Affine and Projective Structure from Motion

Sabine Demey*
Katholieke Universiteit Leuven
Department of Mechanical Engineering, Division PMA
Heverlee, Belgium

Andrew Zisserman and Paul Beardsley†
Robotics Research Group, Department of Engineering Science
Oxford University, OX1 3PJ.

Abstract

We demonstrate the recovery of 3D structure from multiple images, without attempting to determine the motion between views. The structure is recovered up to a transformation by a 3D linear group - the affine and projective group. The recovery does not require knowledge of camera intrinsic parameters or camera motion.

Three methods for recovering such structure based on point correspondences are described and evaluated. The accuracy of recovered structure is assessed by measuring its invariants to the linear transformation, and by predicting image projections.

1 Introduction

A number of recent papers have discussed the advantages of recovering structure alone, rather than structure and motion simultaneously, from image sequences [4, 5, 6]. Briefly, structure can be recovered up to a 3D global linear transformation (affine or projective) without the numerical instabilities and ambiguities which normally plague SFM algorithms. In this paper we compare and evaluate three methods for obtaining such structure. The novelty of this approach is that camera calibration, extrinsic or intrinsic, is not required at any stage. The absence of camera calibration facilitates simple and general acquisition: Structure can be recovered from two images taken with different and unknown cameras. All that is required for unique recovery is point correspondences between images. Here the points are polyhedra vertices.

The three methods are labelled by the minimum number of points required to compute the epipolar geometry (see below). If more points are available a least squares minimisation can be used, though which error measure should be minimised for noise in non-Euclidean structure recovery is an unresolved question.

1. **4 point - affine structure**
   This assumes affine projection[1]. It requires the least number of points

*SD acknowledges the support of ERASMUS

†AZ and PAB acknowledge the support of the SERC

[1] a generalisation of weak perspective or scaled orthography, valid when the object depth is small compared to its distance from the camera, see the Appendix in [13].

BMVC 1992 doi:10.5244/C.6.6

of the three methods. The structure is recovered *modulo* a 3D affine transformation: Lengths and angles are not recovered. However, affine invariants are determined. For example: parallelism, length ratios on parallel lines, ratio of areas.

2. **6 point, 4 coplanar - projective structure**
   The camera model is a perspective pin-hole and structure is recovered *modulo* a 3D projective transformation (i.e. multiplication of the homogeneous 4-vectors representing the 3D points by a 4×4 matrix). Affine structure is not recovered, so parallelism cannot be determined. However, projective invariants, such as intersections, coplanarity and cross ratios can be computed. The invariants are described in more detail below.

   Only two more points (than the affine case) are required to cover perspective projection rather than its affine approximation. However, the planarity requirement is a limitation on the type of object to which the method is applicable.

3. **8 point - projective structure**
   Again perspective pin hole projection is assumed, and structure is recovered *modulo* a 3D projective transformation.

   This method makes no assumption about object structure, but requires more points than the other two methods.

The methods are described in sections 3 and 4.

Structure known only up to a 3D linear transformation, is sufficient to compute images from arbitrary novel viewpoints. The process of rendering new images given only *image(s)* of the original structure is known as *transfer*. This is described in section 2 and evaluated in section 5.2. Transfer has several significant visual applications:

1. **Verification in model based recognition**
   Model based recognition generally proceeds in two stages: first, a recognition hypothesis is generated based on a small number of image features; second, this hypothesis is *verified* by projecting the 3D structure into the image and examining the overlap of the projected structure with image features (edges) not used in generating the hypothesis. This is used routinely for planar objects [15] where the projections can be sourced from an image of the object - it is not necessary to measure the actual objects. The transfer method described here generalises this to 3D objects with similar ease of model acquisition - the model is extracted directly from images, and no camera calibration is required at any stage. In contrast, for 3D structures, previous verification methods (e.g. [9]) have required full Euclidean structure for the model and known camera calibration.

2. **Tracking**
   The performance of trackers, such as snakes or deformable templates, in efficiently tracking 3D structure, is markedly improved if the image position of the tracked features can be estimated from previous motion. This reduces the search region, which facilitates faster tracking, and gives greater immunity to the tracker incorrectly attaching itself to background clutter. The work here demonstrates that by tracking a small number of

features on an object it is possible to predict the image projection of the entire structure.

Since structure is recovered only up to a transformation, invariants to the transformation contain all the available (coordinate free) information. We assess the quality of the recovered structure by measuring these invariants. The invariants are described in sections 3 and 4 and evaluated in section 5.3.

2 Point Transfer

Transfer is most simply understood in terms of epipolar geometry. This is described here for the eight point case. In the other cases the principle is exactly the same, but less reference points are required.

It is assumed that two acquisition images, $imA1$ and $imA2$, have been stored with n known point correspondences ($n > 8$). Consider transfer to a third image, imT. The epipolar geometry between $imA1$ and imT is determined from eight reference point correspondences. A ninth point (and any other point) then generates an epipolar line in imT. Similarly, between $imA2$ and imT the epipolar geometry is determined, and each extra point defines an epipolar line in imT. The transferred point in imT lies at the intersection of these two epipolar lines. (Note, it is not necessary to explicitly compute correspondences between $imA2$ and imT. Once the correspondences between $imA1$ and imT are known, the correspondences required between $imA2$ and imT are determined from the acquisition correspondences between $imA1$ and $imA2$).

3 Affine invariants and point transfer

This approach is similar to that adopted by [6, 14, 17] and Barrett in [13].

3.1 Affine invariants

The 3D affine group is 12 dimensional, so for N general points we would expect[2] $3N - 12$ affine invariants - i.e. 3 invariants for each point over the fourth. The four (non-coplanar) reference points[3] $\mathbf{X}_i, i \in \{0, .., 3\}$ may be considered as defining a 3D affine basis (one for the origin $\mathbf{X}_0$, the other three specifying the axes $\mathbf{E}_i = \mathbf{X}_i - \mathbf{X}_0 \quad i \in \{1, .., 3\}$, and unit point) and the invariants, α, β, γ, thought of as affine coordinates of the point, i.e. $\mathbf{X}_4 = \mathbf{X}_0 + \alpha\mathbf{E}_1 + \beta\mathbf{E}_2 + \gamma\mathbf{E}_3$.

Under a 3D affine transformation (with $\mathtt{A}$ a general 3×3 matrix and $\mathbf{T}$ a 3-vector), $\mathbf{X}' = \mathtt{A}\mathbf{X} + \mathbf{T}$ the transformed vectors are $\mathbf{X}'_4 - \mathbf{X}'_0 = \alpha\mathbf{E}'_1 + \beta\mathbf{E}'_2 + \gamma\mathbf{E}'_3$, which demonstrates that α, β, γ are affine invariant coordinates. Following the basis vectors in this manner (they can be identified after the transformation) allows the retrieval of α, β, γ after the transformation by simple linear methods.

Projection with an affine camera may be represented by $\mathbf{x} = \mathtt{M}\mathbf{X} + \mathbf{t}$, where $\mathbf{x}$ is the two-vector of image coordinates, $\mathtt{M}$ is a general 2×3 matrix, $\mathbf{t}$ a general

[2] By the counting argument in the introduction of [13].

[3] We adopt the notation that corresponding points in the world and image are distinguished by large and small letters. Vectors are written in bold font, e.g. $\mathbf{x}$ and $\mathbf{X}$. $\mathbf{x}$ and $\bar{\mathbf{x}}$ are corresponding image points in two views.

2-vector, and $\mathbf{X}$ a three vector for world coordinates. Differences of vectors eliminate $\mathbf{t}$. For example the basis vectors project as $\mathbf{e}_i = \mathbf{M}\mathbf{E}_i \quad i \in \{1,..,3\}$. Consequently,

$$\mathbf{x}_4 - \mathbf{x}_0 = \alpha\mathbf{e}_1 + \beta\mathbf{e}_2 + \gamma\mathbf{e}_3 \tag{1}$$

A second view gives

$$\bar{\mathbf{x}}_4 - \bar{\mathbf{x}}_0 = \alpha\bar{\mathbf{e}}_1 + \beta\bar{\mathbf{e}}_2 + \gamma\bar{\mathbf{e}}_3 \tag{2}$$

Each equation (1) and (2) imposes two linear constraints on the unknown α, β, γ. All the other terms in the equations are known from image measurements (for example the basis vectors can be constructed from the projection of reference points $\mathbf{X}_i, i = \{0,..,3\}$). Thus, there are four linear simultaneous equations in the three unknown invariants α, β, γ, and the solution is straightforward.

3.2 Epipolar line construction

Equation (1) gives two linear equations in three unknowns, which determines β and γ in terms of α, namely:

$$\begin{aligned} \beta &= [v(\mathbf{x}_4 - \mathbf{x}_0, \mathbf{e}_3) - \alpha v(\mathbf{e}_1, \mathbf{e}_3)]\,/v(\mathbf{e}_2, \mathbf{e}_3) \\ \gamma &= [-v(\mathbf{x}_4 - \mathbf{x}_0, \mathbf{e}_2) + \alpha v(\mathbf{e}_1, \mathbf{e}_2)]\,/v(\mathbf{e}_2, \mathbf{e}_3) \end{aligned}$$

where the notation $v(\mathbf{a}, \mathbf{b}) = a_x b_y - a_y b_x$.

These are used to generate the epipolar line in another view. From (2) $\bar{\mathbf{x}}_4$ lies on the line

$$\begin{aligned} \bar{\mathbf{x}} = \bar{\mathbf{x}}_0 &+ [v(\mathbf{x}_4 - \mathbf{x}_0, \mathbf{e}_3)\bar{\mathbf{e}}_2 - v(\mathbf{x}_4 - \mathbf{x}_0, \mathbf{e}_2)\bar{\mathbf{e}}_3]\,/v(\mathbf{e}_2, \mathbf{e}_3) \\ &+\alpha\,(\bar{\mathbf{e}}_1 + [-v(\mathbf{e}_1, \mathbf{e}_3)\bar{\mathbf{e}}_2 + v(\mathbf{e}_1, \mathbf{e}_2)\bar{\mathbf{e}}_3]\,/v(\mathbf{e}_2, \mathbf{e}_3)) \end{aligned}$$

which is the equation of a line parameterised by α. Note, all epipolar lines are parallel with a direction independent of $\mathbf{x}_4$.

4 Projective invariants and point transfer

4.1 6 point, 4 coplanar, projective transfer

The 3D projective group is 15 dimensional so for N general points we would expect $3N - 15$ independent invariants. However, the coplanarity constraint loses one degree of freedom leaving only two invariants for the 6 points.

The meaning of the 3D projective invariants can most readily be appreciated from figure 1. The line formed from the two off plane points intersects the plane of the four coplanar points in a unique point. This construction is unaffected by projective transformations. There are then 5 coplanar points and consequently two plane projective invariants - which are also invariants of the 3D transformation.

4.1.1 Epipolar geometry

The algorithm for calculating the epipolar geometry is described briefly below, more details are given in [1, 12]

We have 6 corresponding points $\mathbf{x}_i, \bar{\mathbf{x}}_i, i \in \{0,..,5\}$ in two views, with the first 4 $i \in \{0,..,3\}$ the projection of coplanar world points.

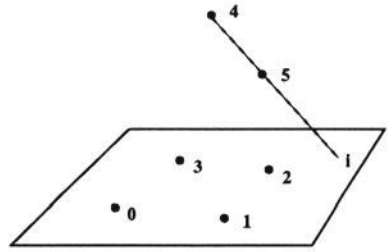

Figure 1: The projective invariant of 6 points, 4 coplanar (points 0-3), can be computed by intersecting the line through the non-planar points (4 and 5) with the common plane. There are then 5 coplanar points, for which two invariants to the plane projective group can be calculated.

1. Calculate plane projective transformation matrix $\mathtt{T}$, such that $\bar{\mathbf{x}}_i = \mathtt{T}\mathbf{x}_i, i \in \{0, .., 3\}$.
2. Determine the epipole, $\bar{\mathbf{p}}$, in the $\bar{\mathbf{x}}$ image as the intersection of the lines $\mathtt{T}\mathbf{x}_i \times \bar{\mathbf{x}}_i,\ i \in \{4, 5\}$.
3. The epipolar line in the $\bar{\mathbf{x}}$ image of any other point $\mathbf{x}$ is given by $\mathtt{T}\mathbf{x} \times \bar{\mathbf{p}}$.

4.1.2 Projective invariants

1. Determine the $\bar{\mathbf{x}}$ image of the intersection, $\bar{\mathbf{x}_I}$, of the plane and the line as the intersection of the lines $\mathtt{T}\mathbf{x}_4 \times \mathtt{T}\mathbf{x}_5$ and $\bar{\mathbf{x}}_4 \times \bar{\mathbf{x}}_5$ [14].
2. Calculate the two plane projective invariants of five points (in this case the four coplanar points and $\bar{\mathbf{x}_I}$) by

$$I_1 = \frac{|m_{320}||m_{I10}|}{|m_{310}||m_{I20}|} \qquad I_2 = \frac{|m_{310}||m_{I21}|}{|m_{321}||m_{I10}|}$$

where m_{jkl} is the matrix $[\bar{\mathbf{x}_j}\bar{\mathbf{x}_k}\bar{\mathbf{x}_l}]$ and $|m|$ its determinant.

4.2 8 point projective transfer

The construction described is a projective version [4, 5] of Longuet-Higgins' 8 point algorithm [7]. As is well known [8, 10] if points lie on a critical surface the epipolar geometry cannot be recovered. The method will clearly fail in these cases.

We have 8 corresponding points $\mathbf{x}_i, \bar{\mathbf{x}}_i, i \in \{0, .., 7\}$ in two views.

1. Calculate essential matrix $\mathtt{Q}$, such that $\bar{\mathbf{x}}_i^t \mathtt{Q} \mathbf{x}_i = 0, i \in \{0, .., 7\}$.
2. The epipolar line in the $\bar{\mathbf{x}}$ image of any other point $\mathbf{x}$ is given by $\mathtt{Q}\mathbf{x}$.

5 Experimental results and discussion

The images used for acquisition and assessment are shown in figure 2.

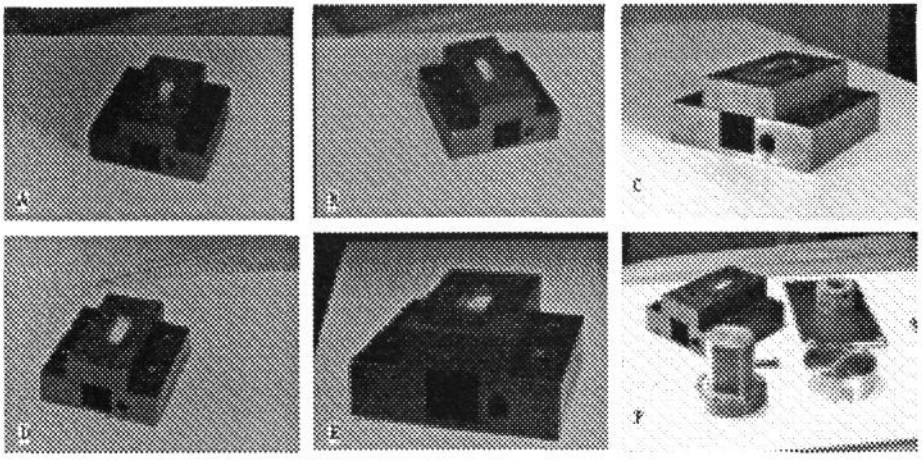

Figure 2: Images of a hole punch captured with different lenses and viewpoints. These are used for structure acquisition and transfer evaluation.

5.1 Segmentation and tracking

For the acquisition images the aim is to obtain a line drawing of the polyhedron. A local implementation of Canny's edge detector [2] is used to find edges to sub-pixel accuracy. These edge chains are linked, extrapolating over any small gaps. A piecewise linear graph is obtained by incremental straight line fitting. Edgels in the vicinity of tangent discontinuities ("corners") are excised before fitting as the edge operator localisation degrades with curvature. Vertices are obtained by extrapolating and intersecting the fitted lines. Figure 3 shows a typical line drawing.

Correspondence between views is achieved by tracking corners with a snake. This stage is currently being developed and some hand matching is necessary at present.

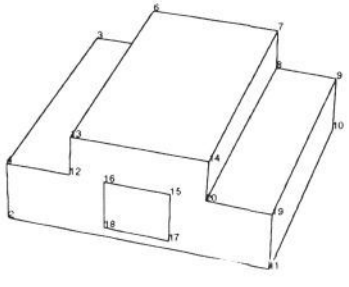

Figure 3: Line drawing of the hole punch extracted from image A in figure 2. Points 1 and 5 are occluded in this view.

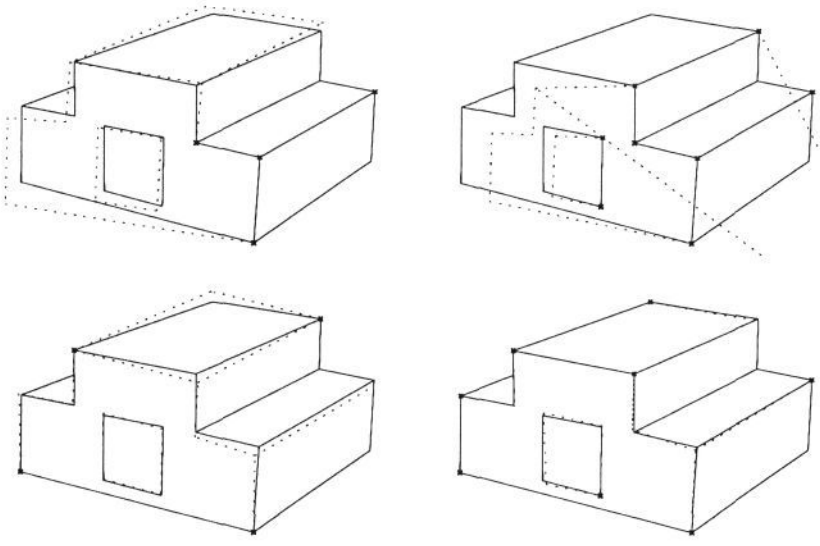

Figure 4: The effect of not spreading base points across the entire object. The transfer is computed from view A and B to view C. The "correct" graph structure of view C is shown by a solid line and the transferred view by a dashed line. "Correct" refers to the corners extracted from the real image. Base points are indicated by a cross. The left figures are affine transfer, the right projective 8 point transfer. Note, the graph structure is for visualization only, points not lines are transferred.

5.2 Transfer Results

The method is evaluated using two acquisition images plus a third image. Correspondence is established between the reference points in the third and acquisition images (e.g. 4 points in the case of affine transfer). Other points in the acquisition image are then transferred, the difference between the transferred points and the actual position of the corresponding points giving the transfer error. This error is taken as the Euclidean distance $d(\mathbf{p}_t, \mathbf{p}_c)$ between the transferred point $\mathbf{p}_t$, and its actual image position $\mathbf{p}_c$ (i.e. the position extracted from the actual image). Two measures are used for the evaluation:

1. mean error $E_{mean} = \frac{1}{n}\sum_{i=1}^{n} d(\mathbf{p}_t^i, \mathbf{p}_c^i)$
2. maximum error $E_{max} = \max_i \; d(\mathbf{p}_t^i, \mathbf{p}_c^i) \quad i \in \{1, .., n\}$

	Spread out points		Not spread out	
Method	mean error	max. error	mean error	max. error
affine	4.09	8.72	9.56	23.72
8 points	1.60	4.64	51.84	421.11

Table 1: Mean and maximum errors for the transfers shown in figure 4.

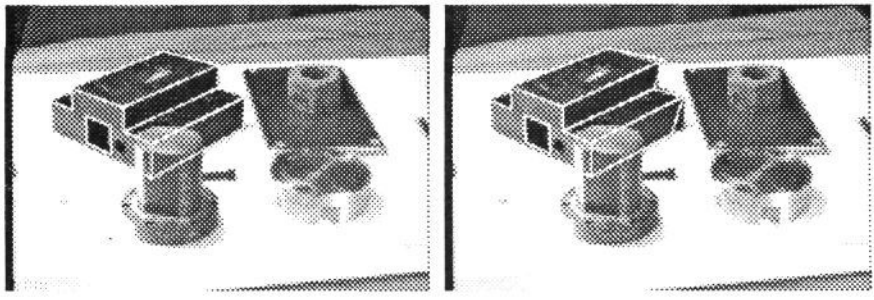

Figure 5: Typical transfers using the 6 point method with different base points. Left figure: using off plane points are 6 and 9, measured mean and max. errors are 3.88 and 15.88. Right figure: using off plane points 7 and 10, mean and max. errors are 1.76 and 4.91. See figure 3 for point numbering.

Method	mean error	max. error
affine	4.09	8.72
8 points	1.60	4.64
6 points (4 coplanar)	3.58	12.09
6+ points (5 coplanar)	2.93	5.58

Table 2: Comparison of typical results of the various transfer methods from view A and B to C. Note, the improvement in the 6 point method obtained by taking an additional coplanar point. The points used for the transfer are 11,13,14,17,(4),6,9. The point in brackets is the additional fifth coplanar point.

In both cases n is the number of points transferred. This varies between transfer methods as differing number of reference points are involved.

We have found that all methods perform poorly when the base points only partially cover the figure, see table 1 and figure 4. A similar result was noted in [11] in the case of planar transfer.

The affine transfer methods is very stable and does not suffer from the dramatic errors shown in the projective case (see figure 4). However, as would be expected, its performance degrades as perspective effects become more significant.

The six point transfer method can produce very good results, but success is very dependent on the "correctness" of the four point planarity. Of course, four real world points are never coplanar, but here there are additional errors from measured image positions. Some typical examples are given in figure 5. Stability can be improved by using more than four coplanar points to estimate the plane to plane projection matrix. This least squares estimate tends to cancel out some of the localisation errors in the image measurements (a total of 7 points for projective transfer, is still an advantage over the eight point method). This improvement is demonstrated in table 2 which also compares the three methods for a typical example. The eight point method does achieve the best performance, but the price paid is additional complexity of finding and matching extra points.

Images	α	β	γ
A,B	0.610	-0.221	-0.009
A,C	0.664	-0.268	-0.023
A,D	0.597	-0.213	-0.002
B,D	0.594	-0.214	-0.003
B,C	0.636	-0.242	-0.012
C,D	0.681	-0.285	-0.031
A,B,D	0.601	-0.217	-0.004
B,C,D	0.637	-0.244	-0.014
A,C,D	0.670	-0.274	-0.027
A,B,C	0.637	-0.243	-0.013

Images	I_1	I_2
D,A	0.440	-0.968
D,B	0.378	-1.117
B,A	0.371	-1.170
C,E	0.370	-1.150
F,A	0.333	-1.314
D,A,B	0.372	-1.151
C,E,D	0.369	-1.148
F,A,C	0.370	-1.196
C,A,B,D,E	0.375	-1.140
F,A,B,C,D,E	0.369	-1.170

Table 3: Invariants for varying sets of images. **Left**: affine coordinates (α, β, γ) of point 20 with respect to the base points 11,13,2,7. Note, measurements are spread over a smaller range when more images are used. **Right**: 6 point invariants using points 2,4,14,17 and the line between points 6 and 13. See figure 3 for point numbering.

5.3 Invariance Results

We find in general that the invariant values are more stable than transfer would suggest. This is probably because extra errors are incurred in measuring reference points in the transfer image.

5.3.1 Affine invariants

Equation (1) and (2) are four linear constraints on the three unknown affine invariants. Least-squares solution (by using singular value decomposition) immediately confers some immunity to noise. Further improvement is obtained by including corresponding equations from additional views. The stability and benefit of additional views is illustrated in table 3. In a tracked sequence robust estimates can be built in real-time using a recursive filter.

5.3.2 Projective invariants

Although invariants obtained from two views are fairly stable, improvements in stability are again achieved by augmenting with measurements from other views. See table 3. In this case by providing a least squares estimate of the line plane intersection.

6 Conclusion

In this paper, we have presented three methods to recover structure from two or more images taken with unknown cameras. All that is required is point correspondences between the images. Structure is recovered up to a linear transformation, but this is sufficient for transfer and computation of invariants of the 3D point set.

Experimental results show the methods perform well except for some sensitivity to corner detection errors. Future work will be based on automatically tracking corners using snakes.

Acknowledgements

We are grateful for helpful discussions with Roberto Cipolla, Richard Hartley, Joe Mundy and David Murray. Rupert Curwen provided the snake tracker and Charlie Rothwell the segmentation software.

References

[1] Beardsley, P., Sinclair, D., Zisserman, A., Ego-motion from Six Points, Insight meeting, Catholic University Leuven, Feb. 1992.

[2] Canny J.F. "A Computational Approach to Edge Detection," *PAMI*-6, No. 6. p.679-698, 1986.

[3] Curwen, R.M., Blake, A. and Cipolla, R. Parallel Implementation of Lagrangian Dynamics for real-time snakes. Proc. BMVC91, Springer Verlag, 29-35, 1991.

[4] Faugeras, O., What can be seen in 3D with an uncalibrated stereo rig?, *ECCV*, 1992.

[5] R. Hartley, R. Gupta and Tom Chang, "Stereo from Uncalibrated Cameras" Proceedings of CVPR92.

[6] Koenderink, J.J. and Van Doorn, A.J., Affine Structure from Motion, *J. Opt. Soc. Am. A*, Vol. 8, No. 2, p.377-385, 1991.

[7] Longuet-Higgins, H.C., A Computer Algorithm for Reconstructing a Scene from Two Projections, *Nature*, Vol. 293, p.133-135, 1981.

[8] Longuet-Higgins, H.C., The Reconstruction of a Scene from two Projections - Configurations that Defeat the 8-point Algorithm, *Proc. 1st IEEE Conference on Artificial Intelligence Applications*, p.395-397, December 1984.

[9] Lowe, D.G., *Perceptual Organization and Visual Recognition*, Kluwer, 1985.

[10] Maybank, S.J., Longuet-Higgins, H.C., The Reconstruction of a Scene from two Projections - Configurations that Defeat the 8-point Algorithm, *Proc. 1st IEEE Conference on Artificial Intelligence Applications*, p.395-397, December 1984.

[11] Mohr, R. and Morin, L., Relative Positioning from Geometric Invariants, *Proc. CVPR*, p.139-144, 1991.

[12] Mohr, R., Projective geometry and computer vision, To appear in *Handbook of Pattern Recognition and Computer Vision*, Chen, Pau and Wang editors, 1992.

[13] Mundy, J.L. and Zisserman, A., editors, *Geometric Invariance in Computer Vision*, MIT Press, 1992.

[14] Quan, L. and Mohr, R., Towards Structure from Motion for Linear Features through Reference Points, *Proc. IEEE Workshop on Visual Motion*, 1991.

[15] Rothwell C.A., Zisserman A., Forsyth D.A., and Mundy J.L., "Using Projective Invariants for Constant Time Library Indexing in Model Based Vision", *Proc. BMVC91, Springer Verlag*, 62-70, 1991.

[16] Semple, J.G. and Kneebone, G.T. *Algebraic Projective Geometry*, Oxford University Press, 1952.

[17] Ullman, S. and Basri, R., Recognition by Linear Combination of Models, *PAMI*-13, No. 10, p.992-1006, October, 1991.