

Geometric grouping of repeated elements within images

F. Schaffalitzky and A. Zisserman
Department of Engineering Science
University of Oxford, UK
{fsm, az}@robots.ox.ac.uk

Abstract

The objective of this work is the automatic detection and grouping of imaged elements which repeat in a scene. We show that structures that repeat in the world (for example wall paper patterns) are related by particular parametrized transformations in perspective images. These image transformations provide powerful grouping constraints, and can be used at the heart of hypothesize and verify grouping algorithms.

Parametrized transformations are given for some classes of repeating operation in the world as well as some groupers based on these. These groupers are demonstrated on a number of real images, where both the elements and the grouping are determined automatically. It is also shown that the repeating element can be learnt from the image, and hence provides an image descriptor.

1 Introduction

The objective of this work is quite simple: suppose a structure is repeated in the world a number of times by some operation (for example a translation); then identify this structure and all its repetitions from a perspective image. The output is the imaged *element*, and a *grouping* over the imaged repetitions.

There are many reasons why such an objective is helpful for computer vision tasks. First, repetitions are common in the world — examples include parquet floor tilings, windows, bricks, patterns on fabrics, wallpaper; Second, the groupings provide a compact image descriptor — essentially a ‘high level’ feature — which may be used where ever a process involves image matching. For example in image database retrieval, model based recognition, and stereo correspondence; Third, the retrieved repeating operation can provide shape and pose information, for example the vanishing line of a plane, in a similar manner to that of shape-from-texture.

Image relationships as a basis for grouping have a healthy tradition in computer vision. The generic relationships, e.g. parallelism, identified by the Gestalt school have influenced several authors [1, 3, 10, 12]. More specialized relationships have been identified for certain classes of curved surface [11, 16, 17] and used for grouping. Specific relationships for repeated elements have been investigated recently by [7, 9, 14].

The repeating operation acts in the world, and we search for the repeated element in the image. Thus image grouping always has two components:

1. **Grouping geometry:** Given a repeating operation in the world, what geometric relationships are induced in the image between the imaged repeated elements?
2. **Grouping strategy:** Given these image relations, how are they best harnessed to facilitate grouping?

We explore grouping strategies for two classes of image relations:

1. Local affine transformations. This is described in section 2.
2. Global parametrized transformations arising from particular repeating operations in the world. This is described in section 3.

Grouping as an objective should have an associated and well defined measure of success — for example, that it enables a subsequent visual task. Here the measure of success is the ratio of the number of grouped elements to the number that can be grouped.

2 Local affine transformations

The idea is that, locally, repeated elements are approximately related in the image by affine transformations [7].

2.1 Overview of algorithm

This is essentially a hypothesize and verify algorithm. The main steps are summarized here, and then expanded on in the following sections.

1. **Getting started:** identify interesting image regions and hypothesize a grouping. The output is the elements together with the putative grouping.
2. **Verification:** a grouping is verified if the elements are mapped under local affine transformations. The output is a locally connected graph whose vertices are the elements and whose edges correspond to matched elements.
3. **Enlarge the groupings:** search for new elements by extrapolating on the estimated affine transformation.

2.2 Getting started

Initially we do not know the elements or the grouping. This is the chicken and egg problem that often arises in computer vision: if we know the elements we can (relatively) easily determine groupings; conversely, if we know the groupings we can (relatively) easily identify elements. We have two strategies for identifying elements: intensity based and contour based.

Intensity based Interesting image points are identified by detecting Harris corners [6]. Each corner is used to hypothesize a potential element in the vicinity of the corner. Typically this image patch is chosen as a small square region centred at the corner. The normalised cross-correlations between patches within a certain distance are computed and, if above a threshold, a grouping is hypothesised. Figure 1 illustrates the process. The patch is 12×12 pixels, the cross-correlation threshold 0.8, and the search distance 100 pixels.

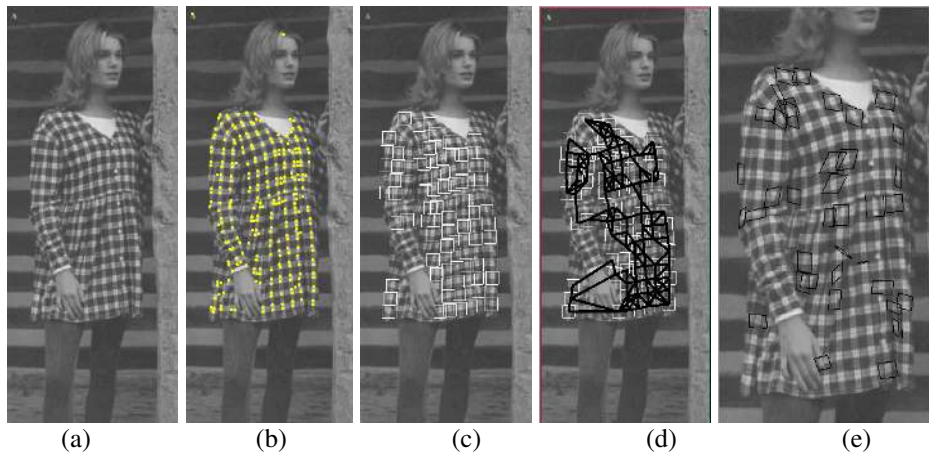


Figure 1: Grouping on local affine transformations I: using corners and cross-correlation. (a) Original image (b) Harris corners. (c) Intensity grouped regions. (d) Verified groupings. (e) New elements found by search.

Closed curves The idea here is to identify interesting regions by detecting closed Canny [2] edge contours, and then determine if these regions are related by affine transformations by computing their affine texture moment invariants [15]. Regions which are related by an affine transformation have the same value for affine invariants. Thus clustering on the invariants yields a putative grouping of regions. Six affine invariants are computed, so each curve gives a point in a 6-dimensional space. The points are clustered in this 6D space by the k-means clustering algorithm. The result is illustrated in figure 2.

2.3 Verification using local affine transformations

The input at this point is a collection of elements forming a putative grouping. The groupings are verified by attempting to register an element to each of its four closest neighbours. Registration is by an iterative warping algorithm [7] which minimizes the squared sum of differences (SSD) between the element and the neighbouring region. The criterion for good registration is the threshold on normalised cross-correlation between the warped element and its neighbour. Here the threshold is set at 0.80. The output of this stage is a graph whose vertices are the elements of the putative grouping and whose edges denote pairs of elements that register well.

2.4 Search for additional groupings

To search for missing elements, take an edge in the grouping graph. This represents an affine transformation T which registers the element A at its starting point with the element B at the end point. We hypothesize the existence of another element at the location given by applying T to the element B , and test this hypothesis by cross correlation as above. Additional elements are found by this method in both of figures 1 and 2.

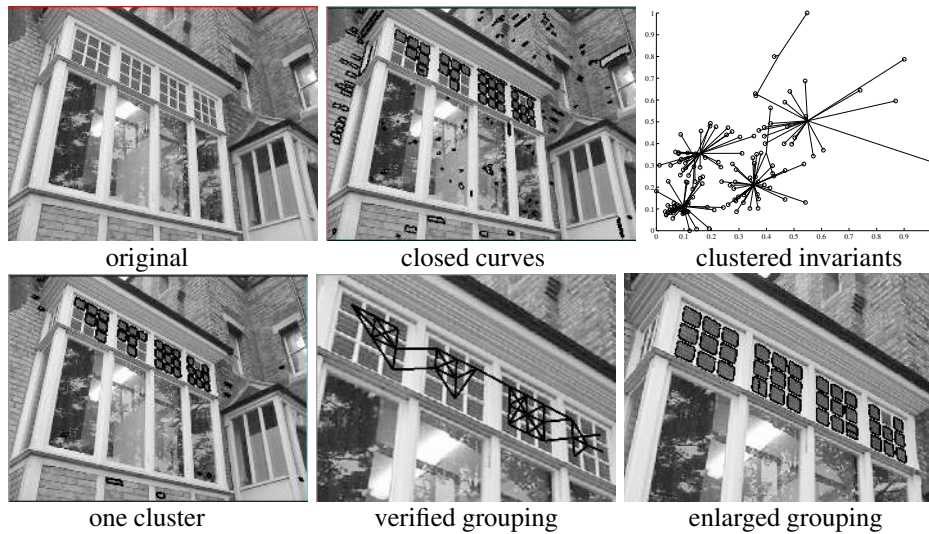


Figure 2: Grouping on local affine transformations II: using closed curves and moment invariants. The plot shows the distribution and clustering of the zeroth order moments of shape (horizontal axis) and intensity (vertical axis). The cluster used as a hypothesised grouping is the bottom left-most one.

3 Parametrized transformations

A particular repeating operation in the world induces a particular image transformation. There are two significant advantages of this class of induced image transformation:

1. The entities fixed by the transformation are geometrically significant. For example for repetitions on a plane the fixed line will correspond to the vanishing line of the plane.
2. There is a simple parametrization, often expressible in terms of the fixed entities.

Here we will investigate in detail one repeating operation, namely translations on a plane, for which the induced transformation is a conjugate translation. This will serve as an exemplar for the other classes of transformations that are described in section 5.

In the case of an imaged planar translation the induced transformation has only four degrees of freedom. This is two less than the canonical and ‘simple’ affine transformation used by many authors in the past for this type of grouping [7], yet the induced transformation exactly models perspective effects which are not accounted for by an affine transformation.

3.1 Conjugate translation

The repeating operation which gives rise to a conjugate translation in the image is illustrated in figure 3. The conjugate translation may be parametrized as:

$$H = I + \lambda \mathbf{v} \mathbf{l}_\infty^\top \quad \text{with} \quad \mathbf{v} \cdot \mathbf{l}_\infty = 0 \quad (1)$$

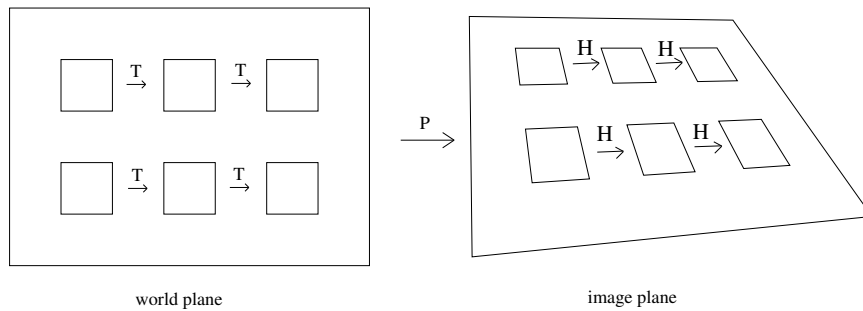


Figure 3: A translation T on a world plane gives rise to a conjugate translation H in the image.

where \mathbf{l}_∞ is a 3-vector representing the vanishing line of the plane, \mathbf{v} a 3-vector representing the vanishing point of the translation direction, and λ a scalar representing the magnitude of the translation. This transformation has four degrees of freedom: two for the fixed line \mathbf{l}_∞ (which is represented by a unit vector), one for the fixed point \mathbf{v} (which is represented by a unit vector orthogonal to \mathbf{l}_∞), and one for λ . A few remarks on this transformation: The transformation applies to two elements repeated by the translation anywhere on the image plane. If there is a line of repetitions (as in figure 4) then the zeroth element is mapped to the n -th as $H = I + n\lambda\mathbf{v}\mathbf{l}_\infty^\top$. The transformation (1) can be determined from two point or two line correspondences. Once the transformation is determined, then so is \mathbf{l}_∞ .

The parametrization (1) can be derived in several ways. One is to start from the homography induced by a plane between two cameras [4]

$$H = C'(\mathbf{R} + \mathbf{t}\mathbf{n}^\top/d)C^{-1}$$

where \mathbf{R} is the rotation and \mathbf{t} the translation between the cameras, C and C' are the camera internal parameters for the first and second view, and the plane has the equation $\mathbf{n}\cdot\mathbf{X} = d$. Repeating by a translation in the plane is equivalent to the images obtained from two identical cameras translated in a direction parallel to the plane. This means that $\mathbf{R} = I$, $\mathbf{n}\cdot\mathbf{t} = 0$, $C = C'$. The parametrization is obtained by setting $\mathbf{v} = C\mathbf{t}$ and $\mathbf{l}_\infty = C^{-\top}\mathbf{n}$.

3.2 Grids

A variation on the conjugate translation is where there is a repetition in two directions so that the world pattern is a grid of repeated elements. The image is then a conjugate grid. This mapping can be thought of as being composed of two elements

$$H_{\mathbf{v}} = I + \lambda\mathbf{v}\mathbf{l}_\infty^\top \quad H_{\mathbf{u}} = I + \mu\mathbf{u}\mathbf{l}_\infty^\top$$

one for each direction \mathbf{u}, \mathbf{v} , i.e. a total of six degrees of freedom. However, note that \mathbf{l}_∞ is common to both, so that once the transformation is determined in one direction only two degrees of freedom remain for the transformation in the other direction. These two degrees of freedom can be determined by one point correspondence.

4 Grouping with parametrized transformations

Since only a small number of parameters are needed to describe the parametrized transformations, only a small number of feature correspondences are required to estimate a putative transformation. Additional feature correspondences then enable this putative transformation to be verified. This facilitates a very efficient grouping algorithm, which is summarized here and then illustrated by examples in figures 4, 5 and 6.

The algorithm is based on RANSAC [5] and can be applied to any type of parametrized transformation. A robust estimation algorithm, like RANSAC, is required because some of the putative elements and groupings may well be wrong, and it is necessary to identify these 'outliers' when estimating the transformation.

4.1 Grouping under conjugate translations

The algorithm may be used to group any feature type. For example, the features could be the interest points and closed contours of section 2.2. Here we specialize the algorithm to grouping the intersections of line segments under a conjugate translation.

1. **Seed elements and groupings:** Find intensity step edges in the image using an edge detector. Fit straight line segments to resulting contours. Find pairs of intersecting line segments. Generate putative ('seed') correspondences using cross-correlation of intensity neighbourhoods.
2. **Ransom sampling from seed groupings:** using the seed match determine the transformation and then verify. The model (the conjugate translation transformation) can be computed from either (a) two line correspondences, no two of which are collinear or (b) two line correspondences, two lines of which are collinear, and one point correspondence on the other two lines. We score this hypothesised model by the number of correspondences consistent with it and keep those models with the highest score.
3. **Maximum Likelihood Estimation (MLE):** Estimate the four parameters of the transformation from the best inlier sets that arose from the random sampling.
4. **Guided search:** using the estimated parameters, search for new elements consistent with the model and reestimate the parameters.

See figures 4 and 5 for illustrations.

4.2 Grouping projective grids

A grid grouper can be implemented with much the same approach as that of a conjugate translation grouper, i.e. hypothesize and verify on the parametrized transformation. An alternative is to partition the problem into two components, and first identify the 1D repetitions of a grid (the lines) and then group these. In outline the latter approach is the following :

The elements are first organised along straight lines, accomplished by a simple robust estimation (RANSAC) of collinear points in a point set. Usually 10 lines are sufficient. On each line found we seek points that are arranged in a regular one-dimensional grid.

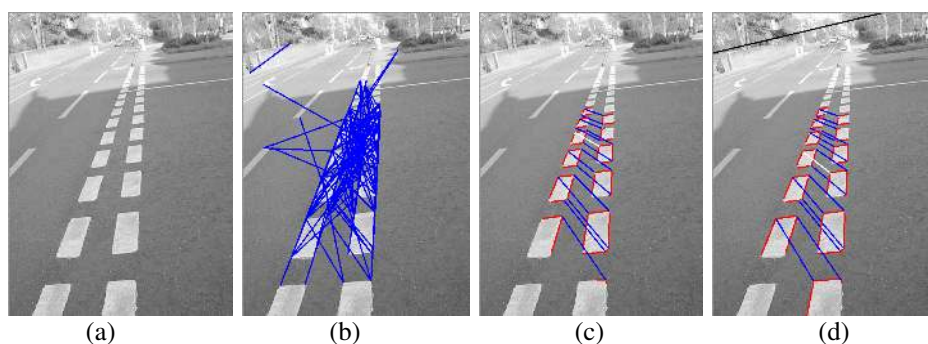


Figure 4: Conjugate translation grouping of line pair intersections. (a) Original image. (b) Seed correspondences based on cross-correlation. (c) Correspondences consistent with a seed (shown in white). (d) Result of guided search and the vanishing line of the plane (shown in black) computed from the MLE.

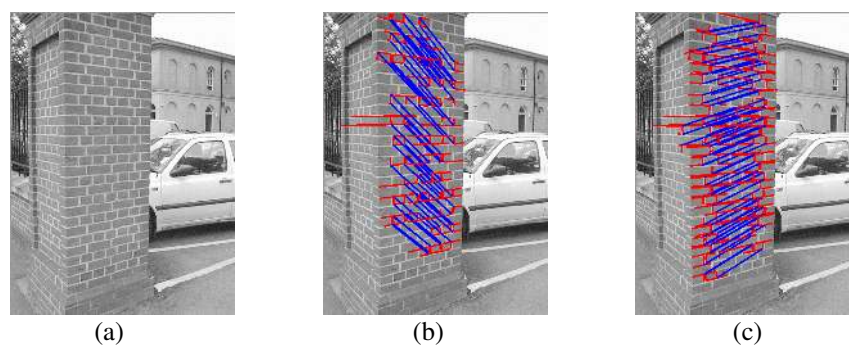


Figure 5: Conjugate translation grouping of line pair intersections. (a) Original image. (b) One grouping after guided search. (c) Another grouping after guided search.

Again, this is accomplished by RANSAC (three points determine the model). The resulting one-dimensional grids can then be clustered according to their vanishing points to yield groupings of one-dimensional grids with the same vanishing point. Lastly, the full two-dimensional grid can be extracted by finding two consistent groupings of one-dimensional grids.

The final parts of the algorithm are then identical to conjugate translation grouping, namely ML estimation and guided search for new elements. An example is shown in figure 6.

Harvesting more elements Once a good estimate of transformation parameters is available, this can be used to guide the *search* for missing elements in the grid.

For each point missing from the grid, the closest (in the image) grid point is identified. An element is verified at the missing point if there is a cross-correlation above threshold with the mapped intensity regions from the closest point. Note, this procedure identifies elements which have been missed in the initial feature detection. There may well not be any features present, but because the transformation and intensity are tightly estimated

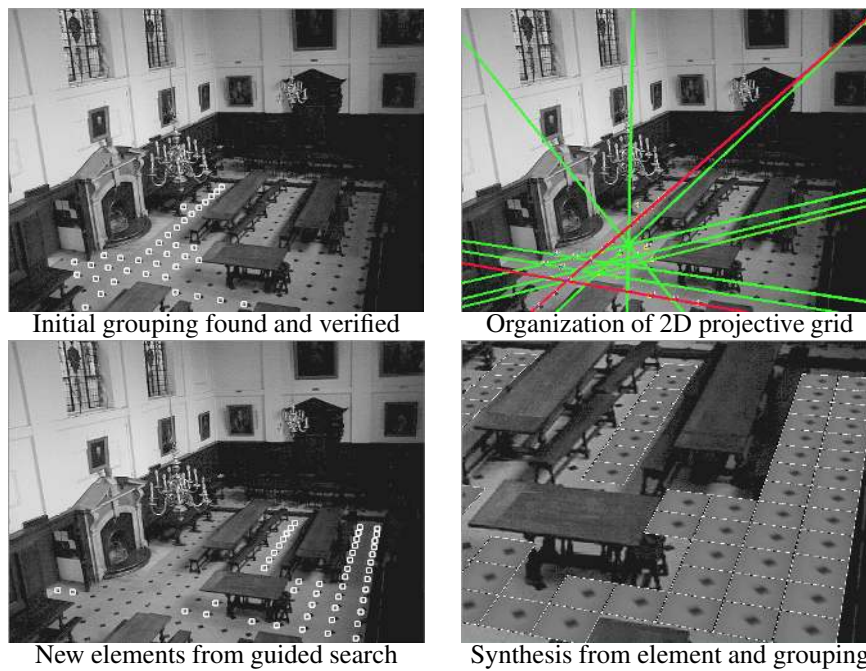


Figure 6: Finding a 2D grid, searching for new elements and estimating the tile. The second image shows the lines extracted from the point set given by the elements in the first image.

false positives are not generated.

Image descriptor Having learnt the element and grouping from the image we can now estimate the frontoparallel intensity on the tile. A simple averaging of the frontoparallel views obtained from the grid structure produced the results shown in the figure. This demonstrates that the element plus grouping does provide a succinct description for part of the image.

5 Conclusions and Extensions

We have shown that parametrized transformations arise from repeating operations in the scene, and explored grouping strategies for several classes of transformation including local affine, conjugate translation, and conjugate grid. These image transformations provide very strong grouping and search cues and enable very successful grouping — the grid grouper harvested all the non-occluded elements, and also computed a compact description of a substantial part of the image. Table 1 shows these and several other important parametrized transformations arising from various commonly occurring repeating operations in the scene. Similar grouping strategies can be applied to these other cases and this is current work.

There are a number of interesting research problems remaining:

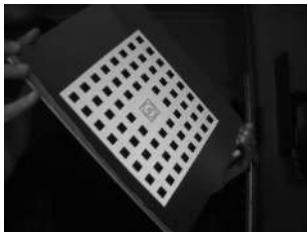
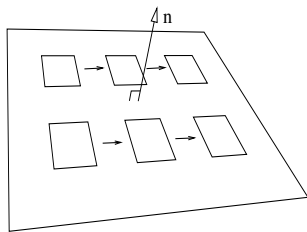

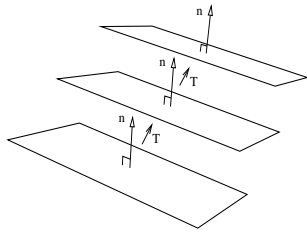

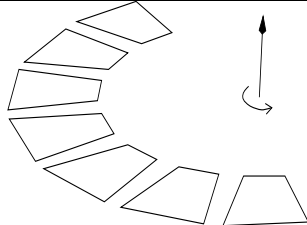
Transformation	Example image	Schematic
Conjugate Translation $\mathbf{I} + \mathbf{v}\mathbf{l}_\infty^\top$, where $\mathbf{l}_\infty \cdot \mathbf{v} = 0$. 4 degrees of freedom		
Family of Planar Homologies $(\alpha + k\beta)\mathbf{I} + \mathbf{v}\mathbf{l}_\infty^\top$, for $k \in \mathbb{Z}$ 5 degrees of freedom		
Conjugate Rotation $\mathbf{H}^n = \mathbf{I}$ for an n -fold symmetry		

Table 1: A menagerie of transformations.

1. There are many other classes of exact repeating operation (e.g. a similarity transformation on a plane) and the induced image transformations from these are yet to be derived.
2. The parametrized transformations model *exact* repeating operations and elements. However, in many natural scenes there is statistical variation about an exact repetition (for example leaves on a tree). Statistical variation can be incorporated at two points: first, the repeated element can be drawn from a distribution on geometry and/or intensity; second, the repeating transformation can be drawn from a distribution.
3. Often groupings can be organised into *meta-groupings*. For example the windows in figure 2 may be organized as four meta-groupings, each consisting of nine grouped elements. Finding such meta-groupings is an AI type problem.
4. At present no *a priori* information is included. For example, there are often sensible limits on the parameters of the conjugate translation. One means for incorporating this information is to sample the parameters from suitable prior distributions.

Acknowledgements

This work was supported by an EPSRC grant and EU ACTS Project Vanguard. The image in figure 1 was provided by T. Leung. We are grateful for help from Dr Andrew Fitzgibbon in supporting the TargetJr software libraries used for this work.

References

- [1] T. O. Binford. Inferring surfaces from images. *Artificial Intelligence*, 17:205–244, 1981.
- [2] J. Canny. A computational approach to edge detection. *IEEE T-PAMI*, 8(6):679–698, 1986.
- [3] I. J. Cox, J. M. Rehg, and S. Hingorani. A bayesian multiple hypothesis approach to contour grouping. In *Proc. ECCV*, LNCS 588, pages 72–77. Springer-Verlag, 1992.
- [4] O. Faugeras. *Three-Dimensional Computer Vision: a Geometric Viewpoint*. MIT Press, 1993.
- [5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.
- [6] C. J. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conf.*, pages 147–151, 1988.
- [7] T. Leung and J. Malik. Detecting, localizing and grouping repeated scene elements from an image. In *Proc. ECCV*, LNCS 1064, pages 546–555. Springer-Verlag, 1996.
- [8] T. Leung and J. Malik. Contour continuity and region based image segmentation. In *Proc. ECCV*, LNCS 1406, pages 544–559. Springer-Verlag, 1998.
- [9] J. Liu, J. Mundy, and A. Zisserman. Grouping and structure recovery for images of objects with finite rotational symmetry. In *Proc. Asian Conf. on Computer Vision*, volume I, pages 379–382, 1995.
- [10] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.
- [11] J. Ponce, D. Chelberg, and W. B. Mann. Invariant properties of straight homogeneous generalized cylinders and their contours. *IEEE T-PAMI*, 11(9):951–966, 1989.
- [12] A. Sarkar and K.L. Boyer. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE T-PAMI*, 15(3):256–274, 1993.
- [13] J. Shi and J. Malik. Self inducing relational distance and its application to image segmentation. In *Proc. ECCV*, LNCS 1406, pages 528–543. Springer-Verlag, 1998.
- [14] L. Van Gool, T. Moons, and M. Proesmans. Groups, fixed sets, symmetries and invariants. Technical Report KUL/ESAT/MI2/9426, Katholieke Universiteit Leuven, ESAT/MI2.
- [15] L. Van Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *Proc. ECCV*, pages 642–651. Springer-Verlag, 1995.
- [16] M. Zerroug and R. Nevatia. From an intensity image to 3-d segmented descriptions. In J. Ponce, A. Zisserman, and M. Hebert, editors, *Object Representation in Computer Vision*, LNCS 1144, pages 11–24. Springer-Verlag, 1996.
- [17] A. Zisserman, J. Mundy, D. Forsyth, J. Liu, N. Pillow, C. Rothwell, and S. Utcke. Class-based grouping in perspective images. In *Proc. ICCV*, 1995.