Detection and tracking of independent motion

J. C. Clarke and A. Zisserman
Department of Engineering Science, University of Oxford,
Parks Road, Oxford, OX1 3PJ, UK
email: [johnc,az]@robots.oxford.ac.uk *

Abstract

We describe an efficient method of using a translating camera to detect and track independently translating objects and assess the likelihood of a collision. By analysing the underlying geometry it is shown that the tracking is reduced to two independent linear searches for a single feature in the image plane. Results are presented for both an off-line implementation and work towards a real time implementation. The method is completely automatic and shown to be accurate and robust.

1 Introduction

Any autonomous vehicle must be able to detect and avoid other moving objects. Previous work on the detection of independent motion has tended to combine the (computationally expensive) optical flow field with a ground plane assumption (Enkelman [4], Carlsson and Eklundh [3]) or weak geometric constraints (Nelson [10]). With the exception of Irani *et al.* [8] this work has ignored the benefits of tracking moving objects to improve the segmentation. In contrast to these optical flow methods Torr and Murray [12] use the epipolar geometry to detect independent motion; we adopt this use of the rigidity constraint, and extend their work – in the case where all motions are pure translations – by incorporating tracking of both the background and the independent objects. The task is easily stated: given a sequence of images of a scene composed of rigid objects taken by a translating camera (figure 1) detect and track any independently translating objects. If any are found, determine whether or not a collision will ensue. This paper shows how to do this accurately, automatically and robustly without requiring camera calibration. Although the method presented uses only image corners the theory can also be expressed as a special case of the trifocal constraint (Hartley [7], Torr *et al.* [13]) thus allowing the possibility of incorporating line segments.

The three dimensional geometry, and image projection of features on the background and moving objects, are described in sections 2 and 3. Next, section 4

*Funding for this work was provided by the EPSRC and Esprit Project 6448 VIVA. We are grateful to Paul Beardsley, Ian Reid and David Murray for helpful discussions and to Paul Beardsley for the use of his software.

BMVC 1995 doi:10.5244/C.9.43

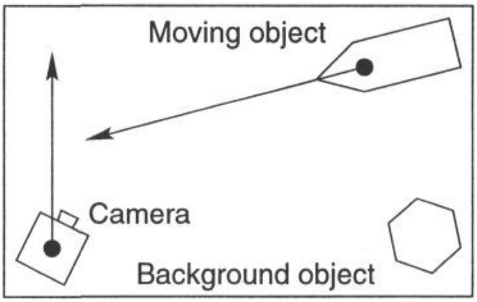


Figure 1: *The motion of the camera and the independent object is assumed to be a translation, but their directions need not be perpendicular or even in the same plane. Both are also free to vary their speed.*

gives a simple and robust test for imminent collisions. Finally sections 5 and 6 describe a complete implementation and a highly efficient frame rate feature tracker respectively. Some extensions of this work are suggested in the conclusion.

2 Model of the background motion

We are principally concerned with tracking points through a sequence of images. The problem of using a point's position in one set of images to predict its whereabouts in another is known as transfer; we shall show that for a pure translation given correspondences in two views, only one match in a third view is required to constrain the position of all the other features. Further, this single feature match can be found by a one dimensional search in the third image.

2.1 Algebraic modelling of transfer

The mechanics of image projection can be used to capture the geometric constraints imposed by the assumption of straight line motion, and predict the position of features in future images after a search for a single parameter. Using homogeneous coordinates the world position $\mathbf{X}_f$ of the f^{th} feature projects to its image position $\mathbf{x}_f^i$ at time i as

$$\mathbf{x}_f^i = \mathtt{P}^i \mathbf{X}_f \tag{1}$$

Here $\mathbf{x}_f^i$ is a 3 vector, $\mathbf{X}_f$ a 4 vector, and $\mathtt{P}^i$ a 3×4 matrix. Since homogeneous entities are equivalent up to scale $\mathbf{x}_f^i$, $\mathbf{X}_f$, and $\mathtt{P}^i$ have 2, 3, and 11 degrees of freedom respectively. A pair of views of a rigid scene separated in time is equivalent to stereo and so it is possible to find the values of $\mathbf{X}_f$ and $\mathtt{P}^i$ given sufficiently many feature correspondences (Faugeras [5], Armstrong [1]). It can be shown that by combining an uncalibrated camera model and the fact that the camera motion is a pure translation the projection matrix at the i^{th} view may be taken to have the form

$$\mathtt{P}_b^i = [\mathtt{I} | s_b^i \mathbf{e}_b] \tag{2}$$

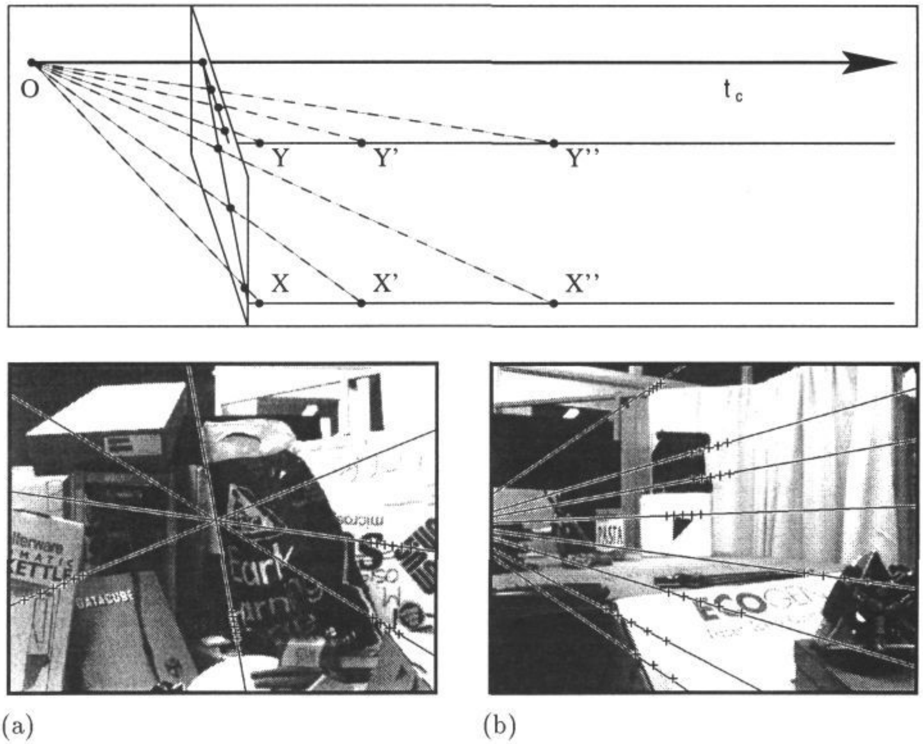


Figure 2: *The epipolar geometry of a translating camera. Upper figure: as the camera's optical centre O translates along $\mathbf{t}_c$, points X and Y in the world appear to take up new positions (X', X'', ...). The feature's epipolar lines are shown in the image, these are the intersections of the image plane with the planes defined by $\mathbf{t}_c$ and the lines $\bar{OX}$ and $\bar{OY}$ respectively. Because these epipolar planes always include the line $\mathbf{t}_c$ the epipole $\mathbf{e}_b$ must lie on every epipolar line. The epipole is the vanishing point for all lines parallel to $\mathbf{t}_c$, and remains fixed so long as the direction of $\mathbf{t}_c$ is constant. Lower figures: image sequences obtained as the camera translated towards a set of objects; a number of features at different instants are superimposed on their epipolar lines. In (a) the camera's optical axis is aligned with the direction of translation. In (b) the camera has been rotated to the right forcing the epipole to the left.*

in which s_b^i is the magnitude of the translation, and $\mathbf{e}_b$ is the background epipole or image plane projection of the direction of motion (figure 2). In this case $\mathbf{X}_f$ will be recovered up to an affinity (Moons *et al.* [9]) allowing the measurement of ratios of distances in parallel directions (in particular s_b^i). For a pure translation the epipole $\mathbf{e}_b$ has a fixed position on the image plane, and can be calculated from the motion of two image features. Once the epipole and all the features' world positions have been calculated only a single new parameter arises at the i^{th} image - the magnitude of translation s_b^i. Clearly s_b^i is determined by a single corner match, and the matching corner must lie on the epipolar line – see figure 2. Once s_b^i is known the position of all other features may be found from equations 1 and 2.

To summarise: **Grouping:** two point matches determine $\mathbf{e}_b$ which is common to all background features. **Initialisation:** two point matches determine $\mathbf{e}_b$ and affine structure, s_b^0 and s_b^1 may be chosen arbitrarily. **Tracking:** a single point match along a feature's epipolar line determines s_b^i and hence the positions of all background features.

3 Model of the independently translating object(s)

We shall show that given the background motion (i.e. $\mathbf{e}_b$ and s_b^i) independently translating objects have the same complexity as the background – the image projection for all points on the object is determined by a single new parameter at each frame and that parameter may be determined by a single feature match found using a one dimensional search.

As in the case of the background we use an uncalibrated camera and it can be shown that the projection matrix for points on the object may be chosen to have the form

$$\mathtt{P}_o^i = [\mathtt{I}|s_b^i\mathbf{e}_b + s_o^i\mathbf{t}_o] \tag{3}$$

in which $\mathbf{t}_o$ is the image plane projection of the object's direction of motion and s_o^i is the magnitude of that translation at the i^{th} frame. The **epipole** for the moving object lies on the ray $s_b^i\mathbf{e}_b + s_o^i\mathbf{t}_o$ which gives the direction of relative motion between the camera and the object, but unlike the background epipole $s_b^i\mathbf{e}_b + s_o^i\mathbf{t}_o$ varies according to the ratio $s_b^i : s_o^i$. Notice that the "ground-plane constraint" was not invoked i.e. the independent object and the camera may move in different planes.

To summarise: **Grouping:** two point matches determine $s_b^i\mathbf{e}_b + s_o^i\mathbf{t}_o$ which is common to all points on the object. **Initialisation:** three point matches determine $\mathbf{t}_o$ and affine structure. **Tracking:** given $s_b^i\mathbf{e}_b$ and $\mathbf{t}_o$ the search is along the line parameterised by s_o^i for a single point match which then determines s_o^i and hence the position of all points on the object.

4 Collision prediction

In order to decide whether or not a collision will occur requires a knowledge of the viewer's size, the potential obstacle's size, and the direction of relative motion. In this section we consider the simpler problem of deciding whether or not the camera's optical centre will collide with a moving object, thus obviating any need for knowledge about the viewer's size.

As was shown in section 3 the direction of the combined motion is given by the epipole defined by points on the moving object, and has already been calculated as part of the tracking process. The test is simple: **a collision will occur if the epipole lies inside the rectangle considered to bound the moving object.** If the epipole lies inside then the camera's relative motion is toward the object and a collision will eventually occur. If the epipole lies outside the rectangle then no collision can occur.

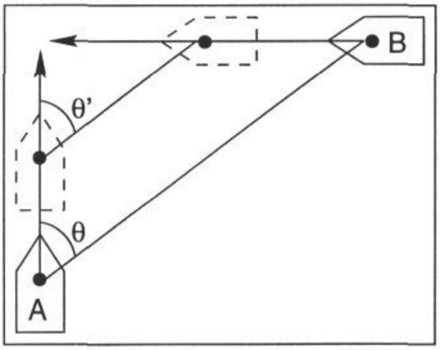


Figure 3: *The sailor's test for a collision. Two ships A and B will collide if the second makes a constant bearing θ as seen from the first, so the condition for a collision is that $\theta = \theta'$. This is a special case of the epipole test for a collision.*

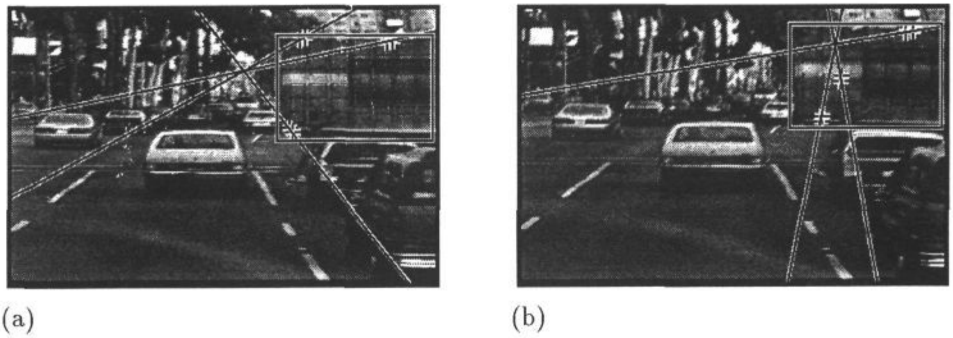


Figure 4: *The test for a collision illustrated by frames from the film* Speed. *The camera translates along its optical axis while the truck on the right moves into its path. Each image shows the bounding rectangle and three epipolar lines intersecting at the epipole for the moving object. In (a) the epipole lies ahead of the object and so at this instant the camera is moving fast enough to pass in front of the object. In (b) the epipole lies directly on the object showing that a collision will occur.*

This test can be seen as a generalisation of the sailor's test for determining whether two ships will collide. Figure 3 shows how in nautical terms a collision will occur if the second vessel makes a constant bearing as seen from the first. However, notice that the only point on the second vessel which will not appear to diverge (hence changing bearing) is the epipole, and that the epipole will remain fixed only if both the camera and moving object have constant velocities. Figure 4 illustrates the epipole test with images taken from the motion picture *Speed.*

5 Implementation I: Off-line

We have developed an off-line implementation of this method for detecting and tracking moving objects by modifying an automated corner matching program developed by Beardsley *et al.* [2]. Image corners are found to sub-pixel accuracy using the Plessey corner detector (Harris and Stephens [6]) and tracked through a sequence of images by matching them to their corresponding corner in the following frame using a two phase process incorporating both pixel value correlations as well as epipolar and structural constraints.

5.1 Robust calculation of the epipole

The assumption of pure translational motion forces the epipolar geometry to have only two degrees of freedom as figure 2 shows, consequently the epipole can be found by intersecting any two feature's image plane trajectories. The small number of parameters makes the RANSAC robust minimisation technique a highly efficient way to estimate the epipole in the presence of (the inevitable) mismatched features and moving objects. The RANSAC algorithm for finding the epipole proceeds by repeatedly using a random sample of two pairs of point matches to determine a putative epipole which is then evaluated for its support from all the feature matches. A point match is deemed to support a potential epipole if the feature's position in both images lies close to its putative epipolar line. The ultimately selected epipole is the one consistent with the most data, the remaining corner matches are ignored. A typical threshold of 1.5 pixels results in $\sim$ 80% of the corner matches having a common epipole if a moving object is not present.

A linear least squares method improves the estimate of the epipole by finding the best estimate of the common intersection of the feature trajectories. The final minimisation typically reduces the average distance of a corner from its epipolar line from ~ 0.5 to ~ 0.4 pixels.

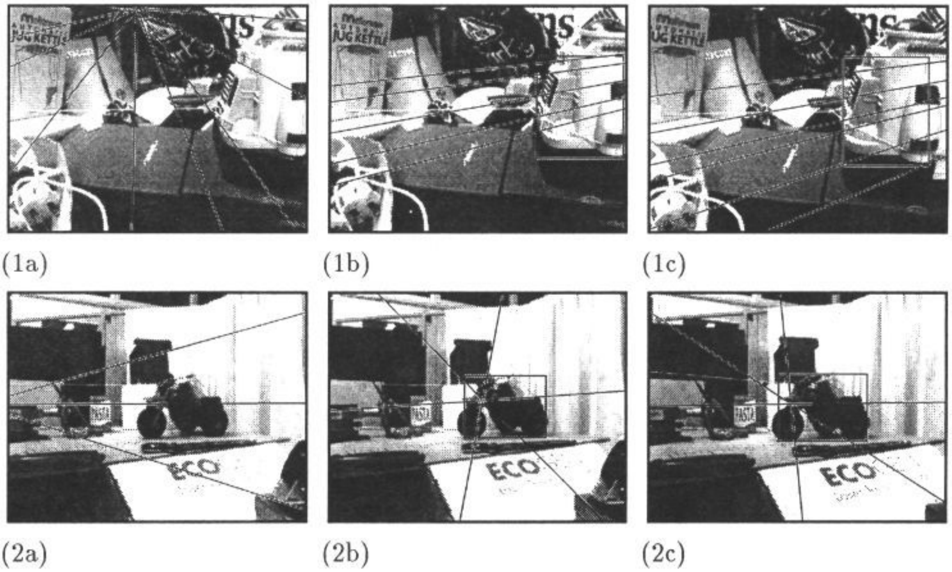
5.2 Tracking points on the background

The background is considered to be that part of the world which generates the most features on the image plane consistent with a single epipole. Given the background epipole and affine structure, points are matched as in Beardsley *et al.* [2] to provide estimates of s_b^i. As tracking proceeds the background epipole is updated as in section 5.1 and the background features' world structure is continuously refined using a least squares method that minimises the image plane errors. Once tracking is suitably advanced only features that have been tracked for more than a minimum number of frames (typically 4) are used when calculating the epipole. This continual re-estimation of the epipole implicitly tests the assumption of pure translation because the epipole's position should be constant.

5.3 The moving object(s)

Once the extent of the camera motion s_b^i is known there is only a single free parameter in equations 1 and 3 to model the moving object.

Detection: Moving objects are found by attempting to fit an epipole to the feature matches not consistent with the background epipole. If the techniques of section 5.1 find a enough features (typically 5 – 10) consistent with a new epipole then an independently moving object is deemed present. If desired the process



(1a) (1b) (1c)

(2a) (2b) (2c)

Figure 5: *Results from two tests of the off-line implementation. In each case the camera and moving object translate along approximately perpendicular paths. In both sequences image (a) shows the epipolar lines for background points, images (b) and (c) show the bounding rectangle and epipolar lines for features on the object. In images (1b) and (1c) the epipole is far from the object indicating safety, whereas in (2b) and (2c) the epipole is on the object hence there will be a collision; both reflect the ground truth.*

may be repeated to search for other moving objects until too few point matches remain, or no object is found.

Tracking a moving object: In addition to the background motion, tracking requires that the direction of the independent translation be known; these extra parameters must be found by fitting the model (equation 3) to earlier point matches; otherwise tracking an object is similar to tracking the background.

Segmentation of the moving object: The image plane extent of the moving object is defined by the smallest rectangle enclosing all the features that have been updated more than a fixed (typically 4) number of times. Other possibilities are the convex hull and methods incorporating image edges e.g. Smith [11].

5.4 Experimental results

Figure 5 shows typical results of the method, which produces a good segmentation and an accurate prediction of whether a collision will occur. In each test a small toy (robot or radio controlled buggy) translates towards the path of a translating camera mounted on an Adept industrial robot. The camera's optical axis is not aligned with the translation – in the first sequence the camera points 20° down, in the second 30° to the right. The algorithm requires $3-4$ frames to initialise an accurate segmentation. About 120 features on the background and up to ~ 50 on the moving object were tracked.

6 Implementation II: Frame rate

This section describes a real time tracker operating on a Sun IPX computer equipped with an S2200 frame grabber and a video camera mounted on an Adept industrial robot. The original off-line implementation takes over four seconds to process a single 512x512 image on a Sun IPX, so running at frame rate requires a speed up by a factor of at least 100. Almost all the processing time used by the off-line version is devoted to the Plessey corner detection algorithm, and so the problem is one of resource allocation; corner detection is slow and must only be performed where it will do the most good. This section shows that real time tracking can be performed on ordinary computers with no loss of performance.

The frame grabber alternately updates one of two 288 line interlaced fields making up a 576 line image. The computer is able to process one field for 20mS while the other is being refreshed. As we have already remarked the corner detection is extremely slow; 20mS is only long enough to detect corners in three 7×7 pixel regions, or one 19×19 pixel region.

6.1 Initialisation

The tracking is initialised by processing two images separated by a step along the camera's direction of motion. The processing in this phase is identical to the off-line version; the Plessey corner detector is run over both entire images to extract around 300 corners, these are then matched based on pixel value correlations and the epipole calculated as in section 5.1.

6.2 Tracking

The key idea is to update tracks only once the corner has moved a significant distance (up to 50 pixels) in the image plane, since the feature's position may be accurately predicted there is little to gain from more frequent updates.

Track initialisation: The full 20mS is used to apply the corner detector to a randomly chosen 19×19 pixel window. The random distribution is slightly biased towards the background epipole as this reflects where new corners will tend to appear. If any corner(s) are found that do not match a currently tracked feature then new track(s) are created. The feature's trajectory is estimated by searching for the corner at the next frame using a 7×7 pixel window centered on the old position. This limits the maximum speed of features that the initialisation can cope with to 3 pixels $\times$ 25Hz $=$ 75 pixels/second, as compared to the maximum measured in our experiments of ~ 99 pixels/second.

Once a second view of the feature is available equations 1 and 2 are used to predict the feature's image position, and the updates become less frequent.

A large number of tracks are also provided by the initialisation process used to determine the first estimate of the background epipole.

Track confirmation: When a newly detected feature has moved more than a fixed distance (typically $4-8$ pixels) and the discrepancy between its measured and predicted positions remains less than a threshold (1.5 pixels) it is considered a valid background feature. Approximately 76% of features tracked for up to 8 pixels pass this test.

Track maintenance: Each track is periodically updated by applying the corner detector to a 7×7 pixel window centered on the feature's predicted position. If a corner is found whose surrounding pixels correlate strongly enough with those when the feature was last found and the corner is close enough (within 1.5 pixels) to the feature's epipolar line then the track is updated. Tracks are serviced sufficiently frequently that the localisation error should not exceed 1 pixel.

While tracking the background epipole is updated at intervals using RANSAC and a linear least squares optimisation.

Track deletion: If the feature consistently fails to be detected at its predicted position for more than a set number of attempts (typically 4) then the track is abandoned. The track is also removed once the feature moves out of the field of view.

6.3 Experimental results

Figure 6 shows the performance of the real time tracker. Typically only one third of the processing time is devoted to maintaining existing tracks, the remainder is available to the search for new corners. The performance is specially impressive because it has been realised on very limited hardware.

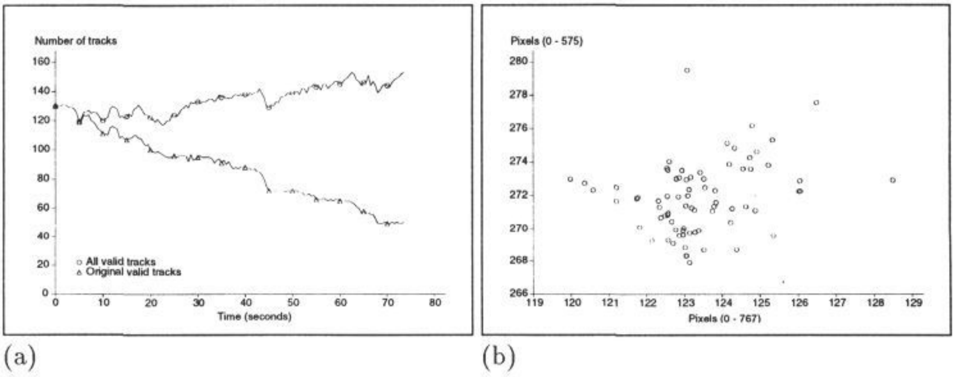


Figure 6: *Performance of the real time tracker. Graph (a) shows both the total number of valid tracks, and the number remaining from the original initialisation. The position of the epipole is shown in (b), its position is seen to be very stable over time.*

7 Conclusion

By concentrating on the geometry of the simplest case of independent motion we have shown how assumptions about the world geometry can have powerful consequences in the image plane. We have shown that for a translating camera and each independent object there is only one parameter that must be determined afresh at each stage, and that this may be found by only a one dimensional search for a single feature in each new image. We have also demonstrated a highly efficient corner tracker that is able to track up to two hundred background features at frame rate while continuously searching for new image features. This work will be extended to detect and track independently moving objects. Slightly longer term tasks are to remove the assumption of translational motion (perhaps by introducing the ground plane assumption), and to address the problem of camera shake.

References

[1] M. Armstrong, A. Zisserman, and P.A. Beardsley. Euclidean structure from uncalibrated images. In E. Hancock, editor, *Proc. 5th British Machine Vision Conf., York*, pages 509–518. BMVA Press, 1994.

[2] P. A. Beardsley, A. Zisserman, and D. W. Murray. Sequential Update of Projective and Affine Structure from Motion. In *Proc. 3rd European Conf. on Computer Vision, Stockholm*, pages 85–96, 1994.

[3] S. Carlsson and J. Eklundh. Object detection using model based prediction and motion parallax. In O. Faugeras, editor, *Proc. 1st European Conf. on Computer Vision, Antibes, France*, pages 207–306, Berlin, 1990. Springer-Verlag.

[4] W. Enkelmann. Obstacle Detection by Evaluation of Optical Flow Fields from Image Sequences. In O. Faugeras, editor, *Proc. 1st European Conf. on Computer Vision, Antibes, France*, pages 134–138, Berlin, 1990. Springer-Verlag.

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

[6] C. J. Harris and M. Stephens. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conf., Manchester*, pages 147–151, 1988.

[7] R. Hartley. Lines and points in three views – a unified approach. In *ARPA Image Understanding Workshop, Monterrey*, 1994.

[8] M. Irani, B. Rousso, and S. Peleg. Detecting and Tracking Multiple Moving Objects Using Temporal Integration. In G. Sandini, editor, *Proc. 2nd European Conf. on Computer Vision, Santa Margharita Ligure, Italy*, pages 282–287. Springer-Verlag, 1992.

[9] T. Moons, L. van Gool, M. van Diest, and A. Oosterlinck. Affine structure from perspective image pairs under relative translations between object and camera. Technical Report KUL/ESAT/M12/9306, Departement Elektrotechniek, Katholiele Universiteit Leuven, Belgium, 1993.

[10] R. C. Nelson. Qualitative detection of motion by a moving observer. *International Journal of Computer Vision*, 7(1):33–46, November 1991.

[11] S. M. Smith. Asset-2: Real-time motion segmentation and shape tracking. In *Proc. 5th Int'l Conf. on Computer Vision, Boston*, 1995.

[12] P. H. S. Torr and D. W. Murray. Statistical Detection of Independent Movement from a Moving Camera. *Image and Vision Computing*, 11(4):180–187, 1993.

[13] P. H. S. Torr, A. Zisserman, and D. W. Murray. Motion clustering using the trilinear constraint over three views. In R. Mohr and C. Wu, editors, *Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision*, pages 118–125, 1995.