Camera Motion Determination from Dynamic Perceptual Grouping of Line Segments

Jonathan Lawn and Roberto Cipolla
Department of Engineering, University of Cambridge, CB2 1PZ, England
jml@eng.cam.ac.uk and cipolla@eng.cam.ac.uk

Abstract

We present here a discussion on the use of perceptual grouping to improve structure and egomotion recovery algorithms for monocular cameras. In particular we look at grouping lines to avoid the need for trinocular algorithms. We also present a number of methods for grouping lines, including a novel method that infers planar groups from those demonstrating a linear deformation.

1 Introduction

Computer vision is usually split into two sections: detection of the image structures (and their motions), and interpretation of these artifacts or *features* in three dimensions. Between these operations are two-dimensional image techniques, though they are often implicitly absorbed into the detection stage. These include the tracking of features using image velocity prediction, and the association of short line segments into curves. The latter is an example of *perceptual* (or pre-attentive) grouping: the association of features by their position, orientation and image intensities.

One form of interpretation is *structure-from-motion*, which studies how the unknown camera motion (*egomotion*) and scene structure can be determined from a series of image frames. The camera must move to provide sufficient information for three-dimensional reconstruction, and therefore all these algorithms use at least the first derivatives of the visual motion (or two views in a discrete formulation). Some also use higher derivatives (or more views) [4, 10, 17] but these require much better tracking of features and rely on small perspective effects or gross changes in camera motion, and are therefore less robust.

This paper discusses various methods of perceptual grouping, some of them novel, and the ways in which the structural inferences they make can be used in the structure-from-motion problem.

We shall be making extensive use of a technique called *equation counting* [31]. Stated simply this compares the number of independent measurements with the number of unknowns in a set of equations. If the former is greater, then generally a unique solution can be found, if it is smaller then there is only enough information to find a locus of possible solutions, and if they are the same then there will be a finite number of solutions.

Equation counting can be applied to calculate how many features and frames are necessary to determine the camera egomotion and the scene structure, without considering any specific algorithm. We shall be considering monocular views from an intrinsically calibrated camera (so that ray directions from each viewpoint are known, but not the distances or *depths*), but this technique is also valid for other camera models. We shall also be assuming that the scene is rigid. Our unknowns are therefore 6 variables for each viewpoint and the unknown structure. However, the choice of origin, orientation and scale for the scene reconstruction are arbitrary (the speed-scale ambiguity), and therefore do not need to be determined. We shall see that perceptual grouping will determine some of the structure of the scene without measurement, decreasing the number of unknowns.

This paper consists of a survey of the available image features and the information that they hold (Section 2), a description of the preliminary experiments undertaken (Section 3) and a review of other related work (Section 4).

2 Useful feature types

There are a number of useful feature types that can be used as representations of the image structure and motion for the algorithms which determine the camera egomotion and scene structure. These can be roughly split into *primitive* features that are actually detected in the image (and in this we include regions, usually referred to as an alternative to features), and *compound* features that are formed from preprocessing the primitive features in 2D, before 3D inference is applied.

2.1 Primitive feature types

There are three major types of primitive feature.

Patterned regions are being researched as a way of extracting the visual (image) motion including the *affine* coefficients (zeroth and first spatial derivatives of the visual motion). By assuming that small regions are planar in space these six velocity coefficients can be used to determine the depth (1 variable), the slant and tilt (2 variables) and the camera motion (3 remaining variables per affine region). Therefore two planes in two images each provide 6 affine coefficients which may, in principle, determine the egomotion (5 variables) and their depth, slant and tilt (6 variables) [21].

2D Fourier transforms of affinely deforming textures also deform affinely (with the coefficients of the two deformations being very closely related) [32]. As features in the frequency domain are often easier to locate and disambiguate, this may be the best way to find the coefficients of the affine deformation. Multi-scale methods have also been produced for segmentation of scenes [27, 28, 36], and these two methods could be combined. However no system has yet been demonstrated to segment images accurately and quickly enough for real-time structure estimation, without using prior knowledge (perhaps from other primitive feature types, as below). The method is very attractive however, because it can represent most of the textured parts of the image.

Corners have 2D image structure and can therefore be tracked to give the local image motion [11]. They are assumed to represent fixed points in space, though often they will not (for instance, where one edge occludes another). Equation counting shows that the image velocities of six of these points (2 measurements each) can determine their depths in space (1 variable each) and the egomotion (5 variables). Unfortunately, 2D tracking is slow and liable to fail because of the increased search dimension. Also sparse sets of points do not describe space well, and are therefore not very useful to navigation algorithms on their own, though Delaunay triangulation is often used to infer shapes [2]. The egomotion and independent motion information extracted may facilitate the interpretation of other parts of the image.

The orientation and rotational velocity of the corner can also be determined (assuming that the deformation of the image region is low). This may be obtainable directly from the corner detector or may be determined once the corner has been located, and may be useful in disambiguation [15]. Unfortunately, the extra coefficient measured is not sufficient to determine any of the extra structure, even assuming the corner is a surface marking on a plane.

Line segments are regions of high image intensity gradient, and are usually assumed to correspond to line segments in space, though they may be occluding contours of curved surfaces. They are comparatively easy to detect and track [9] and, since they naturally connect, they can provide reinforcement for one another in each image, unlike corners. Line segments also delimit space and the image much better than corners, though planes are obviously still better. Unfortunately the normal velocity (perpendicular to the line) and rotational velocity (2 measurements) can only determine the depth and slant of the line in space (2 variables) if the camera motion is already known. But six lines in three frames can, in principle, determine the camera egomotion and scene structure [17, 33].

Some inferences can be made that associate short line segments into compound features however, allowing averaging which increases robustness, or even camera motion determination (without the use of second derivative or trinocular algorithms), as shown below.

2.2 Compound feature types

Compound features are formed by *perceptual grouping*, which is an attempt to invert certain projective identities. Evidence for these groupings can be gathered over a number of frames.

We shall be concentrating on the perceptual grouping of line segments, though work has been done on recognising patterns of corners [12] and segmenting patterned regions [32]. Each grouping method is described below by its implicit perceptual assumption, the inverse projective identity being obvious.

Continuity: If line segments form a continuous curve in the images, then they form one in space. Grouping line segments into curves does not change the number of equations or unknowns.[1] Still, continuity of depth and camera motion implies that the visual motion must change smoothly along the curve, and B-Spline snakes use this to increase the motion estimate accuracy [6, 8].

Linearity: If line segments form a straight line in the images, then they form one in space. Without its endpoints this long line segment will only offer as much information as a single short one, but its image position and motion can be averaged along its length, making the measurements more accurate [9, 33].

Intersection: If three or more close line segments intersect in the images, then they do in space. (Two line segments will always intersect in the image but are

[1] However, in a sufficiently structured scene, interpolation can provide *epipolar tangencies* [23] which are aligned with the direction of motion. At these points the camera motion but not the depths are constrained and the egomotion can be recovered in theory [3].

unlikely to in space.) These will define a vertex of unknown depth (1 variable) with edges of unknown slant (1 variable each) to be determined by the image velocity of the vertex (2 measurements) and the line segments angular velocities (1 measurement each). This implies that the velocities of more than five vertices could determine the camera motion (as with point primitives, though in three images only two three-edge vertices are needed). Because of the undesirability of T-junctions (which represent occlusions, not vertices in general [35]), it is preferable that the linearity grouping is performed before the intersection grouping.

Using the proximity of the endpoints of the line segments as a cue allows L-junctions (with two line segments) to be detected, and decreases the number of false vertices detected. It also reduces the search space of possible vertices, but risks losing occluded vertices and introducing spurious ones.

The order of the edges around each vertex in the image should remain constant for opaque objects. A further perceptual criterion that can be applied is that the visual motion of the vertex inferred is compatible with the inference from other methods (eg. planar groups as below).

Parallelism: Parallel lines in space are well known to intersect at a vanishing point in full perspective images. These vanishing points can be used as *intersections* (at infinite depth) as above. However, in many images the weak perspective approximation [24] is valid, and therefore it is also worthwhile considering the extra assumption that any parallel lines are parallel in space. This association allows averaging of the visual motion for accuracy, and also subgrouping into planar sets.

Planarity: If nearby line segments (or a sufficient number of any features) are deforming affinely in the image, then they are planar in space. This uses the affine or weak perspective camera model, which is only valid in small regions of a full perspective view. Associating the line segments with an affine deformation determines their 2D motion, and this provides a constraint on the compatibility of such groupings.

One proposed method of utilising this involves the fitting of B-Spline snakes to curves in the image (compound features describing associated line segments) [5]. However, whether closed or open, most of the snakes (which have enough structure to give the complete affine transformation) will not be planar. Though this can be detected, it is not efficient to have to reject most candidates.

The method proposed here uses four or more nearby line segments, each providing a 2D constraint in 6D affine deformation Hough space. Obviously this grouping involves considerably more effort than the other methods mentioned, but it is for a much greater gain in information – two planar groups in two frames can determine the egomotion [21]. A simpler method than a search through Hough space is just to test overdetermined groups. (If parallel lines are predominant, then planar parallel lines can be tested for first.) Such a system is described in Section 3.1.

The association of primitive features into compound groups does not only imply structure and determine image motion by averaging overdetermined information. It can also imply a new expectation of the behaviour of the group in the future, allowing improved feature tracking. Another use is to segment the image into objects [34].

The eventual implementation of these methods should use all forms of compound feature in parallel on streams of images. Initially, however, the concept can be proven by testing a number of images with perceptual grouping methods applied separately. It would be preferable to be able to compound all types of primitive feature, but here we continue to restrict discussion to long straight line segments.

3 Experiments

A sequence of experiments have been carried out to test the feasibility of the above grouping methods. They have been kept simple deliberately, using the minimum number of frames, and the most exhaustive combination method that is practical. Eventual implementations will use frames sequentially and more efficient test samples.

3.1 Implementation

Long line segments are selected by hand and matched in four images of the same object. The endpoints are not given accurately (see Figure 1). The test for colinearity is performed first, and the compound features formed replace their components for the following tests for other types of grouping (Figure 3(a)). In the tests where velocities are used (those for planar and parallel planar lines) the motions from frames 1-2, 2-3 and 3-4 were used.

Colinearity: Colinearity is determined by finding a line of best fit for the endpoints of the lines, and then measuring the sum of the squared errors of the endpoints from this line.

Parallelism: Parallelism is determined by finding the parallel lines that best fit the endpoints observed, and then measuring the sum of the squared errors of the endpoints from these lines. Triplets of lines displaying uniform expansion are labelled as coplanar.

Intersection: The best intersection is defined as that point which is closest to the lines (in a least squares sense). The degree of intersection of a set of lines is then determined by measuring the least sum of the squared distances of the endpoints from lines passing through this best intersection.

Planarity: In the proposed final implementation, sets of lines (or colinear groups of lines) are found that correspond to affine transformations. To qualify as a group, a basis subset of the set of lines must define an affine transformation in a well conditioned manner. The groups are *built up* from all the possible minimal basis sets by adding the line that fits the deformation that they define best, and then repeating. However, this generates a large number of matches so, for simplicity, we have tested all sets of five lines for compatibility with an affine transformation, as defined by an error measure similar to those of the other grouping methods above, summing the squared errors from the affine model (see Figure 2).

This still produces a large number of possible groups, but there are a couple of methods of pruning these. Firstly, any two lines that lie in a plane but are not colinear will define that plane, and therefore two planar groups that have more

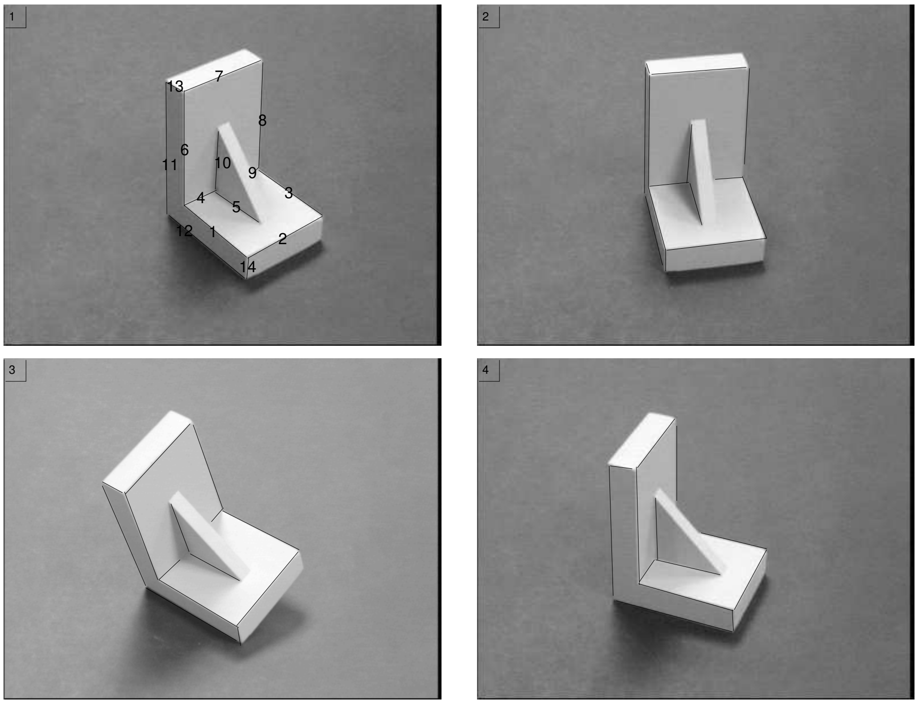

Figure 1: The four frames with 14 line segments selected

than one line in common are coplanar and therefore incompatible. Secondly, each group defines an affine transformation, and therefore the velocities of each lines endpoints, parallel to itself as well as perpendicular. The similarity between these endpoint velocity predictions define a score for each pair of planar groups that have no more than one line in common. The pair with the highest score are selected, as are then any other groups that are still compatible.

For each possible grouping, a best estimate of the line segment endpoints is made given the criterion of the grouping. Then the χ^2 test was used to determine whether the Mahalanobis distance between the measured endpoints and those of the estimate was significant or not [16]. The endpoint image measurements (which combine all the measurements of the lines position) are assumed to have a $\sqrt{2}$ pel (pixel width) standard deviation perpendicular to the line segments, and complete uncertainty parallel to them. The Mahalanobis distance is therefore the sum of the squared errors (normal to the lines) divided by 2 pel^2.

3.2 Results

Colinearity: As one would expect from such a sparse scene, detecting the colinear pair of line segments was not difficult: results were very conclusive (see Figure 3(a)).

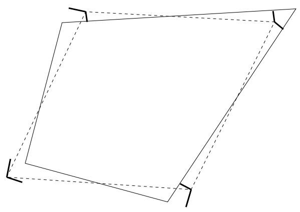

Figure 2: Example of the error estimation in a proposed affine deformation from a square. The diagram shows the measured lines (solid), the best affinely deformed model (dashed), and the errors of the measured lines from the model endpoints (bold).

Parallelism: Lines were correctly grouped (see Figure 3(a)), using a standard deviation of $2\sqrt{2}$ pel to allow for perspective effects. Experiments using images with stronger perspective effects suggest that it may be necessary to use the intersection groupings to find vanishing points instead. The four triplets of coplanar parallel lines were also found correctly.

Intersection: Each triplet of lines was tested for a common intersection in each frame. Four of the five vertices represented were found. The other (lines 4,8,13 in the new scheme) being defined too inaccurately by its lines. In images with strong perspective effects, most of the intersections found are vanishing points (here 17 out of 21 triplets), but there were no other false vertices.

Planarity: All sets of five lines were tested over all three frame pairs and these generated approximately 60 candidate plane groupings. The compatibility criteria scores (see Section 3.1) produced perfect results, identifying all three planes (see Figure 3(b-d)), though other groupings had almost as high scores. We hope to find that when closer frames are used, and velocities better represent the inter-frame displacements, these results will improve, even before more frames are integrated.

3.3 Discussion

The results are encouraging. The single frame groupings (colinearity, parallelism and intersection) perform well in single images, and nearly perfectly over the sequence. This alone could give us sufficient extra information to extract the egomotion and structure using a binocular algorithm (if enough intersections are found) whilst retaining and improving the robustness and accuracy of line tracking.

The planar grouping method is also promising, particularly given that these results came from only four frames. It is unfortunate however that such a large number of candidate planes are produced, and the use of constraints during the search procedure will be vital to keep the computations tractable. Use of the full vertex velocities should reduce the search space, and help ensure that the initial planes accepted are correct, reducing the search space further.

4 Relation to previous work

Though perceptual grouping in vision was considered by Gestalt psychologists, and early vision researchers mentioned it [20], it was not until Witkin and Tenenbaum

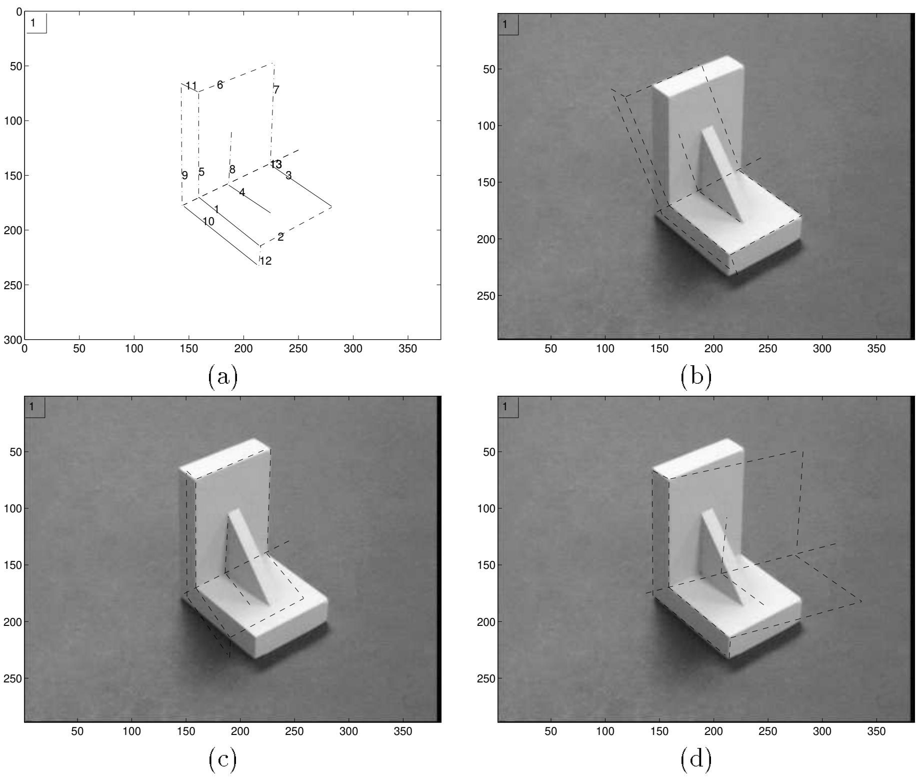

Figure 3: (a) The first frame, with the two colinear line segments (4 and 9 in Figure 1) replaced by their compound feature, showing the groups of parallel lines found. These were correctly grouped into four triplets, each lying in a plane. (b), (c) and (d) show affine deformations associated with the three correct planar groups between Frame 1 (shown) and Frame 2. The motion of the lines lying in the plane has been correctly modelled by the affine deformation, whereas the motion of those out of the plane has not.

[37] and Lowe [18] that it was considered seriously for computer vision. Sarkar and Boyer [26] offer a very good review of what work has been done in what areas of this field, and in what regions there has been little effort, particularly mentioning motion analysis and the role motion segmentation can play in improving algorithms. Sarkar and Boyer [25] have also offered a structure for the organisation of perceptual groups within an image, but there has been little or no work on the integration of a number of frames.

Line segments have also been the focus of a more structural approach, which we adopt here, not only for detecting patterns, but also for considering their meaning. A number of computational methods have been produced for finding vanishing points (common intersections) in sets of lines [7, 13, 14, 19], and Sawhney and Hanson [29] proposed a similar scheme to our affine groups, though they only consider similarity transformations (no shear) and therefore fronto-parallel planes (which they consider most interesting). Smith [30] and Adiv [1] also use affine deformation (of points and patterned regions respectively) to segment scenes. There has also been some variation in the constraints considered in the interpretation, from assuming a cuboid world of three vanishing points [22], to only assuming planar, pairwise-rigid motion [12]. However there has been little that demonstrates the smooth transition from image motion measurement to 3D understanding that can be achieved with perceptual grouping.

5 Conclusion

We have shown how inferences in the image domain imply scene structure, concentrating on those that associate long line segments, and have demonstrated that the criteria are sufficient to judge the validity of the groupings in our preliminary experiments. We have also shown that the extra structure implied by the groupings can allow new structure-from-motion algorithms which should be more reliable than current methods.

Future work is to include investigations into methods for combining the information from the individual compound features in single frames, and algorithms for computing egomotion and structure from the features, and will conclude with a real time implementation using line segment tracking.

Acknowledgements

Thanks to the CUED vision group, and especially Sven Vinther, for fruitful discussions on perceptual grouping.

References

[1] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 7(4):384–401, 1985.

[2] M. Buffa, O. Faugeras, and Z. Zhang. A complete navigation system for a mobile robot, using real-time stereovision and the delaunay triangulation. In *Proc. of MVA'92 (IAPR Workshop)*, pages 191–194, 1992.

[3] S. Carlsson. Sufficient image structure for 3-d motion and shape estimation. In *Proc. 3rd Euro. Conf. on Comp. Vis.*, pages I.83–91, 1994.

[4] R. Cipolla. *Active Visual Inference of Surface Shape.* PhD thesis, University of Oxford, 1991.

[5] R. Cipolla and A. Blake. Surface orientation and time to contact from image divergence and deformation. In G. Sandini, editor, *Proc. 2rd Euro. Conf. on Comp. Vis.*, pages 187–202. Springer–Verlag, 1992.

[6] R. Cipolla and A. Blake. Surface shape from the deformation of apparent contours. *Int. J. of Comp. Vis.*, 9(2):83–112, 1992.

[7] R.T. Collins and R.S. Weiss. Vanishing point calculation as a statistical inference on the unit sphere. In *Proc. 3rd Int. Conf. on Comp. Vis.*, pages 400–405, 1990.

[8] R. Curwen and A. Blake. Dynamic contours: real-time active splines. In A. Blake and A. Yuille, editors, *Active Vision.* MIT Press, 1992.

[9] R. Deriche and O. Faugeras. Tracking line segments. In O. Faugeras, editor, *Proc. 1st Euro. Conf. on Comp. Vis.*, pages 259–268. Springer–Verlag, 1990.

[10] O.D. Faugeras and T. Papadopoulo. A theory of the motion fields of curves. *Int. J. of Comp. Vis.*, 10(2):125–156, 1993.

[11] C.G. Harris. In A. Blake and A. Yuille, editors, *Active Vision*, chapter 16. MIT Press, 1992.

[12] D.D. Hoffman and B.E. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 42:195–204, 1982.

[13] K. Kanatani. Hypothesising and testing geometric properties of image data. *Comp. Vis., Graphics and Image Proc. (Image Understanding)*, 54(3):349–357, 1991.

[14] K. Kanatani. Renormalization for unbiased estimation. In *Proc. 4th Int. Conf. on Comp. Vis.*, pages 599–606, 1993.

[15] T. Lindeberg and J. Garding. Shape from texture from a multi-scale perspective. *Proc. 4th Int. Conf. on Comp. Vis.*, pages 683–691, 1993.

[16] B.W. Lindgren. *Statistical Theory.* Collier Macmillan, 1976.

[17] Y. Liu and T.S. Huang. Estimation of rigid body motion using straight line correspondances. *Comp. Vis., Graphics and Image Proc.*, 43:37–52, 1988.

[18] D.G. Lowe. *Perceptual Organisation and Visual Recognition.* Kluwer Ac., 1985.

[19] M.J. Magee and J.K. Aggarwal. Determining vanishing points in perspectiveimages. *Comp. Vis., Graphics and Image Proc.*, 26:256–267, 1984.

[20] D. Marr. *Vision.* Freeman, San Francisco, 1982.

[21] S. Negahdaripour and S. Lee. Motion recovery from image sequences using only first order optical flow information. *Int. J. of Comp. Vis.*, 9(3):163–184, 1992.

[22] P. Olivieri, M. Gatti, M. Straforini, and V. Torre. A method for the 3D reconstruction of indoor scenes from monocular images. In *Proc. 2rd Euro. Conf. on Comp. Vis.*, pages 696–700, 1992.

[23] J. Porrill and S. Pollard. Curve matching and stereo calibration. *Image and Vision Computing*, 9(1):45–50, 1991.

[24] L.G. Roberts. Machine perception of three - dimensional solids. In J.T. Tippet, editor, *Optical and Electro-optical Information Processing.* MIT Press, 1965.

[25] S. Sarkar and K. L. Boyer. A highly efficient computational structure for perceptual organization. Technical Report SAMPL-90-06, SAMPL-Laboratory, Department of Electrical Engineering, Ohio State University, November 1990.

[26] S. Sarkar and K.L. Boyer. Perceptual organization in computer vision: A review and a proposal for a classificatory structure. *IEEE Trans. on Systems, Man, and Cybernetics*, 23(2):382–399, 1993.

[27] J. Sato and R. Cipolla. Extracting the affine transformation from texture moments. In Jan-Olof Eklundh, editor, *Proc. 3rd Euro. Conf. on Comp. Vis.*, pages II.165–172. Springer–Verlag, 1994.

[28] J. Sato and R. Cipolla. Image registration using multi-scale texture moments. In *Proc. 5th British Machine Vis. Conference*, page (to appear), 1994.

[29] H.S. Sawhney and A.R. Hanson. Trackability as a cue for potential obstacle identification and 3D description. *Int. J. of Comp. Vis.*, 11(3):237–265, 1993.

[30] S.M. Smith. A scene segmenter; visual tracking of moving vehicles. *Engineering Applications of Artificial Intelligence*, 7(2), April 1994.

[31] K. Sugihara. An algebraic approach to shape-from-image problems. *Artificial Intelligence*, 23:59–95, 1984.

[32] T.F. Syeda-Mahmood. Model-driven selection using texture. In *Proc. 4th British Machine Vis. Conference*, pages 63–74, 1993.

[33] T. Vieville. Estimation of 3D-motion and structure from tracking 2D-lines in a sequence of images. In *Proc. 1st Euro. Conf. on Comp. Vis.*, pages 281–291, 1990.

[34] S. Vinther and R. Cipolla. Active 3D object recognition using affine invariants. In *Proc. 3rd Euro. Conf. on Comp. Vis.*, pages II.15–24, 1994.

[35] D. Waltz. Understanding line drawings of scenes with shadows. In P.H. Winston, editor, *The Psychology of Vision.* McGraw-Hill,New York, 1975.

[36] J. Weber and J. Malik. Robust computation of optic flow in a multi-scale differential framework. In *Proc. 4th Int. Conf. on Comp. Vis.*, pages 12–20, 1993.

[37] A. Witkin and J. Tenebaum. On the role of structure in vision. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human And Machine Vision*, pages 481–543. New York: Academic, 1983.