]. The average error over 100 runs for 10 points randomly selected out of the total matched points was used for comparisons.

Several video sequences were used, two of them presented here: the *car* sequence (512×512) and the *foreman* (176×144) sequence used in the video–coding community as a testbed for coding algorithms.

The objective of the first experiment was to establish the accuracy of the estimate of the fundamental matrices and compare it to results obtained by a *typical* feature-based approach. We use the implementation of the Zhang technique [11], which uses correlation and relaxation to obtain matches. The binaries can be obtained form from INRIA ftp server[1]. The selection of this technique as a benchmark was based on the fact that it represents a classical approach to the low-level feature extraction. It should be noted that, at this stage, we have only used the feature extraction/matching front end of the completed system.

[1] `ftp://krakatoa.inria.fr/pub/`

(a) Frame number: 348 (b) Frame number: 358

Figure 2: Two frames from the *foreman* sequence

The values of the error for each approach are presented in Tables 1 and 2 for the *car* and *foreman* sequences respectively. It can be clearly seen that the errors for the proposed algorithm are consistently one to two orders of magnitude better that of the feature-based approach. Indeed, this was the case in all the sequences tested.

We have found that matches selected by the feature-based technique (Figure 1a, 1b) are placed in locations that do not correspond to high confidence levels, as found by the RHT technique (white regions in Figure 3). This is not however the only reason for the poor performance of the feature–based technique - it is felt that the poor localisation of the features might play a more important role here.

	Zhang [11]		Proposed	
Noise (SD)	Matches	Error	Matches	Error
0	91	0.68	93	0.04
1	93	0.83	93	0.04
2	83	0.91	93	0.02
3	87	1.16	94	0.05
4	81	0.89	94	0.05
5	80	0.80	96	0.06
6	49	0.81	98	0.06
7	65	1.19	100	0.07
8	52	1.25	100	0.08
9	52	1.17	100	0.09
10	49	1.09	100	0.11

Table 1: Comparison for the *car* sequence

In the next experiment, the effects of the normalisation of the image coordinate system, as proposed by Hartley [7] is investigated. The graph in Figure 4 shows the error as a function of the number of points used for the estimation of the fundamental matrix. In both cases we used the data and confidence extracted by the RHT technique - the only difference was that in one case the data was normalised. Five different pairs of graphs were plotted for various confidence levels (normalised - solid line and original - dotted line). The first observation is that the error obtained by using high-confidence un-normalised data is smaller than the one calculated with low-confidence normalised data. This proves that selecting high confidence data makes non–linear iterative algorithms redundant and a simple linear algorithm is adequate. Also, when high confidence data is used, the improvement achieved by normalisation is very limited, simply because the normalisation reduces sensitivity to measurement errors. Naturally, in practice high confidence normalised data should be used.

	Errors $\times 10^{-3}$	
Image	Zhang [11]	Proposed
1 - 2	70.0	0.7
2 - 3	56.7	0.6
3 - 4	40.5	0.5
4 - 5	27.5	0.8
5 - 6	50.4	3.1
1 - 8	33.0	5.4

Table 2: Comparison for the *foreman* sequence

One of the arguments often quoted for the use of feature–based algorithms is their robustness to noise. We tested this hypothesis, by adding gaussian noise to the frames in the *car* sequence and looking at the changes in the errors (Table 1). Even under severe noise (standard-deviation equal 10 grey-levels), the proposed algorithm was consistently better than the feature-based one.

4 Conclusions

In this paper we have shown how robust and accurate estimation of the fundamental matrix can be achieved if a low–level, gradient–based (as opposed to feature–based) algorithm is used to extract the correspondences in a pair of images. The selection of reliable estimates based on a well-defined confidence measure in conjunction with the simple and fast 8-point linear algorithm can provide excellent accuracy even without normalisation. Moreover, such solution is more amenable to real–time applications, since the execution time of the linear algorithm is short and constant.

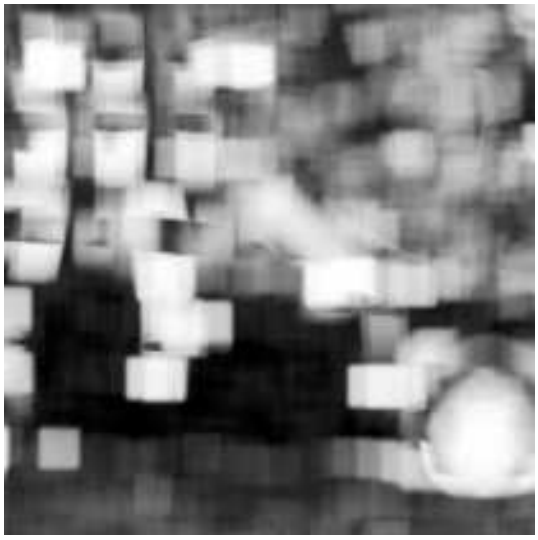

Figure 3: The confidence regions provided by RHT for the *car* sequence

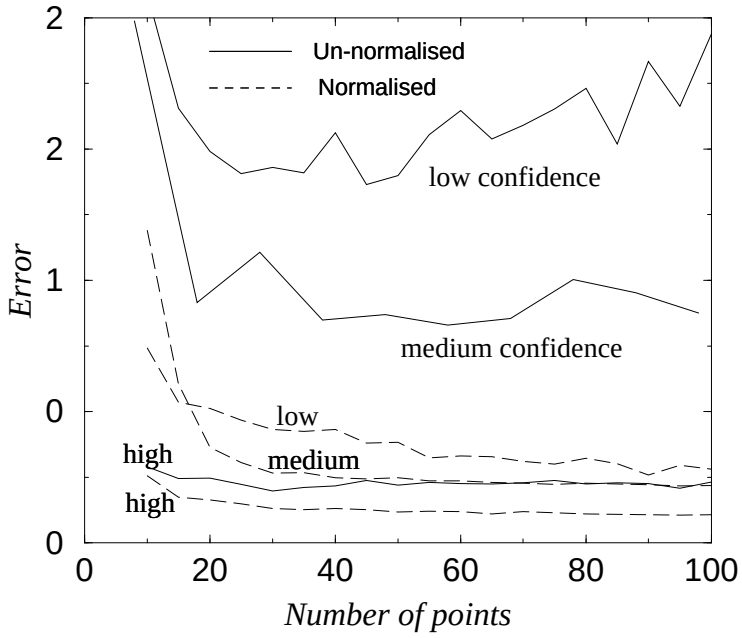

Figure 4: Normalisation is not necessary if good low–level estimates are provided

Acknowledgements

The authors gratefully acknowledge EPSRC for support under grant GR/J52563.

References

[1] M. Bober. *General Motion Estimation and Segmentation form Image Sequences*. PhD thesis, VSSP Group, University of Surrey, United Kingdom, 1994.

[2] M. Bober and J. Kittler. Estimation of general multimodal motion: an approach based on robust statistics and Hough transform. *Image and Vision Computing*, 12(12):661–668, 1994.

[3] M. Bober and J. Kittler. Robust motion analysis. In *Proceedings, CVPR '94 (IEEE Computer Society Conference on Computer Vision and Pattern Recognition), Seattle, June 20-24, 1994*, pages 947–952. IEEE Computer Society Press, 1994.

[4] M. O. F. Eryurtlu and J. Kittler. A comparative study of greylevel corner detectors. In R. Boite J. Vandervalle and A. Oosterlink, editors, *EUSIPCO'92*, volume VI: Theories and Applications of *Signal Processing*, pages 591–594, Amsterdam, August 1992.

[5] N. Georgis, M. Petrou, and J. Kittler. Projective geometry based reconstruction: Limitations and applicability constraints. In Edwin R. Hancock, editor, *British Machine Vision Conference*, volume 2, pages 529–538. British Machine Vision Association Press, York, England, September 1994.

[6] R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *European Conference on Computer Vision*, pages 579–587, 1993.

[7] R. I. Hartley. In defence of the 8–point algorithm. In *ICCV 95*, pages 1064–1070, 1995.

[8] 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, Sophia-Antipolis, France, 1993.

[9] Q. T. Luong and O. D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17(1):43, 1996.

[10] A. Shashua. Projective structure from two uncalibrated images: structure from motion and recognition. A. I. Memo 1363, MIT, September 1992.

[11] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78:87–119, October 1995.