Motion Clustering using the Trilinear Constraint over Three Views

P. H. S. Torr, A. Zisserman and D. W. Murray
Robotics Research Group
Department of Engineering Science
Oxford University Parks Road, Oxford, OX1 3PJ, UK

Abstract

A new method for motion segmentation is presented for clustering features that belong to independently moving objects. It is based on the geometric constraints imposed on the image positions of points and lines arising from rigidly moving objects in the world. The motion of points and lines in the image over three views are linked by the trilinear (or trifocal) constraint, which plays a similar rôle in three views to that played by the fundamental matrix in two. The fundamental matrix only imposes a one dimensional constraint on the location of features in the second image given its location in the first, whereas the trilinear constraint gives the exact location of a feature in a third image given its location in the other two. The trilinear constraint discriminates a wider range of motion than the epipolar geometry. Furthermore the trilinear constraint has the advantage that it constrains the location of lines as well as points. The segmentation problem is transformed into that of grouping the features in the image consistent with different trilinear constraints. Feasible clusters are generated using robust techniques. It is essential that the method be robust due to the prevalence of mismatches generated by state of the art feature matchers. Degenerate cases are explored, with specific emphasis on the three view constraint on points and lines imposed by the affine camera.

Introduction

Motion is a powerful cue for image and scene segmentation in the human visual system as evidenced by the ease with which we see otherwise perfectly camouflaged creatures as soon as they move, and by the strong cohesion perceived when even disparate parts of the image move in a way that could be interpreted in terms of a rigid motion in the scene. Detection of independently moving objects is an essential but often neglected precursor to problems in robotics.

In robotic vision, motion segmentation turns out to be a most demanding problem and has received considerable attention over the years, a review of which may be found in [12]. Previous approaches to motion clustering have failed because the motion models that they employ are too restrictive. For instance, if one tries to cluster based purely on similarity of image velocities then any stream of images from a static scene viewed by a camera undergoing cyclotorsion would be incorrectly segmented. Schemes based on linear variation of the motion flow field will produce false segmentations at depth discontinuities when the camera is translating. Segmentation under the assumption of orthographic or weak perspective imaging conditions will fragment scenes with strong perspective effects, even if no independent motion is present. Some methods require *a priori* knowledge of camera calibration and motion, this may not always be available. Thus the need for a more general framework is apparent.

The work in this paper stemmed from the desire to develop a general motion segmentation and grouping algorithm. That is, given two or more views of a scene, determine any of the objects within the scene which change their relative dispositions. Both the camera and objects' motion are presumed unknown as is the camera calibration. Consider Figure 3 (a)-(c), a jeep is tracked by a rotating camera, causing apparent motion to the left in the image. The jeep hits a pothole giving the semblance of the front moving down theory and algorithms given in this paper represent analysis of certain geometrical and statistical aspects of the problem. Before developing an algorithm several key design issues had to be answered: (a) What data primitives should be used to represent the scene? (b) What decision rule should be used to cluster the primitives chosen? (c) Having arrived at (a) and (b), what algorithm is used to solve the problem? The first problem is one of data reduction, the second of geometry and the third lies in the domain of computational theory. The main contribution of this paper is in the latter two areas. Returning to the first question, a major hindrance to the analysis of motion across an image is the vast amount of data to be managed. Corner and line features are most amenable to geometric and statistical analysis, which is the flavour of this work. Unfortunately they only give a sparse representation and it is envisaged that future work would flesh out the description of the segmentation. With the primitives chosen, a decision rule has to be developed to determine the segmentation. Many previous segmentation algorithms have failed to exploit the geometric reality of the world, which is readily available from image sequences.

The approach espoused in this paper is to adopt a decision rule that segments projected features in accordance with the constraints imposed by the assumption that they are rigidly connected in the world. Rather than adopting heuristics it is observed that the projections of a rigid set of points in the scene are linked over two views by a 3×3 fundamental matrix $[\mathbf{F}]$ [1, 4]. This encapsulates all the information on camera motion, camera parameters and epipolar geometry available from a given set of point correspondences. In particular neither knowledge concerning camera calibration nor the relative motion between camera and object is required to estimate $[\mathbf{F}]$. Each independently moving object will exhibit a unique fundamental matrix and generally may be distinguished on this basis, e. g. the background might exhibit one fundamental matrix and the independently moving object another.

Unfortunately there are cases when the fundamental matrices, and hence epipolar geometries, for two objects are the same and yet their motion in the world is different e.g. two objects moving with different magnitudes of translations in the same direction. This is not the case for the trifocal constraint [5, 11], which plays a similar rôle in three views to that played by the fundamental matrix in two. Rigidly connected point sets observed from three views exhibit a trilinear constraint governed by the relative disposition of object and cameras, and the calibration of the camera. The trilinear constraint provides a greater segmentation power than the fundamental matrix, as explained below, and consequently provides superior segmentation results. The segmentation method used for the fundamental matrix is reviewed first. Then the trifocal constraint and its degeneracies are described. The segmentation algorithm is based on two and three view constraints and as such represents an initialization or bootstrap of the segmentation process. It is envisaged that other processes would be used to augment the segmentation over multiple frames.

The Fundamental Matrix

Torr and Murray [13] first demonstrated that random sampling algorithm could be used to gain highly robust estimates of the fundamental matrix. Outliers have obstructed previous attempts to effect a segmentation process and the use of robust estimators proved a necessity. This lead to Torr and Murray [14] developing a random sampling approach to motion segmentation using the fundamental matrix and RANSAC [2]. Putative fundamental matrices are estimated using samples of seven point correspondences (yielding one or three solutions). The distance of each point from its epipolar line, defined by the fundamental matrix, is calculated, and if it is below a threshold, that correspondence is deemed to support that fundamental matrix. The process is repeated for a set number of times and the putative fundamental matrix with the most correspondences extracted as a cluster. This robust estimation process is repeated until there are no clusters larger than a second threshold, the unassigned correspondences are deemed outliers.

When tested on large data bases of real image sequences several problems were revealed concerning this approach. Foremost was the problem of degeneracy which leads to erroneous segmentations in which the data are incorrectly clustered. Correspondences arising from certain motions or small objects may be degenerate in that multiple fundamental matrices explain the data equally well, in this case a model with fewer degrees of freedom should be used. A full study and presentation of the problem of estimating degeneracy, in the presence of outliers, is given in [12, 15], a robust estimator is developed that simultaneously flushes outliers and selects the correct model.

A drawback with the two view approach is that the fundamental matrix defines only a one dimensional (either due to mismatching or due to independent movement) that coincidently lies near the epipolar line will not be clustered correctly.

The exploitation of the trifocal constraint on the feature locations in three views, leads to significantly superior results. The trifocal constraints provides better discrimination for motion segmentation than the fundamental matrix. Consider the following situation, the camera translates from the origin of the coordinate system to $\mathbf{t}$. Synchronous with the camera movement an object rigidly translates in the scene by $\alpha\mathbf{t}$. Presuming that there are sufficient features on the object and in the background to recover their respective fundamental matrices, then it will be found they are the same (up to a scaling). This means that the object cannot be distinguished as moving relative to the background on the basis of the projected positions over two views alone. An attempt to recover structure would lead to a spurious positioning of the independently moving object. The introduction of a third view can resolve this ambiguity [12]. As the trifocal constraint intrinsically encodes the exact position of observed features relative to the camera; the object and background would exhibit differing constraints. Thus all the features may be clustered according to their indigenous constraint.

The Trilinear Constraint

In a key paper Hartley [5] elegantly links Shashua's [11] constraint on points with Weng, Ahuja and Huang's [17] constraint on lines over three views. The projected motion of points and lines in the image over three views are linked by a trilinear constraint. In the following segmentation is effected by clustering projected features together that conform to the same trilinear constraint. The various constraints on the motion of points and lines over three views are now discussed. Firstly the trifocal constraint is stated, then degenerate cases, which typically arise from independently moving objects subtending a small field of view, are listed.

The trifocal constraint constrains the positions of each feature (points and lines) over the three images: given the corresponding locations of a feature in two images and the constraint, its location in the third can be computed. Following the notation in [5] the trilinear constraint may be represented by a tensor, the coefficients of which are represented by the triply indexed quantity T_{ijk}. Let the image of a world point $\mathbf{X}$ be $\mathbf{x}$, $\mathbf{x}^{'}$ and $\mathbf{x}^{''}$ in each of three images, where $\mathbf{x} = (x_1, x_2, x_3)^{\top}$ is the noise free homogeneous coordinate in the image plane. It can be shown [5, 11] that

$$x_l^{''} = x_i^{'} \sum_{k=1}^{k=3} x_k T_{kjl} + x_j^{'} \sum_{k=1}^{k=3} x_k T_{kil} \quad , \tag{1}$$

for all $i, j = 1 \ldots 3$. Hence there is a constraint linking rigid motion in the world to homogeneous image coordinates in image one, two, and three. Each triplet of point correspondences provides nine constraints, of which four are independent. There are 26 parameters to be estimated in T_{ijk} (not all independent). Using eigenvector methods one may solve for T_{ijk} given seven points in correspondences over three images, this is exploited in the clustering technique [1]. To solve for T a 27×27 moment matrix is formed and the solution obtained from the eigenvector corresponding to the smallest eigenvalue in the usual way.

A similar constraint may be obtained for lines [6]. If the projections of a world line $\mathbf{L}$ in the three views is $\mathbf{l}_1$, $\mathbf{l}_2$, $\mathbf{l}_3$, where $\mathbf{l} = (l_1, l_2, l_3)^{\top}$ are the homogeneous coordinates of the line, then

$$l_i = \sum_{j=1}^{j=3} \sum_{k=1}^{k=3} l_j^{'} l_k^{''} T_{ijk} \quad . \tag{2}$$

Each line gives two linear equations in the entries of T_{ijk}. To estimate the tensor a minimum of $4m_p + 2m_l > 26$ is required, where m_p is the number of point correspondences and m_l is the number of line correspondences.

The above model is valid for general motion. In many real applications independently moving objects, especially those subtending a small field of view, do not have enough indigenous correspondences to estimate the trifocal constraint. Degeneracy also occurs either when two of the camera centres are coincident or upon observing a planar surface. If degeneracy is ignored then over fitting might occur, in this instance the

[1] Using non-linear methods T_{ijk} may be recovered given six points. (e.g. [8]).

Torr [12, 15]. To avoid this, each cluster in the segmentation is described by one of the following models: full trifocal (undegenerate) constraint, or the following degenerate models: the affine camera constraint, image-image projectivity, image-image affinity, image-image translation, or no motion. The rest of this section briefly details these models.

For the modelling of independently moving objects which are typically distant and subtend a small field of view the affine camera [7, 10] is appropriate. The equivalent trifocal relations for the affine camera are as follows. For inhomogeneous coordinates (x, y):

$$\begin{aligned} p_1 x'' + p_5 x + p_9 y + p_3 x' + p_{13} &= 0 \\ p_2 x'' + p_7 x + p_{11} y + p_3 y' + p_{14} &= 0 \\ p_1 y'' + p_6 x + p_{10} y + p_4 x' + p_{15} &= 0 \\ p_2 y'' + p_8 x + p_{12} y + p_4 y' + p_{16} &= 0 \; . \end{aligned}$$

This is the extension of the result of Ullman [16] for orthography to the affine camera. For lines

$$\begin{aligned} l_3(l'_1 l''_1 p_5 + l'_1 l''_2 p_6 + l'_2 l''_1 p_7 + l'_2 l''_2 p_8) &= l_1(l''_3 l'_1 p_1 + l''_3 l'_2 p_2 - l'_3 l''_1 p_3 - l'_3 l''_2 p_4 \\ &+ l'_1 l''_1 p_{13} + l'_1 l''_2 p_{15} + l'_2 l''_1 p_{14} + l'_2 l''_2 p_{16}) \\ l_3(l'_1 l''_1 p_9 + l'_1 l''_2 p_{10} + l'_2 l''_1 p_{11} + l'_2 l''_2 p_{12}) &= l_2(l''_3 l'_1 p_1 + l''_3 l'_2 p_2 - l'_3 l''_1 p_3 - l'_3 l''_2 p_4 \\ &+ l'_1 l''_1 p_{13} + l'_1 l''_2 p_{15} + l'_2 l''_1 p_{14} + l'_2 l''_2 p_{16}) \end{aligned}$$

Therefore we need n_p point correspondences over the three images and n_l lines where $4n_p + 8n_l \geq 16$, to estimate the affine trifocal constraint. Proof in Torr [12].

If the camera motion is pure rotation, or a plane is observed, then the line and point correspondences are determined by image-image projectivities. If the projective image point transformation between image one and three is given by $\mathbf{x}'' = [\mathbf{Q}]_{13}\mathbf{x}$ then for lines $\mathbf{l} = [\mathbf{Q}]^\top \mathbf{l}''$. We shall make a simplifying assumption: As the segmentation methods are designed to take consecutive images from a sequence, we assume that the same motion model may be used between the first and second, and second and third image. This assumption is made to reduce the amount of computation necessary, it is a triviality to generalise the system, but good results have been obtained without this resort. Thus if the projectivity model is selected we assume the following relations hold:

$$\mathbf{x}'' = [\mathbf{Q}]_{13}\mathbf{x} \quad \mathbf{x}'' = [\mathbf{Q}]_{23}\mathbf{x}' \quad \mathbf{x}' = [\mathbf{Q}]_{12}\mathbf{x} \; .$$

Three other distinct models are used to describe the data, all special cases of the projectivity, an image-image affinity where

$$[\mathbf{Q}] = \begin{bmatrix} q_{ij1} & q_{ij2} & q_{ij3} \\ q_{ij4} & q_{ij5} & q_{ij6} \\ 0 & 0 & 1 \end{bmatrix} \; . \tag{3}$$

an image-image translation:

$$[\mathbf{Q}] = \begin{bmatrix} 1 & 0 & t_{ijx} \\ 0 & 1 & t_{ijy} \\ 0 & 0 & 1 \end{bmatrix} \; . \tag{4}$$

and no motion when the image correspondences are stationary. Generally, a projectivity models distant data.

Segmentation Algorithm

The segmentation algorithm has its basis in a robust estimator, dubbed PLUNDER [15], which extracts models recursively from the set of correspondences. The algorithm is quite detailed and only a brief description is given here. PLUNDER is founded on a highly robust fitting algorithm - the random sample consensus paradigm (RANSAC). Given that a large proportion of the data may belong to different populations the small a subset of the data as is feasible to estimate the parameters repeatedly to generate hypotheses, e. g. for point correspondences: 7 for undegenerate trilinear, 4 for affine, 4 for projectivity, 3 for affinity, 1 for image translation. Each point gives four linear equations in the entries of T_{ijk}, each line two. From this it is clear that half as many points as line correspondences are needed to instantiate a undegenerate solution. Taking into consideration that the line matcher used is less reliable than the point matcher this elicits the decision that only points should be used to instantiate a solution. Sampled sets of points instantiate solutions and their veracity is indicated by the number of points *and* lines consistent with each solution.

A cluster is now grown by determining how many features over the three images are consistent with the constraint estimated. To determine this two error criteria are defined: one each for lines and points. Given the trifocal constraint and the location of a scene point in two images its location in the third image may be predicted. From empirical tests the best results have been obtained by the minimization of the distance of a predicted point from its actual location in the image plane. The error criterion that is used is the average of this distance over the three images. A trio of feature correspondences is deemed consistent with a given constraint if the error criterion is below a threshold of 1.96σ, where σ is the estimated standard deviation of the error in locating a point. For lines, the root mean square of the distances of the end points of the actual line to the predicted line is used as the error criterion.

How many samples are required? Fischler and Bolles [2] and Rousseeuw [9] propose slightly different means of calculation, but both give broadly similar numbers. In this paper the method of calculation given in [9] is employed. Ideally every possible subsample would be considered, but this is usually computationally infeasible, and so m, the number of samples, is chosen sufficiently high to give a probability, Υ, in excess of 95% that at least all the correspondences in one sample are consistent with a cluster (and hence the constraint estimated from those features describe that cluster). The expression for Υ is

$$\Upsilon = 1 - (1 - (1 - \epsilon)^{p_f})^m, \tag{5}$$

where ϵ is the probability that each triplet of correspondences is inconsistent with the rest, and p_f is the number of features in each sample. Table 1 gives examples of the number m of subsamples required to ensure $\Upsilon \geq 0.95$ for given p and ϵ. It can also been seen that the smaller the data set needed to instantiate a model, the less samples are required for a given level of confidence Υ.

p_f	Fraction of Contaminated Data ϵ						
2	2	2	3	4	5	7	11
3	2	3	5	6	8	13	23
4	2	3	6	8	11	22	47
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 consistent.*

Information about spatial proximity may be used to reduce the number of samplings that are necessary, this is because independently moving objects usually possess spatial cohesion. The selection process exploits spatial cohesion by randomly selecting one feature, then selecting six features randomly from the N (typically $N = 20$) nearest features (the distance being determined by the average of the image distance in the two images). For example an object may possess only 10% of the features within an image, but if a feature on the object is selected there is a much greater chance that adjacent features will also be on that object.

To estimate the best fitting constraint for each model type would be computationally inefficient. PLUNDER uses a guided search to reduce computation time providing the biggest estimate set consistent with each model. It also provides a mechanism to decide which of the models is most descriptive of the data. An ingenuous approach selects only the model consistent with the most correspondences. Generally, in the presence of noise and outliers, this scheme would always lead to the selection of the non-degenerate model, which might not necessarily be correct, in probabilistic decision process is used to determine the best model, following a minimum description length. The PLUNDER algorithm is applied repeatedly to extract models in order of size, hence effecting a segmentation that accounts for degeneracy.

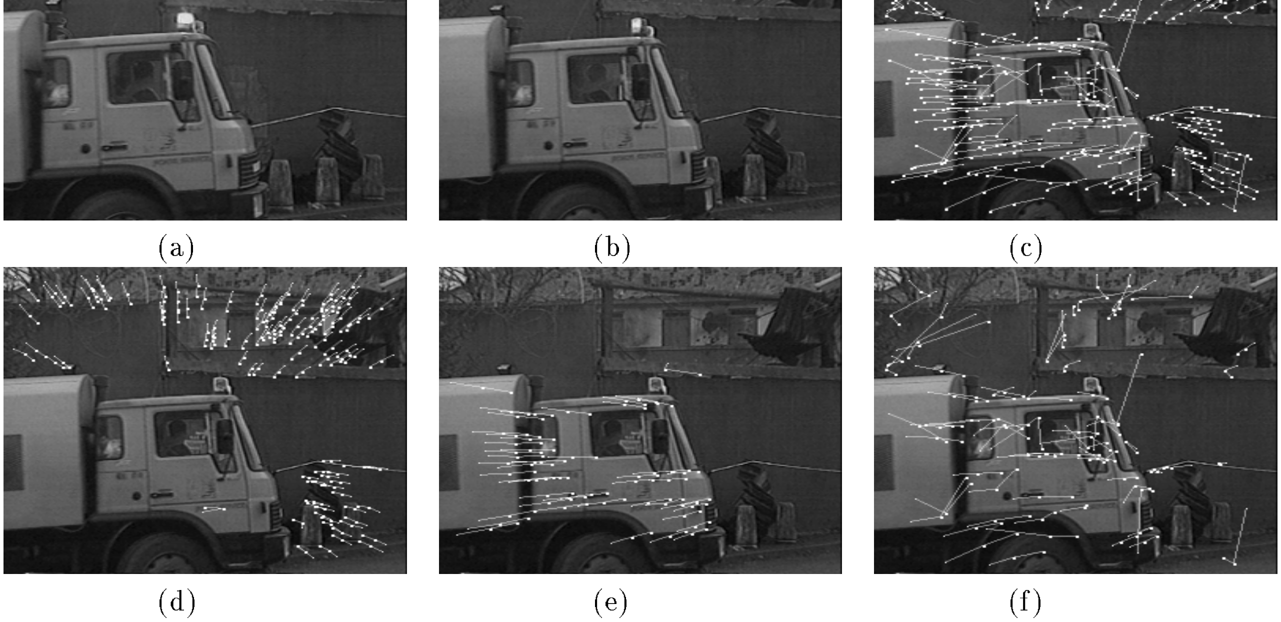


Figure 2: *(a)-(c) three consecutive images of a lorry translating to the right as the camera pulls away. In (c) the point correspondences made over the three views are recorded. (d) the algorithm correctly classifies correspondences on the background, (e) the lorry in the foreground, and (f) the outliers.*

Results

The corner detector we use is that suggested by Harris and Stephens [3] which calculates an interest operator defined according to an auto-correlation of Gaussian smoothed images. The initial matching is done on the basis of cross correlation. All processing is automatic.

Lorry Sequence: In Figure 2 (a)-(c) three consecutive images of a lorry translating to the right as the camera pulls away, together with the point matches each in superposition on the third lorry image (c). Epipolar geometry alone proved insufficient to segment this scene, application of the algorithm in Torr and Murray [14] gave rise to only one object. This is because, within the vicinity of the observed objects, the epipolar geometry of the lorry and background are locally similar. The use of the trifocal constraint gives much better results. Figure 2 (d) (e) show the clusters generated, neither is degenerate, (d) can be seen to be consistent with the background, (e) with the lorry, (f) gives the set of outliers or mis-matches.

Jeep Sequence: Figure 3 (a)-(c) shows a jeep tracked by a rotating camera, causing apparent motion to the left in the image. The jeep hits a pothole giving the semblance of the front moving down towards the bottom of the image. The output of the segmentation algorithm is shown as three disjoint sets of point and line features. The point and line correspondences for each set are shown superimposed on the last scene. Use of the trifocal constraint alone, without consideration of degeneracy, fails to segment this scene correctly. Because of the degenerate nature of the camera motion (pure rotation) the correspondences of the background and many of those on the independently moving jeep are all consistent with a single trifocal constraint. In fact the background is consistent with an image-image translation (a special case of image-image projectivities). This is because the majority of background scene points are quite distant and the camera is rotating about an axis parallel to the image plane. The PLUNDER algorithm extracts the largest group as the background and indicates that this is indeed consistent with an image translation, the points and lines shown in Figure 3 (d) (e) respectively. The detected lines in each view, consistent with the background, are shown together in (e). The motion is parallel to the bottom of the image from right to left, vertical lines over the three images may be seen as distinct, horizontal lines, on the background, are coincident. The other group detected is that on the jeep indicated as consistent with an image-image affinity, its features shown in Figure 3 (f) (g), outliers (h) (i). It can be seen that the line segments are correctly partitioned between the jeep and background.

Conclusion and Discussion

The topics of parameter estimation, outlier detection, model selection are intertwined and a motion segmentation algorithm must consider all three together to avoid failure. Use of the trifocal constraint emphasizes the need to detect degeneracy–the increased number of parameters increases the likelihood of instability. In this paper it is shown how robust methods may be used to solve the problem of motion segmentation. Our methods are based on real world constraints (without the necessity of recovering structure) rather than image based approximations. The three view constraint gives a better segmentation than the two view, as it provides a better discriminatory ability between similar motions. The algorithm works well on real image data, and allows a unified approach for points and lines. The segmentation algorithm uses only three views and is suggested as a bootstrap or initialization process. Other filtering mechanisms need to be developed to improve the segmentation through subsequent images.

Acknowledgement Financial support was provided by UK EPSRC grant GR/J65372 and EU Esprit Project 6448 'VIVA'.

References

[1] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision , LNCS 588*, pages 563–578, 1992.

[2] 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.

[3] C. Harris and M.. Stephens. A combined corner and edge detector. In *Proc. Alvey Conf.*, pages 189–192, 1987.

[4] R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In G. Sandini, editor, *Proc. 2nd European Conference on Computer Vision , LNCS 588*, pages 579–87. Springer–Verlag, 1992.

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

[6] R. I. Hartley. Projective reconstruction from line correspondences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1994.

[7] J. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision.* MIT press, 1992.

[8] L. Quan. Invariants of 6 points from 3 uncalibrated images. In J. O. Eckland, editor, *Proc. 3rd European Conference on Computer Vision, LNCS 800/801*, pages 459–469. Springer-Verlag, 1994.

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

[10] L. S. Shapiro, A. Zisserman, and M. Brady. Motion from point matches using affine epipolar geometry. In J. O. Eckland, editor, *Proc. 3rd European Conference on Computer Vision, LNCS 800/801*, pages 161–166. Springer–Verlag, 1994.

[11] A. Shashua. Trilinearity in visual recognition by alignment. In *Proc. 3rd European Conference on Computer Vision, LNCS 800/801*, volume 1, pages 479–484, May 1994.

[12] P. H. S. Torr. *Outlier Detection and Motion Segmentation.* PhD thesis, University of Oxford, 1995. In preparation.

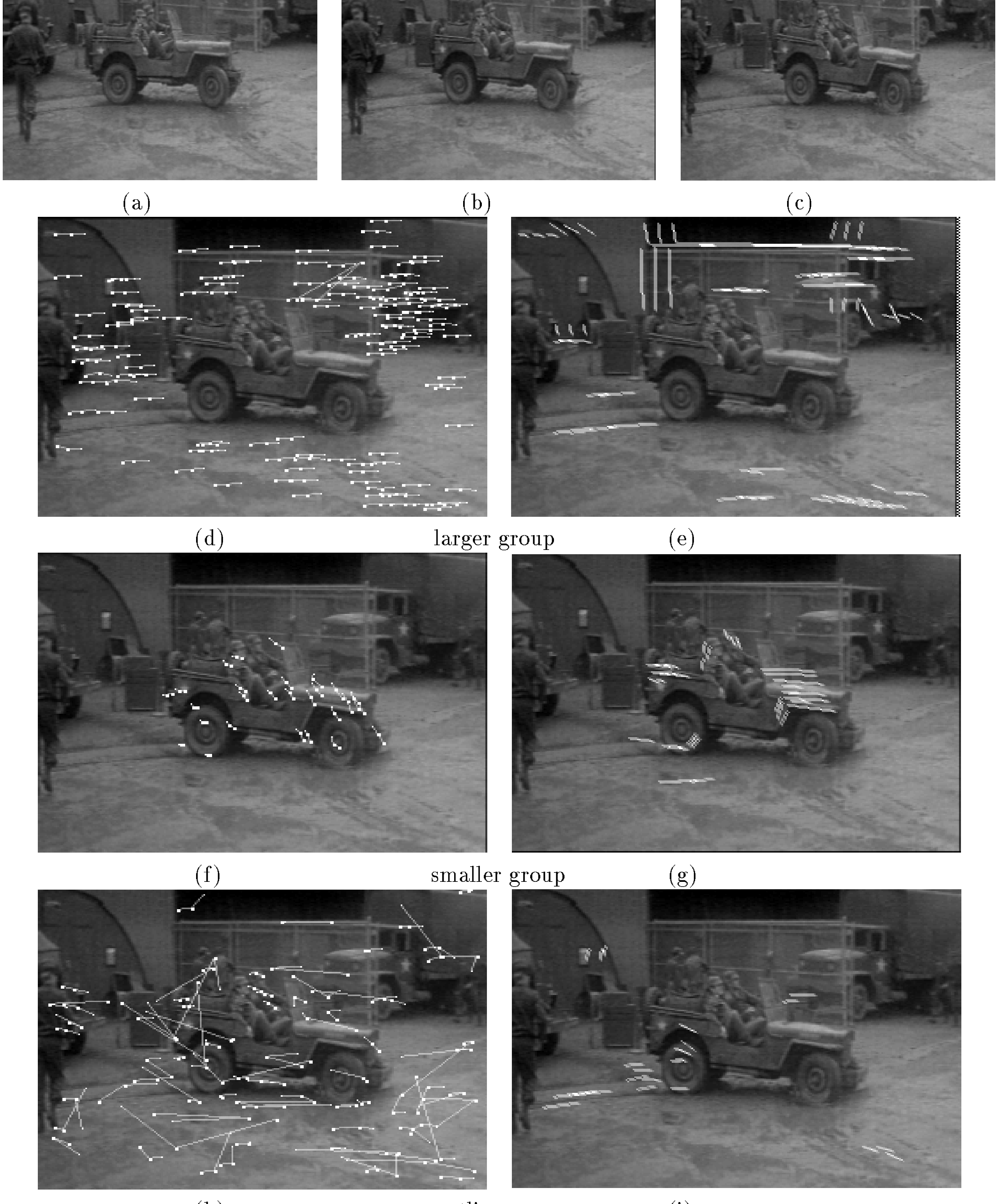
[13] P. H. S. Torr and D. W. Murray. Outlier detection and motion segmentation. In P. S. Schenker, editor, *Sensor Fusion VI*, pages 432–443. SPIE volume 2059, 1993. Boston.

[14] P. H. S. Torr and D. W. Murray. Stochastic motion segmentation. In J. O. Eckland, editor, *Proc. 3rd European Conference on Computer Vision, LNCS 800/801*, pages 328–338. Springer–Verlag, 1994.

[15] P. H. S. Torr, A. Zisserman, and S. Maybank. Robust detection of degeneracy. to appear in ICCV95, 1995.

[16] S. Ullman and R. Basri. Recognition by linear combination of models. *PAMI*, vol.13(10):992–1006, 1991.

[17] J. Weng, N. Ahuja, and T. Huang. Optimal motion and structure estimation. *IEEE PAMI*, vol.15(9):864–884, 1993.


(a) (b) (c)

(d) larger group (e)

(f) smaller group (g)

(h) outliers (i)

Figure 3: *(a)–(c) three consecutive images of a jeep translating to the right as the camera rotates to keep it in view. (d) correspondences in the most substantial group–the background. The line correspondences detected for this group in each of the three images are displayed together as triplets in (e). The background motion is to the left, hence vertical lines appear superimposed and horizontal lines, over the three images are distinct. The background is found to be consistent with an image translation. (f) point (g) line correspondences in the second group, consistent with an image-image affinity. (h) point correspondence outliers and (i) line correspondence outliers. Note the small cluster of outlying features on the soldier running towards the jeep to the left of the image.*