Epipole Estimation under Pure Camera Translation*

Zezhi Chen[1,2], Nick Pears[1] John McDermid[1] and Thomas Heseltine[1]

[1] Department of Computer Science, University of York, York, YO10 5DD, UK
[2] ISN National Key Lab., Xidian University, Xiían 710071, P.R.China
{Zezhi.Chen, Nep, John.McDermid, tomh}@cs.york.ac.uk

Abstract. The position of the epipole (or focus of expansion), when a camera moves under pure translation, provides useful information in a range of computer vision applications. Here we present a robust method to estimate the epipole, which is based on the relation between the epipole and the fundamental matrix and which uses both a binning technique and random sample consensus (RANSAC). The required input data is only two uncalibrated images. No prior knowledge of either the parameters of the camera, or camera motion is required. Firstly, we use a linear method to get an initial estimate of the epipole. This is then used to initialise a non-linear optimization method, based on the minimization of the epipolar distance, in order to refine this estimate and yield a highly accurate epipole. Simultaneously, the method computes a highly accurate fundamental matrix. Extensive experimental results on real images and simulated data illustrate that the new method, which leads to an enormous improvement on the accuracy of the epipole, performs very well in terms of robustness to *outliers* and noises.

1 Introduction

In considering pure translations of the camera, one may consider the equivalent situation in which the camera is stationary, and the world undergoes a translation **-t**. In this situation points in 3D space move on straight lines parallel to **t**, and the imaged intersection of this parallel lines is the vanishing point **v,** which is in the direction of **t**. It is evident that **v** is also the epipole for both views, and the imaged parallel lines are epipolar lines. The epipole in this case is also termed the focus of expansion (FOE). In this case, it is possible to carry out an affine reconstruction from two images. A simple way of seeing this is to observe that a point **X** on the plane at infinity will map to the same point in two images related by a translation. At the same time, the fundamental matrix, which can be described by a 3×3 singular matrix, can be obtained. It is well known that fundamental matrix contains all geometric information that is necessary for establishing correspondences between two images, from which 3D structure of the perceived scene can be inferred, but in this case only up to a projective transformation [1,2,3,4,5,6]. If the intrinsic parameters of the cameras (e.g. the focal length, the coordinates of the principal point, etc) are known, we can work with normalized image coordinates, and the matrix relating the two images is known as the essential matrix [7].

Once the FOE is obtained, a homography (**H** matrix) is determined simply. It can be used to segment ground plane from the rest of the imaged scene and furthermore

* The authors acknowledge the support of the UK DTI Aeronautics Research Programme.

the affine height of objects above that ground plane can be recovered. Obviously, this is very useful in many applications, for example, where vehicles need to avoid obstacles on a ground plane. This technique has been applied to mobile robot monocular, uncalibrated obstacle avoidance [8,9]. Much previous research has addressed the automatic detection of epipole [10,11,12,13], but the epipole is highly sensitive to noise.

In this paper, a robust method to estimate epipole under pure translation is developed. On the basis of considering the relation between the epipole and the fundamental matrix, a cost function based on epipolar distance is introduced. This method does not need any prior knowledge either about the parameters of the cameras, or about their motion. Extensive experimental results on real images and simulated data illustrates that the new method leads to an enormous improvement in the accuracy of the epipole over previous methods and illustrates that it performs very well in terms of robustness to *outliers* and noises. At the same time, a highly accurate estimate of the fundamental matrix is obtained.

2 Relation Between the Fundamental Matrix and FOE

The camera model widely used is the pinhole model and, in the general case, the camera performs a projection (a linear transformation), rather than a mere perspective transformation. The epipolar constraints are the basic constraints that arise from the existence of two viewpoints. It is well known that, in stereovision, for each point $\mathbf{x}$ in the first image, its corresponding point $\mathbf{x}'$ lies on its epipolar line $\mathbf{l}'$ [17].

Let us now use retinal (pixel) coordinates. For a given point $\mathbf{x}(x, y, 1)^T$ in the first image, the epipolar line in the second image is given by $\mathbf{l}' = \mathbf{Fx}$. Since the point $\mathbf{x}'(x', y', 1)^T$ corresponding $\mathbf{x}$ belongs to the line $\mathbf{l}'$, by definition, it follows that:

$$\mathbf{x}'^T \mathbf{Fx} = 0 \qquad (1)$$

The 3×3 matrix $\mathbf{F}$ describes this correspondence; it is called the fundamental matrix and its rank is 2.

Suppose the camera matrices are those of a calibrated stereo rig with the world origin at the first camera

$$\mathbf{P} = \mathbf{K}[\mathbf{I}|\mathbf{0}] \qquad \mathbf{P}' = \mathbf{K}'[\mathbf{R}|\mathbf{t}] \qquad (2)$$

Then

$$\mathbf{F} = [\mathbf{e}']_\times \mathbf{K}'\mathbf{R}\mathbf{K}^{-1} = \mathbf{K}'^{-T}\mathbf{R}\mathbf{K}^T[\mathbf{e}]_\times \qquad (3)$$

Suppose the motion of the camera is a pure translation with no rotation and no change in the intrinsic parameters. Equation (3) then becomes

$$\mathbf{F} = [\mathbf{e}']_\times = [\mathbf{e}]_\times \qquad (4)$$

where $[\mathbf{e}]_\times$ is a skew-symmetric matrix corresponding to vector $\mathbf{e}$. $\mathbf{e}$ and $\mathbf{e}'$ are two epipoles. Obviously, according to the definition of the epipole, $\mathbf{e} = \mathbf{e}' = \mathbf{v}$ (FOE) under pure translation.

Note that in the case of pure translation $\mathbf{F} = [\mathbf{e}']_{\times} = [\mathbf{e}]_{\times}$ is a skew-symmetric matrix and has only 2 degree of freedom, which correspond to the position of the FOE. Each point correspondence provides one linear constraint on the homogenous parameters, and the FOE and fundamental matrix can be determined uniquely from two point correspondences.

3. Estimating the FOE and Fundamental Matrix

Suppose $\mathbf{x} \leftrightarrow \mathbf{x}'$ is any pair of matching points in two images. The FOE **v** must lie on the line $\mathbf{l} = \mathbf{x} \times \mathbf{x}'$, where $\times$ represents the vector cross product. That is, the triple scalar product identity

$$(\mathbf{x} \times \mathbf{x}') \bullet \mathbf{v} = 0 \tag{5}$$

This is a linear equation. Given at least two pairs of matching points, equation (5) can be used to compute the unknown FOE. If the data is not exact, because of noise in the point coordinates, then sufficiently many pairs of matching points are needed. From a set of n pairs matching points, we obtain a set of linear equations of the form

$$\begin{cases} (\mathbf{x}_1 \times \mathbf{x}'_1) \bullet \mathbf{v} = 0 \\ (\mathbf{x}_2 \times \mathbf{x}'_2) \bullet \mathbf{v} = 0 \\ \quad\vdots \\ (\mathbf{x}_n \times \mathbf{x}'_n) \bullet \mathbf{v} = 0 \end{cases} \tag{6}$$

Let the coefficient matrix, to the left of the set of equations (6), be **A**. The least-square solution for **v** is the singular vector corresponding to the smallest singular value of **A,** that is, the last column of V in: $svd(\mathbf{A}) = \mathbf{UDV}^T$.

Then we obtain:

$$\mathbf{F} = [\mathbf{v}]_{\times} \tag{7}$$

From Eq. (1), we get $\mathbf{x}'^T \mathbf{Fx} = 0$. Whereas, due to noise and the existence of *outliers*, $\mathbf{x}'^T \mathbf{Fx} \neq 0$, in general. So we define an optimisation objective function as:

$$f(\mathbf{v}) = \left(\frac{1}{\sqrt{(\mathbf{Fx}_i)_1^2 + (\mathbf{Fx}_i)_2^2}} + \frac{1}{\sqrt{(\mathbf{F}^T\mathbf{x}'_i)_1^2 + (\mathbf{F}^T\mathbf{x}'_i)_2^2}} \right) \left| \mathbf{x}'^T_i \mathbf{Fx}_i \right| \tag{8}$$

Where $(\mathbf{Fx}_i)_j^2$, (j=1,2), is the *j-th* component of vector $\mathbf{Fx}_i$ [1]. The initial value is the solution of the set of linear equations (6).

Among the matches established, we may find two types of outliers due to

1) Bad location. In the estimation of the FOE, the location error of a point of interest is assumed to exhibit Gaussian behaviour. This assumption is reasonable since the error in localization for most points of interest is small (with one or two pixels), but a few points are possibly incorrectly localized

(three or more pixels). The latter points will severely degrade the accuracy of the FOE.

2) False matches. In the establishment of correspondences, only heuristics have been used. Because the only geometric constraint (the epipolar constraint) is not yet available, many matches are possibly false. These will completely corrupt the estimation process and the final estimate of the FOE will be highly inaccurate.

The outliers will severely affect the precision of the FOE if we directly apply the method described above. Recently, computer vision researchers have paid attention to the robustness algorithm-RANSAC [14] (the RANdom Sample Consensus algorithm), because image data and the process of interest point matching is unavoidably error prone. The RANSAC algorithm is able to cope with a large proportion of outliers. It must be solved by a search in the space of possible estimates generated from the data. Since this space is so large, for example, if we assume 100 pairs of matching point, the number of samples is $C_{100}^{2} = 4950$, only a randomly chosen subset of the data can be analysed.

The question now is: how do we determine the number of samples we should use *m*? A subsample is *good* if it consists of *p* good correspondences. Assuming that the whole set of correspondences may contain up to a fraction ε of outliers, the probability **P** that at least one of the *m* subsamples is good is given by

$$\mathbf{P} = 1 - \left[1 - (1-\varepsilon)^{p}\right]^{m} \quad \textbf{(9)}$$

In implementation, we assume *p*=2, $\varepsilon = 40\%$ and require P=0.99, thus m=11. if $\varepsilon = 20\%$ then m=5.

However, two points of a subsample thus generated may be very close to each other. Such a situation should be avoided because the estimation of the FOE from such points is highly unstable and the result is highly inaccurate. In order to achieve higher stability and efficiency, we use a regular random selection method based on binning techniques [1], which works as follows. We first calculate the minimum and maximum of the coordinates of the points in the first image. The region is then evenly divided into $b \times b$ bins. To each bin is attached a set of points, and indirectly a set of matches, which fall in it. The bins having no matches attached are excluded (**Fig. 1**). But one question remains: if we assume that bad matches are uniformly distributed in space, and if each bin has the same number of matches and the random selection is uniform, does equation (9) still hold? However, the number of matches in one bin may be quite different from that in other. As a result, matches belonging to a bin having fewer matches has a higher probability of being selected. It is thus preferred that a bin having many matches has a higher probability to be selected than a bin having fewer matches. In order for each match to have almost the same probability to be selected, we implement the following procedure. If we have in total *l* bins, we divide the range [0,1] into *l* intervals so that the width of the *i*th interval is equal to $n_i \big/ \sum_i n_i$, where n_i is the number of matches attached to the *i*th bin (**Fig. 2**). During the bin selection procedure, a number produced by a [0,1] uniform random number generator falling in the *i*th interval implies that the *i*th bin is selected.

The robust algorithm to estimate the FOE is summarized in Table 1.

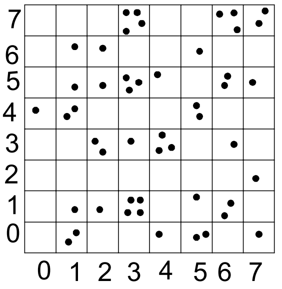


Fig. 1. Illustration of a binning technique

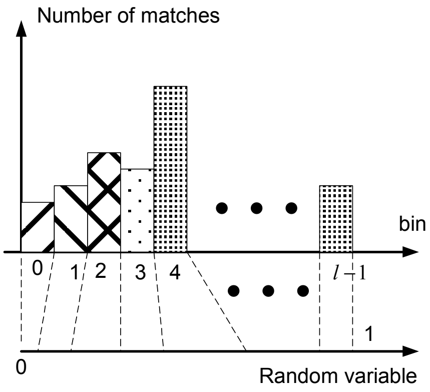


Fig. 2. Interval and bin mapping.

Table 1. A Robust Method to Estimate FOE

(1). **Extract interest points:** Compute interest points in each images by using KLT algorithm[15] or SUSAN[16] method.

(2). **Putative correspondences:** Compute a set of interest point matches based on proximity and similarity of their intensity neighbourhood.

(3**). RANSAC robust estimation:** Repeat for *m* samples, where *m* is determined adaptively by using binning technique.

(a). Select a random sample of at least 2 correspondences and compute the FOE by using Eq. (6).

(b). Calculate the epipolar distance $f(\mathbf{v})$ for each putative correspondence.

(c). Compute the number of inliers consistent with **v** by the number of correspondences for which $f(\mathbf{v})$<*threshold*.

Choose the FOE with the largest number of *inliers*.

(4). **Optimal estimation:** re-estimate the FOE from all correspondences classified as inliers, by optimizing the objective function (8).

(5). **Repeat** steps (3)-(4) until the number of correspondences is stable.

4. Experiments

A large number of synthetic data and real images were selected and intensive experimental work was carried out in order to test the robustness and the accuracy of the FOE as well as the fundamental matrix. In the paper we only give one simulated image and two real images.

4.1 Simulated Experimental Results

The simulated experiment was carried out on a 3D Euclidean model. The scene consists of 66 points (**Fig. 3**). The intrinsic camera parameters are chosen as follows. The focal length is invariant, and the synthetic camera had an aspect ratio of one and no skew, $f/dx = f/dy = 500$, the principal point has coordinates (225,225). The image size is 512×512 (**Fig. 4**).

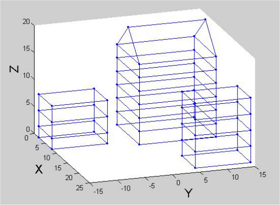
Let the first camera projective matrix be $\mathbf{P}_1 = \mathbf{K}[\mathbf{I}|\mathbf{0}]$. After translation $\mathbf{t} = (0, -2.04, -4.56)^T$, we get the second camera projective matrix $\mathbf{P}_2 = \mathbf{K}[\mathbf{I}|-\mathbf{t}]$. Where

$$\mathbf{K} = \begin{bmatrix} 500 & 0 & 225 \\ 0 & 500 & 225 \\ 0 & 0 & 1 \end{bmatrix}$$

Then, the accurate (ground truth) FOE is $\mathbf{v} = \mathbf{Kt} = (225.00, 448.60, 1)^T$. To analyse the influence of noise on the algorithm, an array of random numbers, whose elements are normally distributed with mean 0 and variance between 0 and 10.0 pixels, were used.

The results of the experiment can be seen in **Fig. 5**. Method A is the robust method proposed in this paper, method B is the linear method (defined by Eq. (6)), and method C is a method which uses the strategy of estimating the fundamental matrix first and then computes the epipole, which is also the FOE. **Fig. 5** gives the Euclidean distance and the average epipolar distance, which is the distance between real FOE and estimated FOE. This is computed by adding Gaussian noise with mean 0 and variance between 0 and 10.0 pixels respectively. In **Fig. 5** (a) the two dashed lines at the top show the left and right average epipolar distance, respectively. Because left and right epipoles are not the same in general, the epipolar distance, which shows the stability and accuracy of the fundamental matrix, is computed by using Eq. (8). These two figures show that we can use any method to get the epipole with high accuracy matching points. But if the match accuracy is not very high, the methods perform differently. Even when the matches include Gaussian noise with mean 0 and variance 6.0, the distance between the real FOE and FOE estimated with noise is only 26.61, and the average epipolar distance is only 4.42. However, the distance computed by method C is 184.99, and the average epipolar distance is 5.68. When the variance of Gaussian noise increases to 10.0, we can use method A to get a rough estimation of FOE. However, we canít use method C to get FOE, because we havenít enough good matches to use to compute the fundamental matrix. From **Fig. 5** and 6, we arrive at the following conclusions:

- Method C gives poor performance.
- Method B (the linear method) gives quite reasonable results.
- Method A (the nonlinear method based on the minimization of epipolar distance) is very robust to noise. This method not only gives a stable FOE, but also gives a highly accurate fundamental matrix.



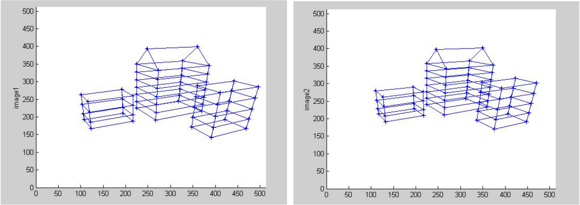


Fig. 3. A synthetic 3D scene.

Fig. 4. The synthetic images used for simulations.

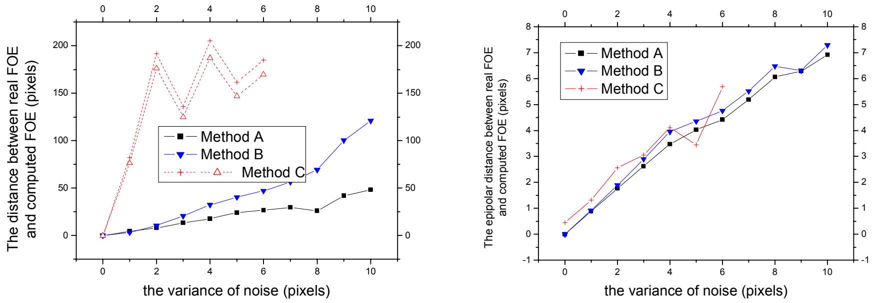


Fig. 5. (a) The distance between real FOE and estimated FOE, (b) The epipolar distance between real FOE and estimated FOE. The results are computed by add Gaussian noise with mean 0, variance between 0 and 10.0 pixels.

4.2 Experimental Results of real images

Figure 6 shows two images of an indoor scene, their matching points (+), epipoles (*) and FOE (7). We use the method proposed by Zhengyou Zhang [1] get the fundamental matrix. The left and right epipole are (325.96, 208.97) and (319.30, 210.80) respectively. The average epipolar distance is 0.38 pixels. But the FOE by using the method proposed in this paper is (412.00, 175.04). The average epipolar distance is 0.36 pixels. The trace of feature points, left epipole, right epipole and FOE are shown in figure 7. Intuitively, Figure 7 shows that the result by using the method proposed in this paper is correct.

Figure 8 shows an image of an outdoor scene, and a part of the epipolar lines. The epipole of image (a) is (-3217.7, 127.3), the epipole of image (b) is (-3194.2, 138.0), and the average epipolar distance is 0.35. But the FOE is (826.22, 406.23) and the epipolar distance corresponding to FOE is 0.69. Actually, after moving the camera on the right forwards, we get the second image. So the FOE and the fundamental matrix corresponding to it are correct.

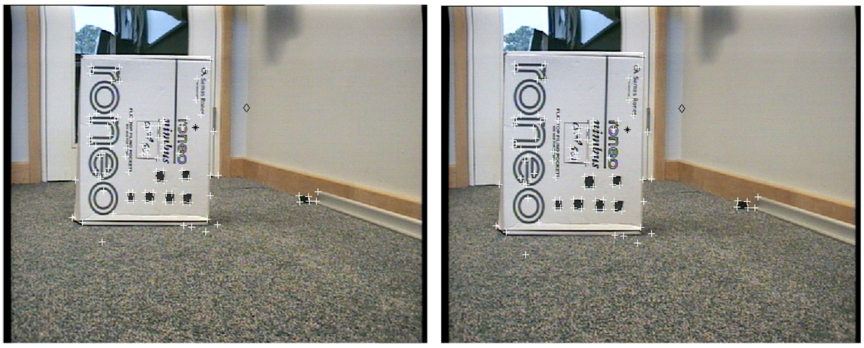


Fig. 6. . Two images of an indoor scene (the sign ë+í expresses matching point, the sign ë*í expresses the epipole and the sign 7 express the FOE.

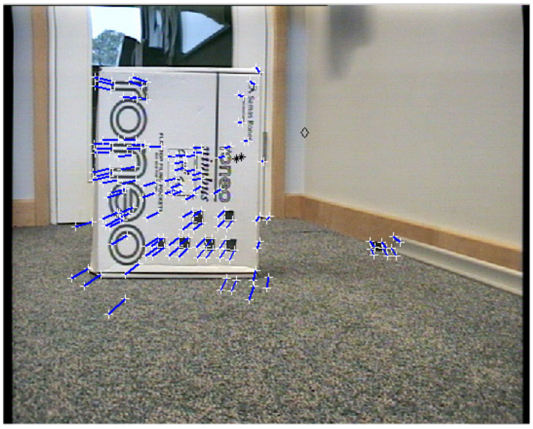
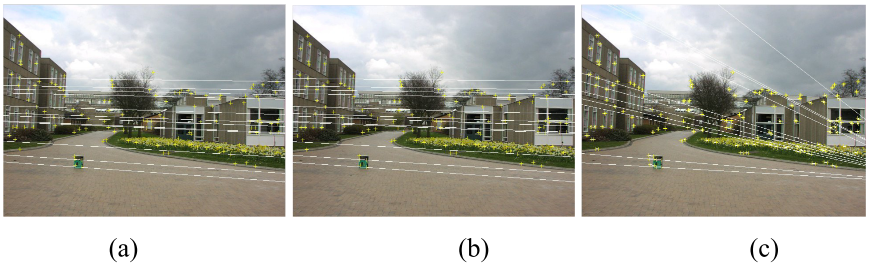


Fig. 7. The trace of feature points, left epipole, right epipole and the FOE.



(a) (b) (c)

Fig. 8. (a) and (b) are a pair of images. Matching points and a part of the epipolar lines are show in it. (c) shows the motion trace of matching points and a part of the epipolar lines.

4. Conclusion

Based on the relation between the FOE and the fundamental matrix, a robust method to estimate epipole for pure translation is proposed. The method uses a linear initial estimate stage, a non-linear refining stage, a binning technique and RANSAC. The input data is only two uncalibrated images without any prior knowledge either about the parameters of the cameras, or about their motion. Firstly, a linear method is used to get an initial value of FOE. Because the epipole is so unstable and sensitive to noise, the result vibrates greatly with various noisy experimental data. Sometimes the range of vibration can be as large as about ten thousand pixels. This means that the method of using the fundamental matrix to get the FOE is hard to use in some cases. Secondly, using a non-linear optimization method, based on the minimization of epipolar distance, we can get a highly accurate FOE. At the same time, we can get a highly accurate fundamental matrix. Both simulated and real images show that the method proposed in this paper is viable. Even when the input data is contaminated with zero mean Gaussian noise with variance 8.0 pixels, we can get an improved estimate of both epipole and fundamental matrix.

References

1. Zhang Zhengyou: Determining the epipolar geometry and its uncertainty: a review. International Journal of Computer Vision, **27(**2), (1998) 161-195
2. Chen Zezhi, Wu Chengke and Liu Yong: From an uncalibrated image sequence of a building to virtual reality modeling language (VRML). The Journal of Imaging Science and Technology. **46**(4), (2002) 365-374
3. Faugeras O.: Three-Dimensional Computer Vision: A Geometric Viewpoint. The MIT Press, 1993.
4. Faugeras O.: Stratification of 3-D vision: affine and metric representations. Journal of the Optical Society of America A, **12**(3), (1995) 465-484
5. Luong Q. -T. and Faugeras O.: The fundamental matrix: theory, algorithms and stability analysis. The Intenational Journal of Computer Vision, **1**(17), (1996) 43-76
6. Faugeras O.: What can be seen in three dimensions with an unclibrated stereo rig? In Computer Vision ECCVi92, LNCS-Series Vol. **588**, Springer-verlag: Santa Margherita Ligure, Italy, (1992), 563-578
7. Quan Long and Kanade Takeo: Affine structure from line correspondences with uncalibrated affine cameras. IEEE Trans. PAMI, **19**(8), (1997) 834-845
8. Liang B. and Pears N. E.: Visual navigation using planar homographies. Proc. of IEEE International Conference on Robotics and Automation, Washington DC, USA, Vol. **1**, (2002) 205-210
9. Liang B. and Pears N. E.: Ground plane segmentation from multiple visual cues. 2nd International Conference on Image and Graphics, Hefei, China, Proc. of SPIE, No. **4875**, (2002) 822-829
10. Liebowitz D. and Zisserman A.: Metric rectification for perspective images of planes. In Proc. of International Conference of Computer Vision and Pattern Recognition, (1998) 482-488
11. Chen Zezhi, Wu Chengke, Shen Peiyi, Liu Yong, Quan Long: A robust algorithm to estimate the fundamental matrix. Pattern Recognition Letters. **21** (2000) 851-861
12. Hartley Richard and Zissermain Andrew: Multiple View Geometry in Computer Vision. Cambridge University Press. reprinted 2001.
13. McLean G. F. and Kotturi D.: Vanishing point detection by line clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, **17**(11) (1995) 1090-1095

14. Fischer M. A. and Bolles R. C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. Assoc. Comp. Mach., **24**(6) (1981) 381-395

15. Shi Jianbo and Tomasi Carlo: Good features to track. IEEE Conference on Computer and Pattern Recognition, Seattle, USA, (1994) 593-600

16. Smith S. and Brady J.: SUSAN-a new approach to low level image processing. International Journal of Computer Vision, **23**(1) (1997) 45-78

17. Faugeras O., Luong Q-T and Papadopoulo T.: The Geometry of Multiple Images: The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications. The MIT Press, Cambridge, Massachusetts, London, England. 2001.