3D Object Recognition using Invariance

by

Andrew Zisserman, David Forsyth, Joe Mundy,
Charlie Rothwell, Jane Liu and Nic Pillow

Robotics Research Group

Report No. OUEL 2027/94

November 1994

University of Oxford
Department of Engineering Science
19 Parks Road
Oxford OX1 3PJ
U.K.

Tel: +44 865 273926
Fax: +44 865 273908
Email: az@robots.oxford.ac.uk

Contents

1 Introduction **1**
- 1.1 Related approaches to object recognition . . . 2
- 1.2 Geometric invariants in modelling and recognition . . . 3
  - 1.2.1 Indexing . . . 4
  - 1.2.2 Invariance and representation . . . 4
  - 1.2.3 Model acquisition from images . . . 5

2 The planar recognition system **5**
- 2.1 Projective invariants used . . . 6
- 2.2 Architecture . . . 6
  - 2.2.1 Feature extraction and invariant formation . . . 7
  - 2.2.2 Indexing to generate recognition hypotheses . . . 8
  - 2.2.3 Hypothesis merging . . . 8
  - 2.2.4 Verification . . . 8
- 2.3 Model acquisition and library formation . . . 9
- 2.4 Recognition examples . . . 9
- 2.5 Summary of performance . . . 10
- 2.6 Appraisal . . . 12

3 Extending invariant descriptions to 3D structures **12**
- 3.1 Object classes . . . 12
- 3.2 Definitions — 3D projective invariants . . . 13
- 3.3 Constrained Point Sets . . . 14
- 3.4 Repeated Structure . . . 15
  - 3.4.1 Bilateral Symmetry . . . 16
  - 3.4.2 Translational Repetition . . . 19
  - 3.4.3 Other Repeated Structures . . . 20
- 3.5 Rotational Symmetry . . . 21
- 3.6 Canal Surfaces . . . 23
- 3.7 Polyhedra . . . 26
- 3.8 Extruded Surfaces . . . 27
- 3.9 Algebraic Surfaces . . . 28

4 An architecture for a 3D recognition system **28**
- 4.1 Fundamental Principles . . . 28
  - 4.1.1 Class . . . 29
  - 4.1.2 Consistency . . . 29
- 4.2 The Architecture . . . 30
- 4.3 Model acquisition . . . 33
- 4.4 MORSE . . . 33

5 Discussion **35**
- 5.1 Critique of the invariance approach . . . 35
- 5.2 Avenues of future research . . . 36

Abstract

The systems and concepts described in this paper document the evolution of the geometric invariance approach to object recognition over the last five years. Invariance overcomes one of the fundamental difficulties in recognising objects from images: that the appearance of an object depends on viewpoint. This problem is entirely avoided if the geometric description is unaffected by the imaging transformation. Such invariant descriptions can be measured from images without any prior knowledge of the position, orientation and calibration of the camera. These invariant measurements can be used to index a library of object models for recognition and provide a principled basis for the other stages of the recognition process such as feature grouping and hypothesis verification. Object models can be acquired directly from images, allowing efficient construction of model libraries without manual intervention.

A significant part of the paper is a summary of recent results on the construction of invariants for 3D objects from a single perspective view. A proposed recognition architecture is described which enables the integration of multiple general object classes and provides a means for enforcing global scene consistency.

Various criticisms of the invariant approach are articulated and addressed.

1 Introduction

The computer recognition of objects has attracted considerable research effort over the last 25 years. It is now widely accepted that object recognition, in the setting of real world scenes and based on a single perspective view, is a difficult problem and cannot be achieved without the use of object models to guide the processing of image data and to confirm object hypotheses. It is also accepted that the most reliable information which is available in a scene is derived from a geometric description of the object based on its projection in the form of 2D geometric image features, as opposed to, for example, its intensity shading. Thus, object recognition systems draw on a library of geometric models, which usually contain information about the shape and appearance of a set of known objects, to determine which, if any, of those objects appear in a given image or image sequence. Recognition is considered successful if the geometric configuration in an image can be explained as a perspective projection of a geometric model of the object.

At present, 3D recognition systems generally have small modelbases containing relatively simple objects. Progress is needed on three fronts:

- **Larger modelbases**: Systems should be able to deal with modelbases containing hundreds to thousands of models. The methods of pose consistency (reviewed in section 1.1), which are commonly used for modelbases with only a few objects, are infeasible for large modelbases because of the computational expense. Coping with such sizes clearly requires some partitioning of the modelbase.

- **More general shape models**: Typically polyhedra are used, which are a poor model for curved objects. A direct representation for non-trivial curved objects is required.

- **Automatic segmentation and grouping**: This is the process, also called figure-ground separation, of extracting image feature groups which correspond to individual object outlines without including the background and other occluding objects. The lack of such grouping is a significant barrier to successful recognition in current systems. In addition to representing the shape of 3D objects, models will have to provide *mechanisms* for their feature segmentation and grouping.

This paper establishes a framework for the next generation of 3D model based vision recognition systems which will have large modelbases, with objects partitioned into a number of different 3D object *classes*. Recognition is from single perspective images of scenes, where the camera is uncalibrated, the objects could be partially occluded, and the scene might contain objects not in the model library. The object classes are defined geometrically in terms of symmetry or other 3D geometric constraints. The *constraint* enables *invariants* of a 3D object in the class to be extracted from a single image of the object outline; and also generates invariant relations on the image outline that enable *grouping*.

Although the paper concentrates on perspective images, the methods are, of course, applicable in weak-perspective (or "affine") imaging situations. Weak-perspective, a linear approximation to perspective, is appropriate as a camera model when object relief is small compared to distance from the camera. A consequence is that parallel world lines are imaged as parallel lines. Invariants computed for perspective imaging are also valid for weak-perspective.

A major constraint underlying the work presented here is that recognition is based on one uncalibrated view of a scene. Our motivation is that this restriction applies in many of the current and future applications for object recognition, such as aerial surveillance, image database query processing, and image-hypertext editing. Even if more images are available, for example in the case of video processing, camera calibration will not generally be known initially. Any grouping, recognition hypothesis, or object recovered up to some ambiguity from a single image, can be propagated to advantage to subsequent views.

A central question explored in this paper is the nature of the shape representation necessary for recognition. Euclidean (metric) representations are routinely used in many existing recognition systems. However, under the most general imaging conditions, structure is recovered up to a projective transformation (i.e., a more general transformation than Euclidean). We demonstrate that projective representations are adequate for recognition. A stratification of representations is provided by the hierarchy of transformation groups: projective, affine, similarity (scaled Euclidean), and Euclidean. This representation hierarchy is progressively more restrictive; for example, two objects that are projectively equivalent need not be affine or similarity equivalent. We will be primarily concerned with the projective stratum, since this covers the "worst case" ambiguity. The other strata will be used to advantage at particular stages of the recognition process.

A related area is the use of *quasi-invariants* [4]. A quasi-invariant is an object property or relation that is not invariant to projective transformations, but is stable over a useful range of views. Invariants of other transformation groups in the hierarchy given above are sometimes quasi-invariants [5]. Quasi-invariants can be very effective in grouping and partial indexing even though they vary under perspective projection. Examples of quasi-invariants are given in the paper.

Our geometric notion of class differs from the more usual functional one. For example, in our definitions, a vase is considered as a surface of revolution as opposed to a container for flowers and water. A geometric class is not specific to a particular object but instead describes a family of objects which are unified by their common 3D constraint relations. A number of examples of these 3D object classes are given in section 3.

We have defined a recognition architecture which integrates these ideas. Class influences each level of the architecture, from image grouping through to organisation of the model base and 3D scene constraints. Recognition is class based, proceeding first by a classification based on image curves, and subsequently the identification of a particular model within the class using values of geometric attributes. This contrasts with many existing recognition systems where a particular object is directly identified. The architecture, combined with the success of existing implementations, demonstrates that a large-scale system implementation based on an invariant framework is now warranted. This effort will culminate in an object recognition system that can recognise a broad class of 3D structures with thousands of individual object instances in the model library.

1.1 Related approaches to object recognition

Recognition is the establishment of a correspondence between image and model features. Most recent approaches to recognition have been implemented in three stages (similar to those defined in [28]): grouping, indexing, and verification.

The aim of grouping (also called *perceptual organisation* [37], *selection*, or *figure-ground discrimination*) is to provide an association of features that are likely to have come from a single object in a scene. Features are typically grouped together using cues such as proximity, parallelism [3, 37] collinearity, and approximate continuity in curvature [12, 62]. The indexing stage hypothesises an association between the grouped image features, and features on a model in the library. The final stage, verification, determines the consistency of this hypothesis with the image data. The image-model match is used to project the model onto the image, and to test the validity of the model hypothesis and model-to-image feature

correspondences determined by measuring image support.

There are three distinct categories of algorithm that have been used to compute correspondence:

1. **Interpretation trees** frame the model-to-image correspondence task as a search tree to allow all possible model and image feature associations, and then control and prune this search process. Although inefficient, this has proved reliable for planar object recognition for a small modelbase when single images are used [27], and has been extended by Ettinger to include useful notions about how hierarchical object descriptions can be realised [14]. However, interpretation trees are not generally able to work with single images of three dimensional objects (though effective when 3D data is provided as direct input to the system [2, 27, 48, 49, 52]). Interpretation trees are not restricted to rigid objects; the *sup-inf* framework for geometric reasoning used in ACRONYM [8] allows the interpretation tree to account for tolerance interval constraints on parameterised objects. Brooks' work has been extended by both Fisher [19] and Reid [54] for different types of sensor and constraint framework. Other ways to treat parameterisations have been suggested by Grimson [28].

2. **Hypothesise and test**, also called *alignment*, first aligns a model-to-image features [31] to yield an initial estimate of pose. This hypothesised alignment is tested by searching for other model-to-image correspondence predicted by the model pose (verification). This algorithm has been implemented for a variety of data formats and feature types [1, 6, 15, 24, 38, 69]. In fact, extensions to 3D curved surfaces have even been created [13, 33].

3. **Pose clustering** is implemented by computing the object pose from a group of features corresponding to a particular model, and storing the estimate in an accumulator in pose space; if enough local groups have the same pose, a hypothesis for the model is formed. This approach (frequently called *generalised Hough*) has the disadvantage that the pose space is high dimensional (six *dof* for 3D Euclidean space), so searching for consistent pose is expensive. Two ways round this are to use a decomposition of the pose space into separable parameters [44, 68], or to use an adaptive Hough transform [66]. Another approach eliminates the requirement to quantise the pose space into rectangular cells by constructing a quantisation that depends both on the estimates of pose, and on the expected error bounds of the pose measurements [10].

For a small number of models, for example two or three, it is reasonable simply to try to find image feature support for each model. This approach is typical of many existing systems [1, 2, 28, 31, 38, 48, 52]. As the size of the model library increases, this approach becomes computationally too expensive. It is then more effective to choose potential models from the library based on the observed image features. That is, image feature measurements are used to *index* into the model base. In constructing such *index functions*, invariance plays a major role, since a model should be identified irrespective of object pose.

1.2 Geometric invariants in modelling and recognition

Invariants are properties of geometric configurations which remain unchanged under an appropriate class of transformations. Within the context of vision we are interested in determining the invariants of an object under perspective projection onto an image. For example, for a planar object the perspective projection between object and image planes is a *projective* transformation. Properties such as intersection, collinearity, and tangency are unaffected by a projective transformation; however, invariant *values* can also be computed. Examples are given in section 2.1.

More formally, under a linear transformation of coordinates, $\mathbf{X}' = \mathtt{T}\mathbf{X}$, the invariant, $I(\mathbf{P})$, of a configuration $\mathbf{P}$ transforms as

$$I(\mathbf{P}') = |\mathtt{T}|^w I(\mathbf{P})$$

and is called a *relative* invariant of weight w, where $\mathbf{P}'$ is the transformed configuration. If $w = 0$, the invariant is unchanged under transformations and is called a *scalar* invariant. We will only be interested in scalar invariants in this paper.

In general we seek invariance to *projective* transformations, so $\mathtt{T}$ is a general non-singular square matrix acting on homogeneous coordinates. For planar configurations it is 3×3, and for 3D configurations 4×4. Note that invariants are computed with respect to a *transformation*, which is a mapping between

spaces of the same dimension. The goal is to measure the invariants from a perspective *projection* of the configuration, where the image may have a lower dimension than the object. We write $\mathtt{P}$ for the projection matrix that covers a 3D Euclidean transformation of the object followed by perspective projection onto the image. For *planar* objects the original and image spaces are the same dimension and $\mathtt{P}$ is simply a projective transformation represented by a 3×3 matrix. This is discussed in detail in section 2. For *three-dimensional* objects, the original and image spaces are no longer of the same dimension and $\mathtt{P}$ is a 3×4 matrix mapping 3D homogeneous coordinates onto the image plane. This is described in detail in section 3.2.

1.2.1 Indexing

One of the most important uses of invariants in vision is as indexing functions. In traditional model-based recognition systems (section 1.1), recognition proceeds by hypothesising a correspondence between image and object features, and then evaluating the hypothesis based on the consistency of the best projection of the model onto the image features. This constitutes simultaneously finding pose and performing recognition, and is generally of a complexity *linear* in the number of models in the library, since each model must be evaluated.

An index function provides direct access to a certain model in the model base without using specific information about the model, or model pose in advance. Ideally, the index function should uniquely retrieve a model from the library (thus facilitating *constant* time, as opposed to linear, access to the library), but in practice it is likely that a small number of models are retrieved with the same index. Even so, the search cost is considerably reduced below that of testing the full library. The index is typically a vector of independent invariant measurements.

More formally: the index is considered to be a vector, $\mathbf{M}$, which selects a particular model from the library. The index is a function $\mathbf{M}(\mathbf{f})$ of a set of *projected* object features only, where $\mathbf{f} = \mathtt{P}\mathbf{F}$, with $\mathbf{F}$ object features, and $\mathbf{f}$ the corresponding image features. Assuming that $\mathbf{M}$ can be computed from any image projection of the object features, then library values for $\mathbf{M}$ can be constructed simply by acquiring one or a few images of the object in isolation.

For planar objects, $\mathtt{P}$ is a planar projective transformation, $\mathtt{T}$, from the object in an arbitrary pose onto the image plane, and

$$\mathbf{M}(\mathtt{T}(\mathbf{F})) = \mathbf{M}(\mathbf{F})$$

i.e., the index has the same value computed on the original object and after the transformation (a scalar invariant). Each element of the index vector $\mathbf{M}$ is an invariant measure computed from a group of image features such as conics, lines, points and plane curve segments. A typical example is shown in figure 1. For 3D objects the same function cannot be applied to object and image, since they differ in dimension. However, again $\mathbf{M}$ is defined so that each element is a projective invariant of the 3D structure that is measured from the perspective image. Examples are given in section 3.2.

1.2.2 Invariance and representation

The term "invariance" does not simply refer to the viewpoint-invariant measurement vector described above. The term also includes the idea of an invariant *relation*, which is distinct from an invariant *value*. For example, the cross-ratio is an invariant value of four collinear points. The collinearity of the points is a projectively invariant relation between the points which is independent of the cross-ratio value. In the definition of generic geometric classes, the identification of invariant relations is often a more important issue for representation than the computation of specific invariant indexing values.

Another general aspect of the invariant approach is the symbiotic application of geometric and algebraic analysis. It is often the case that geometric insights provide the first clue to the nature of invariants for a particular object class. Then subsequent algebraic analysis can generalise and simplify invariant computation, and in turn provide additional insight.

1.2.3 Model acquisition from images

A model consists of the set of significant geometric features of the object boundary known up to a projective, or more restrictive, transformation (for example affine). Projective models can be constructed from images without requiring knowledge of the intrinsic camera parameters or known 3D ground control points. In the case of 2D objects the model can be acquired from a single image, for 3D objects more images are generally required. Model acquisition is discussed further in sections 2.3 and 4.3.

Before proceeding to the case of more general 3D object recognition, we review a mature system for 2D object recognition. This review will illustrate many of the issues in object recognition by invariants and provide a context for our more general discussion of recognition architectures at the end of the paper.

Notation

We adopt the notation that corresponding entities in two different coordinate frames are distinguished by upper and lower case. In general lower case is used for image quantities, and upper for 3D quantities. Vectors are written in bold font, e.g., $\mathbf{x}$ and $\mathbf{X}$. Matrices are written in typewriter font, e.g., $\mathtt{c}$ and $\mathtt{C}$. With homogeneous quantities, equality is up to a non-zero scale factor.

For smooth surfaces the *profile* (also called the *apparent contour*) is the outline of the surface in the image. It is the image projection of a surface curve, the *contour generator*, where rays from the optical centre are contained in the surface tangent plane.

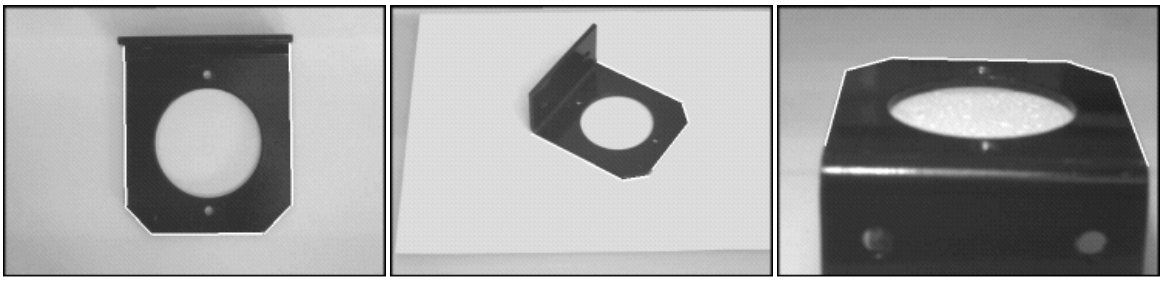
2 The planar recognition system

The use of *planar projective* invariants for planar object recognition is particularly appropriate and straightforward because a projective transformation between object and image planes covers all the major imaging transformations: the plane to plane projectivity models the composed effects of 3D rigid rotation and translation of the world plane (camera extrinsic parameters), perspective projection to the image plane, and an affine transformation of the final image which covers the effects of camera intrinsic parameters. Consequently, projective invariants, which are unaffected with respect to all of these parameters, have a high currency for this domain [40, 50, 55, 57, 58, 60, 70, 71].

Here we summarise the main features of a planar object recognition system that has been developed during the past four years. The projective representation of shape used in the system has the key advantages of simple model acquisition (direct from images), no need for camera calibration or object pose computation, and the use of index functions. Recognition proceeds by measuring invariants in the target image. The invariants are used to construct index vectors to select models from the library. If the index value coincides with that associated with a model, a recognition hypothesis is generated. Recognition hypotheses corresponding to the same object are merged to form *joint hypotheses*, provided they are geometrically compatible. The (joint) hypotheses are then verified. The system[1] has been tested on a large set of images and under varying levels of occlusion and clutter. A detailed description of this system appears in [60].

The projective nature of the representation is utilised at a number of stages in the recognition process, for example in both model acquisition and verification. In acquisition, any image provides a projective model of the object outline because the image and object planes are related by a projective transformation. This is because the object is mapped by a perspective transformation onto the image, and perspective is a restricted form of a projective transformation. In verification, the target image outline is projectively related to the model image outline. This follows because the target image outline is a projective transformation of the object outline, which is a projective transformation of the model image outline. Plane projective transformations are a group, and a sequence of projective transformations is equivalent to a single projective transformation (group closure).

[1] The system is called LEWIS. The motivation for this name is explained in section 4.4.



a **b** **c**

Figure 1: The lines used to compute the five line planar projective invariant for the above images are highlighted in white. The values are given in table 1.

Five line invariants		
Measured on	I_1	I_2
Object	0.840	1.236
Figure 1(a)	0.842	1.234
Figure 1(b)	0.840	1.232
Figure 1(c)	0.843	1.234

Table 1: Values of plane projective invariants measured on the object, and from images with varying perspective effects. The values vary (due to measurement noise) by less than 0.4%.

2.1 Projective invariants used

There are three different algebraic invariant constructions used in the system: five lines; a conic and two lines; and a conic pair. For example the two invariants of five lines are given by

$$I_1 = \frac{|\mathbf{N}_{431}||\mathbf{N}_{521}|}{|\mathbf{N}_{421}||\mathbf{N}_{531}|} \qquad \text{and} \qquad I_2 = \frac{|\mathbf{N}_{421}||\mathbf{N}_{532}|}{|\mathbf{N}_{432}||\mathbf{N}_{521}|}, \tag{1}$$

where $\mathbf{N}_{ijk} = (\mathbf{l}_i, \mathbf{l}_j, \mathbf{l}_k)$, $|\mathbf{N}_{ijk}|$ is the determinant, and $\mathbf{l} = (l_1, l_2, l_3)$ is the homogeneous representation of a line: $l_1 x + l_2 y + l_3 = 0$. (See [45] for the other invariants.) Table 1 gives examples of these invariants computed from the images shown in figure 1, which have varying degrees of perspective distortion. These are applicable to image curves that are "algebraic" (lines, conics). For non-convex smooth curve segments canonical frame invariants [45, 57] are used. These are constructed from projective coordinates of a concavity delineated by a bitangent.

In all cases there is tolerance to partial occlusion, i.e., the invariants can still be formed if part of the outline is occluded. This is a result of using *semi-local* invariant descriptions, i.e., not global, like moments of the entire boundary, and *redundancy*: there are a number of different descriptors for each object so that there is not an excessive requirement for any single object region to be visible. In the algebraic case lines and conics can still be extracted if part of the curve is occluded.

2.2 Architecture

The stages of recognition are shown in figure 2. In the following sections we describe these stages in sufficient detail to expose the important issues for consideration in extending these ideas to 3D object recognition.

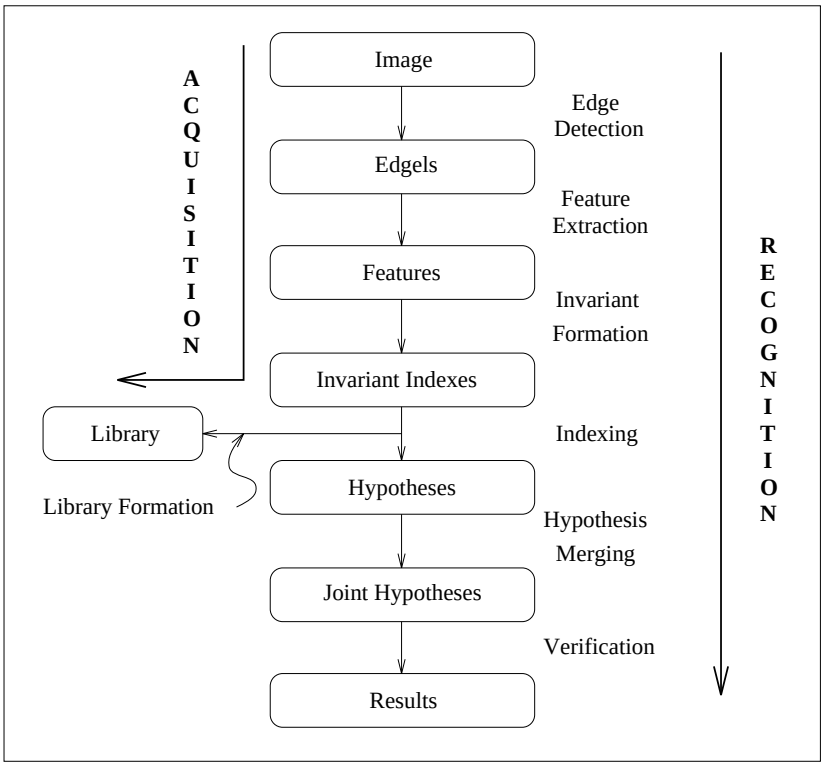


Figure 2: The recognition system has a single grey scale image as input and the outputs are verified hypotheses with associated confidence values. Many of the processes are shared by the acquisition and the recognition paths. The recognition system is similar to previous systems [28] in all but the indexing and hypothesis merging stages.

2.2.1 Feature extraction and invariant formation

The goal of the segmentation is the extraction of geometric primitives suitable for constructing invariants. In the algebraic case this involves straight lines and conics, and for non-algebraic curves, concavities delineated by bitangents. An example of algebraic segmentation is shown in figure 4.

A local implementation of Canny's edge detector is used to find edgels to sub-pixel accuracy. These edgels are linked into chains, extrapolating over any small gaps. Considerable advantage is made of local image feature topology. In many recognition systems, the local connectivity of edgel chains and fitted features is ignored; but we have found that feature grouping, based on the connectivity provided by edgel chains and proximity, allows index formation to have a low complexity with respect to the number of image features.

For algebraic invariants, connectivity enables efficient linking and ordering of line segments. For example, five line invariants are formed from sets of consecutive lines within single edgel chains at a cost that is linear in the number of lines in the scene (i.e., $O(l)$, compared to $O(l^5)$ if all groupings are attempted). For concavities, the curve again provides an ordering for the feature points used (bitangent and cast tangent points [73]) and only the two cases of global curve reversal have to be considered.

Once sets of grouped features, $\mathbf{f}$, have been produced, the algebraic and canonical invariants are computed. Each set of grouped features, or concavity curve, generally produces a number of invariant values which are collected into a vector $\mathbf{M}(\mathbf{f})$. The invariant vector formed by the above process represents a point in the multidimensional invariant space. The space is quantised to enable hashing. Each object feature group is represented by a collection of points that define a region in the invariant space, the size of which depends upon the measured variance in the invariant value[2].

[2] See section 2.3.

2.2.2 Indexing to generate recognition hypotheses

The invariant values computed from the target image are used to index against invariant values in the library. If the value is in the library a preliminary recognition hypothesis is generated for the corresponding object. Each type of invariant (e.g., five lines, conic pair) separately generate hypotheses.

This process is made more efficient using a hash table that allows simultaneous indexing on all elements of the measurement vector. In the experiments to date there has not been any significant problem with collisions in the hash table. Hash table collisions[3] should not be confused with the intersection of object invariant measurements in index space. These intersections lead to erroneous hypotheses which cost some effort during the verification stage, but are usually eliminated.

2.2.3 Hypothesis merging

Many collections of primitives may come from the same model instance: for example, an object consisting of a square plate with a circular hole in it admits four collections, each consisting of a conic and two connected lines. Each collection has an invariant which may generate a recognition hypothesis. Such a set of recognition hypotheses is *compatible* if a single model instance could explain all of them simultaneously. Prior to verification, compatible hypotheses are combined into *joint hypotheses*. There are number of reasons why hypothesis merging is desirable:

1. Backprojection and searching for image support is computationally expensive and it is more efficient to validate several hypotheses of the same object together.

2. More features facilitates more accurate least squares calculation of the back projection transformation (there are more matched model and image features), and consequently a reduced error in measuring image support.

3. Many hypotheses indexing the same object in a single part of the scene significantly increase confidence that the match is correct.

The hypothesis merging process is equivalent to forming an interpretation tree for the indexed object based on the features which index a particular model. The merging is controlled by topological and geometric compatibility. The topological consistency (ordering and connectedness) is illustrated in figure 3a. Geometric consistency is implemented efficiently by a second use of invariants — this time joint invariants between the feature groups used to compute each individual hypothesis. This is illustrated in figure 3b.

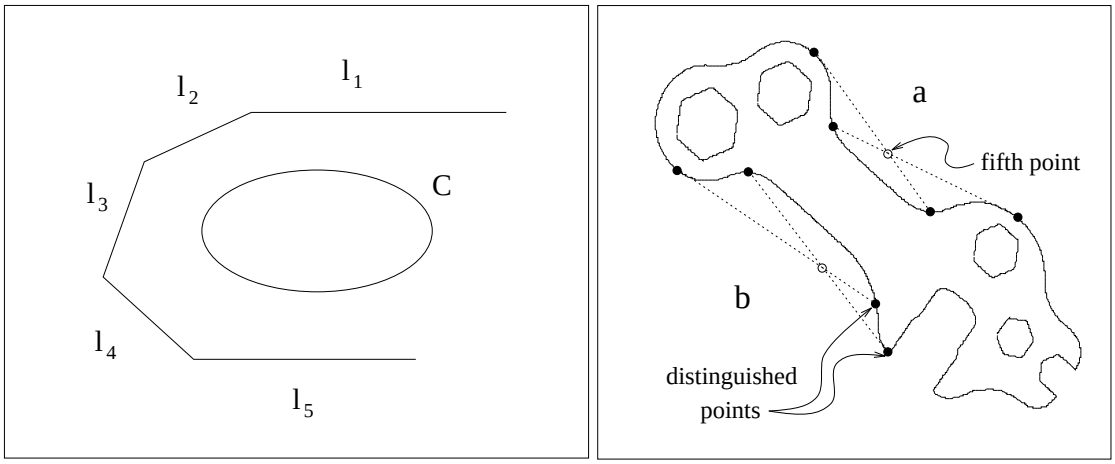
2.2.4 Verification

There are two steps involved in verification, both of which can reject a (joint) recognition hypothesis. The first is to attempt to compute a common projective transformation between the model features and the putative corresponding features in the target image. The second is to use this transformation to project the entire model onto the target image, and then *measure* image support.

Incorrect hypotheses arise because grouped image features happen to have an invariant value that coincides (within the error bounds) with one in the library. The features used to produce the matching model and image invariants provide sufficient constraints to compute the projective transformation between the model and image. In general this will be over constrained — many more constraints than the eight unknowns of the projective transformation are available. Consequently, if a common transformation cannot be computed the features are not projectively equivalent and the hypothesis is rejected.

Backprojection and subsequent searching involves the entire model boundary, not just the features used to form the invariant. Projected model edgels must lies close to image edgels with similar orientation (within 5 pixels and 15^0). If more than a certain proportion of the projected model data is supported (the threshold used is 50%), there is sufficient support for the model, and the recognition hypothesis is confirmed. The final part of the process is expensive as $O(10^3)$ edgels need to be mapped onto the image. Efficiency is improved by approximating the distances using the 3-4 distance transform of Borgefors [7].

[3] A hash table collision occurs when a number of models have the same hash index. Such a collision can occur when the number of hash buckets is smaller than the model population or when the hashing function is not uniform and causes many models to hash to the same bucket.



a **b**

Figure 3: Hypothesis compatibility: (a) If the same model is indexed by a five-line invariant (due to lines $\mathbf{l}_i$, $i \in \{1, \ldots, 5\}$), and a conic three-line invariant that is *compatible* with it (due to C and $\mathbf{l}_i$, $i \in \{2, \ldots, 4\}$), then it is wise to verify both hypotheses together. The invariants are compatible if the ordering of the image lines are consistent with those on the model. (b) For a pair of concavity curves there are 8 distinguished points which could be used to form $2 \times 8 - 8 = 8$ different five point invariants. Rather than computing so many, which is unnecessary, invariants are computed between the four distinguished points of each concavity, and the 'central' point of the other. This yields four invariants, and does so using a symmetric construction. These invariants are sufficient to hypothesise compatibility.

2.3 Model acquisition and library formation

One benefit of using only projective representations, rather than Euclidean ones, is that a model can be acquired directly from an image. No special orientations or calibrations are required. Acquisition is simple and semi-automatic (for instance, curves do not have to be matched entirely by hand between images), using the same software for segmentation and invariant computation as used during recognition.

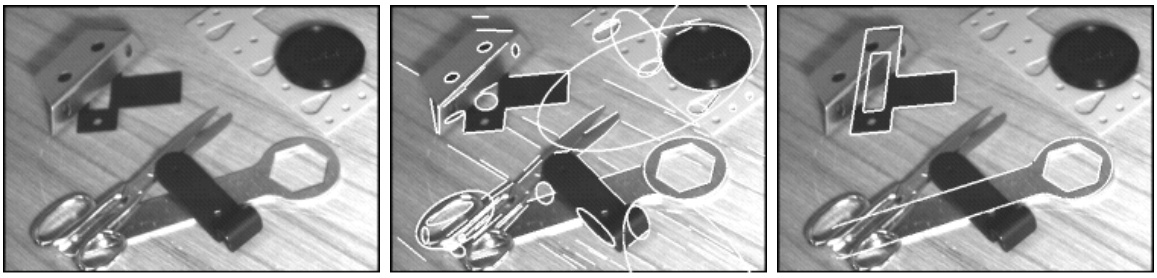
A model consists of the following: a name; a set of edges from an acquisition view of the object (used in the backprojection stage of verification); the lines, conics and concavities fitted to the edges; the expected invariant values and to which algebraic features and curve portions they correspond. (The mean and variance of the invariant values are computed from a variety of 'standard' viewpoints of the object.); and, finally, topological connectivity and geometric relations between feature groups used in the construction of joint invariants.

The library is partitioned into different sub-libraries, one for each type of invariant (e.g., one for the five-line invariant, another for the conic pair). Each sub-library then has a list of each of the invariant values tagged with an object name, and is structured as a hash table.

2.4 Recognition examples

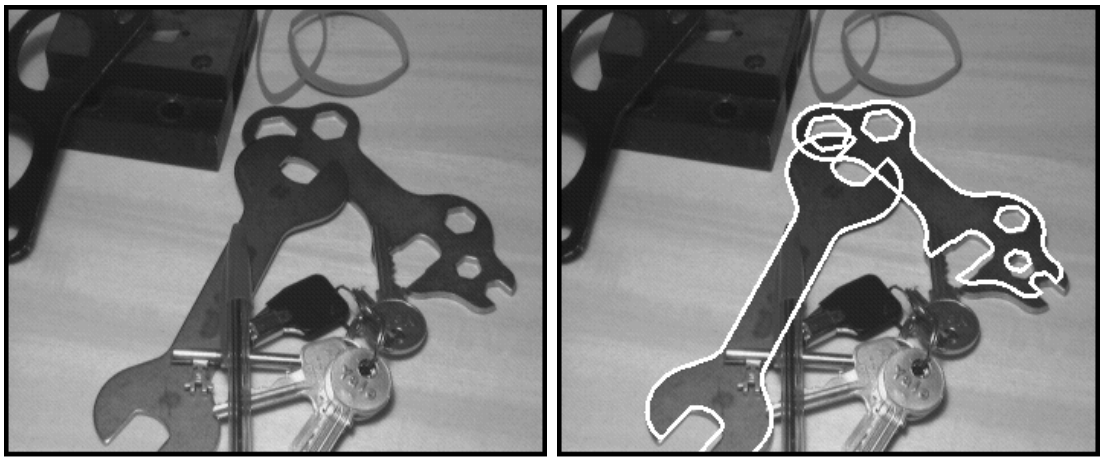
Only a small number of examples are included since others appear elsewhere [45, 58, 60]. In each case successful recognition is demonstrated by projecting the model outline onto the image. Segmentation for algebraic features is shown in figure 4. The two objects in the scene which are contained in the library are successfully recognised using algebraic invariants computed from these features despite substantial occlusion and clutter. 1049 invariants are computed which index 41 hypotheses. These are converted into 131 joint hypotheses[4] that have to be verified, of which 13 are rejected by first stage verification, based on valid projective transformations, and 78 require the second stage, based on image support. Figure 5 shows recognition based on canonical frame invariants. The algebraic and canonical frame invariants can be independently applied to an image to recognise objects of both types. Figure 6 shows an example of

[4] The joint hypothesis list consists of combinations of compatible hypotheses, together with all the original hypotheses.



a **b** **c**

Figure 4: (a) A scene containing two objects from the model base, with fitted lines (100 of them) and conics (27) superimposed in (b). These numbers are typical for images of this type. Note that many lines are caused by texture, and that some of the conics correspond to edge data over only a small section. The lines form 70 different line groups. (c) shows the two objects correctly recognised, the lock striker plate matched with a single invariant and 50.9% edge match, and the spanner with three invariants and 70.7% edge match.



a **b**

Figure 5: Single concavities are sufficient to recognise the two model instances shown in (b). The redundancy of the canonical frame representation gives much better tolerance to occlusion than global shape methods. The left hand object gained 67.1% boundary support, and the right object 81.6%.

recognition for both index methods together.

2.5 Summary of performance

Figure 7 shows data collected over fifty evaluations of the recognition system in which a single object from the model base was placed in a scene and partially occluded by other objects that are not in the model base (clutter). The average number of hypotheses computed as more models were added to the library is plotted. The first model added to the library always corresponded to the actual object in the scene. With 33 models in the library, on average 15.8% of the hypotheses were for the correct model. Although predominately linear, the graph has a very low gradient.

The real benefit of indexing becomes apparent when one considers how many hypotheses would be produced if an alignment technique is used, maintaining the same grouping methods. On average, over 2000 feature groups are produced for each image, and so 2000 hypotheses would be generated for each model feature group in the library (generally there are four or five feature groups per object and so the situation would be far worse). This would result in about 7×10^4 hypotheses for the entire model base compared to fewer than 60 produced when indexing is used. As these all have to be verified it is clear that indexing produces a dramatic improvement in the system efficiency.

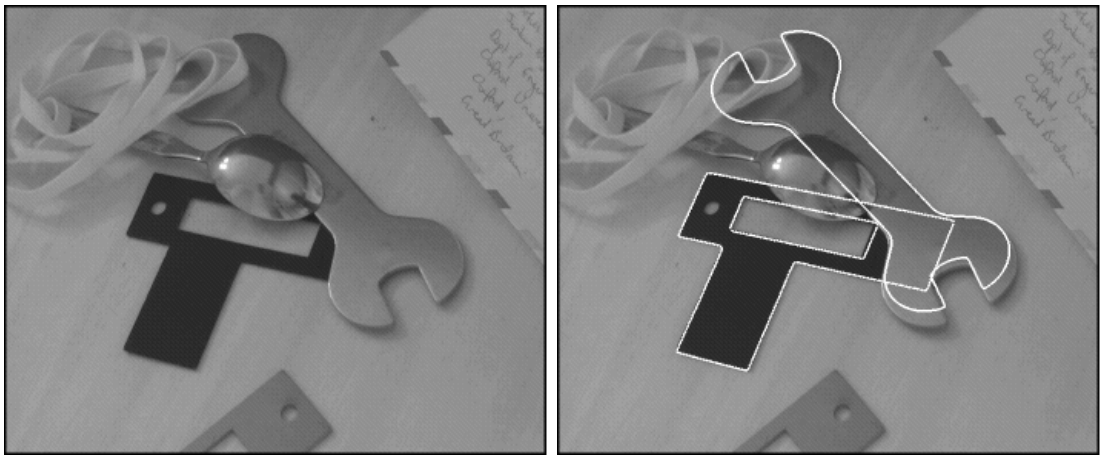


Figure 6: A demonstration that both types of invariant index can be used to recognise objects in a single image. The bracket is indexed using algebraic invariants and the spanner is indexed using the canonical frame signature.

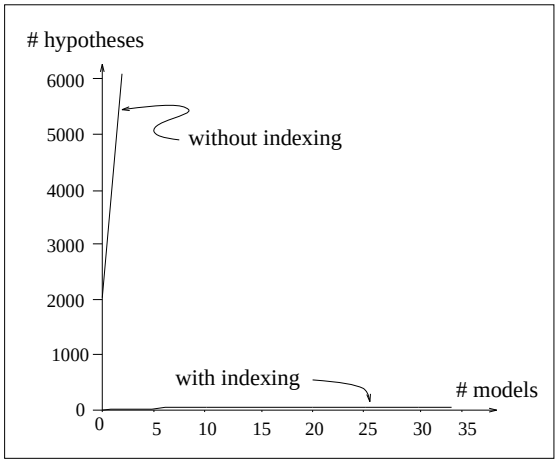


Figure 7: The number of hypotheses that have to be verified as the number of models in the library is varied. The results show an average over fifty scenes containing only one object in the library, but with other clutter and occlusion present. Over 2000 indexes are created for the scene, which corresponds to the number of hypotheses that would have to be verified **per model feature group** if alignment is used. Therefore, there is a rapid linear growth in the number of hypotheses created as the model base is expanded. However, the number of hypotheses created through indexing remains substantially lower — there is a linear growth, but with a very low constant of proportionality.

2.6 Appraisal

This system is an effective and reliable recognition system, and demonstrates a number of features that are likely to be important in building the next-generation system:

- **Hypothesis combination:** simply verifying each indexed model is prohibitive, particularly for complex objects with many features. Hypothesis combination is an effective way of combining semi-local information from different parts of the scene to obtain a single recognition hypothesis.
- **Untrustworthy and expensive verification:** verification is neither cheap nor reliable, as it involves back-projecting a large number of features, and testing for distance between those features and possibly unrelated image events. Verification scores can be incorrectly high, due to background clutter and texture which leads to false positives. The next generation system must have more extensive verification mechanisms using region properties as well as edge geometry. Also much more careful analysis of edge and junction intensity events must be carried out with respect to constraints imposed by the model. For example, specialised corner detection can be supervised by the model hypothesis.
- **A need for global scene analysis:** in many cases, ambiguities arise which must be settled globally by a scene analysis approach: for example, does a given image line come from object A or object B? Are the recognition hypotheses consistent? The lack of local support for a model hypothesis can be augmented by global relationships, e.g., A is on top of and partially occluding B. In this case we can predict the features which are potentially available to support hypotheses for B, once A is recognised.

Next, we take up the problem of 3D object recognition. First, the central question of the existence of invariants for the perspective projection of general 3D structures is discussed.

3 Extending invariant descriptions to 3D structures

Much recent debate has focused around a theorem, proven by a number of authors [9, 11, 42], which states that invariants can not be measured for a 3D set of points in general position from a single view. The theorem has frequently been misinterpreted to mean that *no* invariants can be formed for three dimensional objects from a single image. For the theorem to hold, however, the points must be completely unconstrained (like a cloud of gnats). If a 3D structure is *constrained*, then invariants are available. For example, six points constrained to lie on two planes in a "butterfly" configuration, as in figure 8, have a cross-ratio that can be measured in the image. This is a projective invariant of the entire 3D structure, and not simply a disguised planar invariant, since each plane contains only four points (five coplanar points are required to form a plane projective invariant from points alone).

In fact, a set of points in general position in space is a poor model of what we see: the world is full of curves, polyhedra, and surfaces; sets of isolated points are an irregular occurrence. An analogue of the above "no-invariants" theorem, in the case of surfaces, would be to ask whether a *generic* surface has invariants measurable in a single image from its profile. Other than qualitative descriptions such as non-convexity (from the sign of the profile curvature [32]) and the Euler characteristic (from the profile of a *transparent* surface) no projective invariant can be obtained. A similar result holds for space curves. However, if the surface satisfies constraints, much can be recovered from a single image, as the following section demonstrates.

3.1 Object classes

The form of the constraint on the object defines an object *class*. The class determines both the process by which the 3D invariants are measured in images, and the particular segmentation and grouping strategies that are applied during "early vision". For example, surfaces of revolution define a class, with a specific vase or wine glass being particular instances of the class. Projective invariants of the 3D surface can be recovered from the image profile, and further the two matching "sides" of the profile are projectively equivalent (section 3.5). That is, one side can be mapped onto the other by a projective transformation.

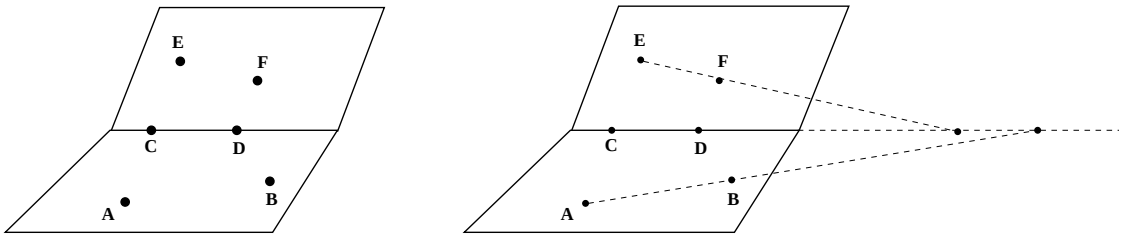


Figure 8: A "butterfly" configuration of six 3D points with a projective invariant measurable from a single perspective image. Points $ABCD$ and $CDEF$ lie on two planes intersecting in the line CD. The lines AB and EF intersect the line CD generating four collinear points. This construction can be carried out in 3D and the image to generate corresponding points. The cross-ratio of these points is the projective invariant. Note that the planes can articulate about the line CD without altering the value of the cross-ratio. Many analogous structures exist e.g., if the points AB are replaced by a line.

The segmentation and grouping for this class is guided by the association of these projectively related image contours.

It is important to distinguish this notion of geometric *class* from the idea of a generic *type*. For example, the class of rotationally symmetric objects is not the same as the generic type category of *wine glass*. There can be many different shapes of wine glass but the class of rotationally symmetric objects is still larger and does not capture the functional notion of a wine drinking container. A related discussion of class is given by Moses and Ullman [42] who contrast the notions of generic and specific classes with regard to recognition functions.

Another significant aspect of the class definition is its imposition of constraints which can be measured and verified in the image. This consideration is a significant departure from the hypothesise and test paradigm of conventional model-based recognition systems operating on specific objects. Here, the class assumption can be immediately confirmed without committing to the full chain of recognition processing. For example, for a rotationally symmetric object, the two "sides" of the image outline are related by a planar projective transformation (section 3.5). This relation can be immediately tested when a pair of image profile curves are hypothesised as belonging to an object of class *rotationally symmetric*.

In the following sections we catalogue a number of object classes where each is defined by an associated constraint. In each case the recovery of invariants is illustrated and other geometric consequences, such as invariant relations, described.

3.2 Definitions — 3D projective invariants

In what follows, we assume a perspective camera with unknown internal parameters, and measure only projective properties in the image. In turn, this means in general that only projective properties of the 3D objects can be recovered. Algebraically, the camera is modelled as $\mathbf{x} = \mathtt{P}\mathbf{X}$:

$$k \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2}$$

where $(x, y)^T$ are image coordinates, and $(X, Y, Z)^T$ world coordinates, and k is a scaling: in this case, $k = (p_{31}X + p_{32}Y + p_{33}Z + p_{34})^{-1}$.

We now introduce 3D projective invariants because they are the basis for image invariants that we can hope to recover from a single view of a constrained structure. These are invariants under projective transformations of $\mathcal{P}^3$. A projective transformation of $\mathcal{P}^3$ can be written as:

$$k \begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t_{21} & t_{22} & t_{23} & t_{24} \\ t_{31} & t_{32} & t_{33} & t_{34} \\ t_{41} & t_{42} & t_{43} & t_{44} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

where k is again the appropriate scaling to ensure the fourth coordinate is one. Fifteen parameters are required to define the 3D projective transformation matrix up to an arbitrary scale factor. Thus five 3D points are sufficient to construct a projective coordinate system. A sixth point will then have invariant 3D coordinates in the projective basis defined by the other five. These 3D point invariants can also be interpreted as the cross-ratio of tetrahedral volumes computed by taking determinants of point coordinates, four at a time.

For example, an invariant for six 3D points is given by

$$I_{6pts}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6) = \frac{|\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3\ \mathbf{X}_4|\,|\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_5\ \mathbf{X}_6|}{|\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3\ \mathbf{X}_5|\,|\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_4\ \mathbf{X}_6|}$$

where $\mathbf{X}_i = (X_i, Y_i, Z_i, 1)^t$. This invariant has the familiar property of invariants that

$$I_{6pts}(\mathbf{X}'_1, \mathbf{X}'_2, \mathbf{X}'_3, \mathbf{X}'_4, \mathbf{X}'_5, \mathbf{X}'_6) = I_{6pts}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5, \mathbf{X}_6),$$

i.e., both the value and the form of the expression are unaffected by the transformation.

By assuming that a set of constraints hold among the 3D projective invariants of a point set, it becomes possible to measure 3D projective invariants in a single view. The following section illustrates the nature of these constraints and provides a geometric interpretation for the measurable invariants.

3.3 Constrained Point Sets

It is possible in general to predict whether invariants of a three dimensional structure can be measured from images, by counting the number of image measurements available. While such counting arguments cannot cover every degeneracy, and therefore never offer a proof that an invariant is or is not possible, they offer a useful guide to what is likely to be true. A complication in counting the degrees of freedom of a geometric configuration, and the number of parameters of a transformation, is the existence of isotropies. An isotropy is an action of a transformation which does not alter the geometry of a configuration. For example, translation along the tangent direction of a line or rotation about the centre of a circle does not affect the structure. Therefore an isotropy reduces the effective number of transform parameters, and generally increases the number of invariants.

Consider a 3D configuration M. Plane projective invariants are denoted I_2, and projective invariants of 3D denoted I_3. Then,

> For m perspective images of M, if there is no isotropy group acting in $\mathcal{P}^3$, then to recover n_{I_3} functionally independent invariants of the three-dimensional structure M from image information alone, the following inequality must be satisfied:
>
> $$m \times n_{I_2} \geq n_{I_3} + 3m$$
>
> where n_{I_2} is the number of functionally independent plane projective invariants of the image of M. If there is an isotropy group of dimension $(\dim I_S)$ acting, then (provided $\dim I_S \leq 3$) the following inequality must be satisfied:
>
> $$m \times n_{I_2} \geq n_{I_3} + m(3 - \dim I_S)$$

We sketch the reasoning when there is no isotropy group acting for the case of a single image. The image projective invariants are functions only of the projective invariants of the configuration consisting of M *taken together* with the optical centre, $\mathbf{O}$. To see this, consider a projective transformation of $\mathcal{P}^3$. This projectively distorts the $\{M, \mathbf{O}\}$ configuration and the image plane. However, the image plane geometry is transformed by only a plane projective transformation. This means that the projective invariants of both the image configuration and the 3D configuration are unaffected. The image projective invariants can depend only on the rays linking $\mathbf{O}$ to points of M, and depend, therefore, on the optical centre $\mathbf{O}$ as well as on M. Since the image projective invariants are unaffected by the position of the image plane, the relationship between the 2D projective image invariants and the 3D projective object invariants is a function only of the three unknown coordinates of the centre of projection. Provided there are three or

	A	B	C
DOF	16	14	12
$\dim I_S$	0	2	4
n_{I_3}	1	1	1
n_{I_2}	4	2	1
dof of **O** that matter	3	1	0
Counting Relation	$4 = 1 + 3$	$2 = 1 + 1$	$1 = 1 + 0$

Table 2: Examples of the counting argument for various butterfly like structures. A = 6 point butterfly; B = butterfly with two points of wing replaced by line (4 points one line); C = butterfly with lines on both wings (2 points 2 lines).

more such image invariants, it is possible (in principle) to eliminate the (unknown) contribution of the optical centre.

The counting argument then simply relates the number of unknowns and the number of measurements: in m views there are $3m$ unknowns for the optical centres, and n_{I_3} unknown 3D projective invariants for the configuration M; the number of measurements is n_{I_2} in each of the m images. Note that, like most such arguments, the condition is necessary but may not be sufficient; this means that there could be cases where the counting argument indicates that invariants can be measured from the image, when in fact they cannot. The significance of the argument is that it indicates where a further analysis may be useful.

As an example, consider the case of six points in space which have three 3D projective invariants, as discussed above. If we specify, or assume, the values for two of the invariants, then we can compute the value of the third from a single image. The so-called butterfly configuration in figure 8 is an example where we assume that two of the 3D invariants are zero, which corresponds to the coplanarity of two sets of four points in the six point configuration. The counting argument goes as follows: the number of degrees of freedom for the image points is 12 (2 for each point) less 8 for the plane projective group gives $n_{I_2} = 4$. For 6 points in space on two planes there are 16 degrees of freedom (3 for each point, less two for the planarity constraints) less 15 for the 3D projective group, gives $n_{I_3} = 1$. There are also three unknown coordinates of the centre of projection. Thus the counting argument shows that the unknown 3D invariant can be measured in a single view, i.e.,

$$1 \times 4 \geq 1 + 1 \times 3,$$

Similar counts for a number of butterfly analogues in the case of isotropies are given in table 2. Sparr [63] has constructed many other examples of butterfly-like configurations, and provides a method for generating such invariants algebraically.

The counting argument is used in this manner to focus attention on configurations where invariants may be available. As a further example, consider recognising algebraic surfaces from their profiles. In this case, the surface has degree d, and has $[(1/6)(d+3)(d+2)(d+1)-1]-15$ functionally independent projective invariants. The profile has degree $d(d-1)$, and has $[(1/2)(1-d+d^2)(2-d+d^2)-1]-8$ functionally independent projective invariants. For $d > 2$, the number of invariants of the profile substantially exceeds the number of invariants of the surface, and so it is reasonable to expect to recover invariants from the profile of an algebraic surface. In fact, such invariants can be recovered, though the procedure is complicated; details are given in [22].

3.4 Repeated Structure

Structures that repeat in a single image of a scene are equivalent to multiple views of a single instance of the structure. Thus, for example, a view of two similar cars in a car park where the cars are parked within translations of one another, is equivalent to a stereo pair of images of one such car, with the cameras related by a pure translation. The 3D shape of the car can be recovered by the familiar techniques of stereopsis. More formally,

A repeated structure is defined by a geometric structure $\mathcal{S}$, and a 3D transformation $\mathcal{T}$, which generates a transformed *copy* of $\mathcal{S}$, i.e., $\mathcal{S}' = \mathcal{T}(\mathcal{S})$. Both $\mathcal{S}$ and $\mathcal{S}'$ are viewed in the same perspective image.

In many cases the internal calibration parameters of the camera will be unknown. In this case a single image of a repeated structure is mathematically identical to an uncalibrated stereo pair where the two cameras are related by the transformation between $\mathcal{S}$ and $\mathcal{S}'$. It has been shown by Faugeras [16] and Hartley *et al.* [29] that if one carries out stereo reconstruction from two uncalibrated perspective images, the reconstruction can differ from the actual 3D Euclidean geometry of the object by a 3D projective transformation. Thus, 3D projective invariants of this recovered structure have the same value as projective invariants measured on the actual Euclidean structure.

The equivalence with stereo means that epipolar structure can be defined within a single image and represents the geometric relationship between corresponding features on the object copies. As a simple example, consider the case where $\mathcal{T}$ is a 3D translation, i.e., $\mathcal{S}$ and $\mathcal{S}'$ are related by a simple 3D translation. In this case, it can be shown [41] that *affine*, rather than projective, 3D structure can be recovered. Lines joining corresponding points on $\mathcal{S}$ and $\mathcal{S}'$ are parallel in 3D and are imaged as a set of lines converging to a vanishing point. These imaged correspondence lines and vanishing point are the analogue of "epipolar lines" and "epipole", and these terms will be used from now on. For translation only, there is a single epipole and corresponding points in $\mathcal{S}$ and $\mathcal{S}'$ lie on the same epipolar line. We call this convenient correspondence relation *auto-epipolar correspondence*. This correspondence relation is an example of the more general idea that repeated 3D geometric structure imposes 2D constraints on corresponding image features which can be used to advantage in grouping and verification.

We reserve the epipolar terminology for the case where the centres of projection of the two cameras are displaced. For some repeated structures, the transformation between $\mathcal{S}$ and $\mathcal{S}'$ does not alter the camera centre and thus does not yield an epipolar structure. However, it is still possible to construct a correspondence structure in the image which is not an epipolar geometry but has many similar advantages. An example of this is given in section 3.5 for surfaces of revolution.

Thus we have two recurring issues that arise in the context of repeated structures and in most cases where object class produces invariants that can be measured in a single image view:

1. The computation of image-measureable 3D invariants.
2. The correspondence relationship between the *imaged* features of the 3D structure, and associated grouping strategies.

In the next few sections we review some mature examples of repeated structure [36, 47] where the discussion is organised around these two issues.

3.4.1 Bilateral Symmetry

In the case of a single bilateral symmetry, the repeated structure is the half object on one side of the symmetry plane. A single camera imaging a bilaterally symmetric object is equivalent to two identical cameras, viewing the half structure, where one camera is transformed to the other by a reflection in the object symmetry plane. A similar observation was made by [26, 39], though in the context of a calibrated camera. Below we give examples of 3D point sets and space curves with a single bilateral symmetry.

3D geometry Lines joining corresponding points (on either side of the symmetry plane) are parallel and orthogonal to the plane of symmetry. There is a natural coordinate system provided by these correspondence directions and the symmetry plane (figure 9). The correspondence lines intersect the symmetry plane at the midpoint of the corresponding points.

Measuring Invariants Perspective projection does not preserve mid-points. However, the images of the 3D midpoints can be computed (see below). All 3D mid-points are co-planar (they lie on the symmetry plane). There is a projective transformation between the set of imaged mid-points and the 3D points on the plane of symmetry. Thus planar projective invariants can be measured in the image from the computed midpoints.

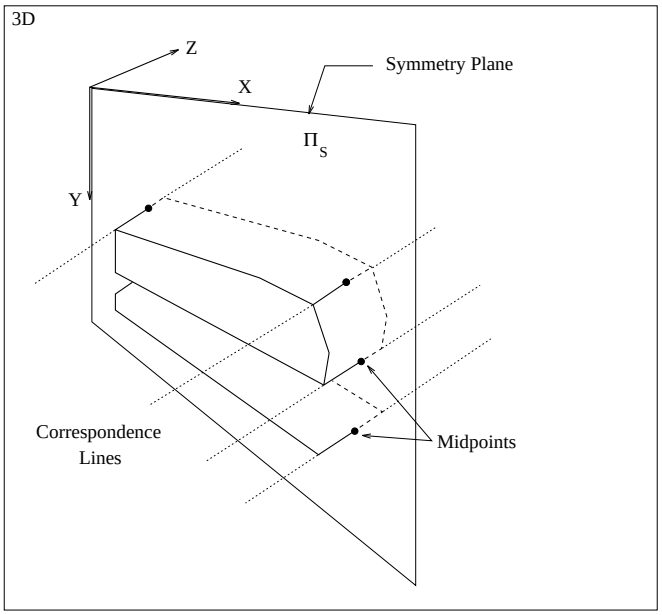


Figure 9: The natural coordinate frame for an object with bilateral symmetry. The XY plane is the plane of reflection, and the Z axis is parallel to lines joining corresponding points. Note, all mid-points are coplanar.

The image of the 3D midpoints can be computed using a property of equally spaced points (see [64]): three collinear points, separated by the same distance, and taken with a point at infinity have a harmonic cross-ratio. Since the point at infinity on the line joining two corresponding points is imaged as a vanishing point (see below, and figure 10), it can be observed. Thus, the position of the midpoint in the image can be computed from the image coordinates of the corresponding points and from the image coordinates of the vanishing point. Furthermore, since computing a point that has a fixed cross-ratio with respect to three other points is linear, there is a unique solution. Other geometric methods for computing the imaged mid-point are available, based on point pairs or triplets.

The 3D structure can be reconstructed up to a projective ambiguity, based on the equivalence with uncalibrated stereo [16, 29]. 3D projective invariants can be measured from this recovered structure. In the case of bilateral symmetry, structure is recovered to better than a projective ambiguity because of the orthogonality constraints available in the "natural" coordinate frame [18, 60].

Correspondence and Grouping The epipolar structure in this case arises from the parallel lines joining corresponding points (on each side of the object). Under perspective projection, these correspondence lines (the epipolars) image to a family of lines converging to a single vanishing point (the epipole). The epipole can be determined using two pairs of corresponding points (figure 10). Once the epipole has been computed, further correspondences are found by a 1D search on the epipolar line. This is a recurring theme — the constraints that define the object class not only show how invariants may be recovered, but also facilitate and direct the image grouping.

We demonstrate two examples of bilateral symmetry, a polyhedral point set and a space curve.

1. **3D polyhedron:** figure 11 shows different views of the 3D reconstruction of a stapler obtained from a single view. The reconstruction is placed in a Euclidean frame to give a normal presentation of the object shape, but any projective frame could be used.

2. **A space curve:** corresponding points on the two imaged space curves are determined using the epipolar geometry. Four different views of the reconstruction for the outline of a spoon are shown in figure 12 (the 3D projective representation has again been constrained to lie in a believable Euclidean frame).

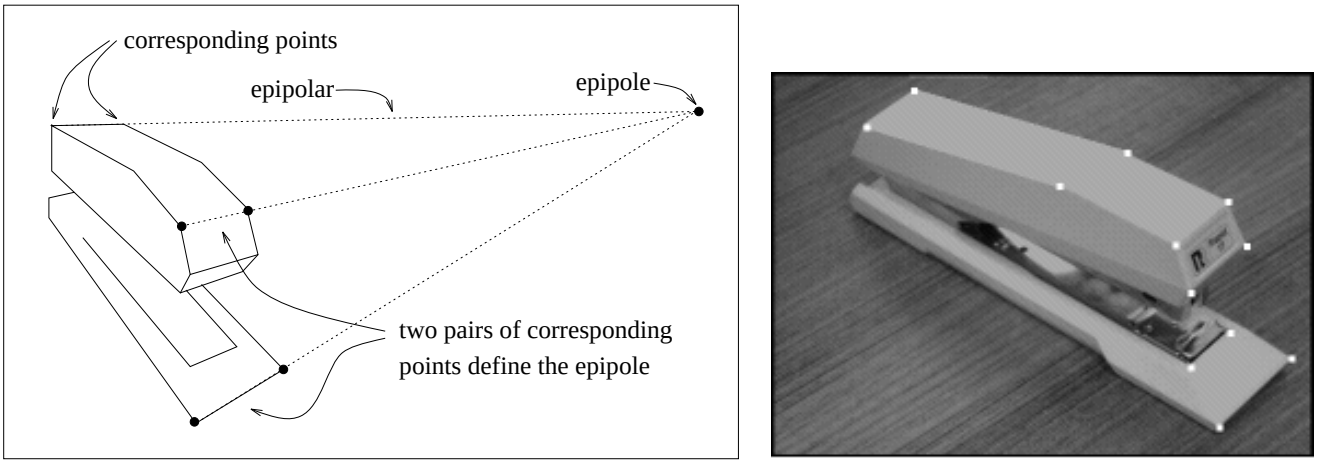


Figure 10: (a) The epipole can be located using the intersection of lines between two corresponding points on a bilaterally symmetric object (the points are marked by solid circles). Epipolars can then be constructed through the epipole to aid correspondence. (b) Typical corresponding points determined in this manner.

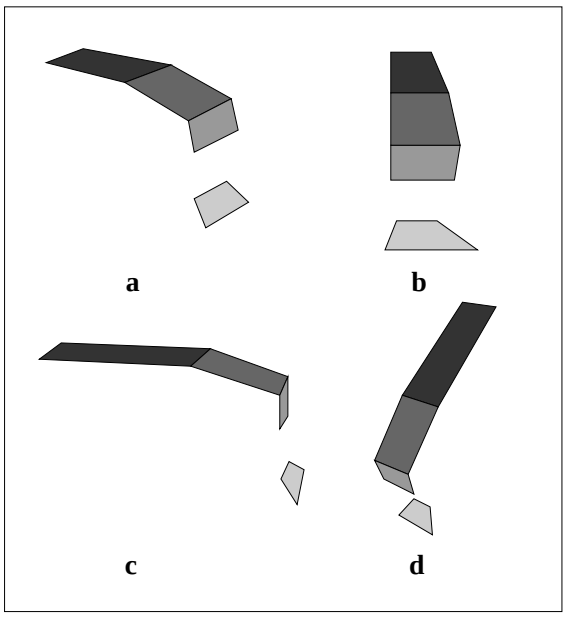


Figure 11: Three dimensional structure is recovered, modulo a projectivity, from the single view of the points marked on the stapler in figure 10b. Four typical views are shown with only (a) at a viewpoint close to that of the original image. Note the collinearity of the line segments in (b), this demonstrates the accuracy of the recovered structure.

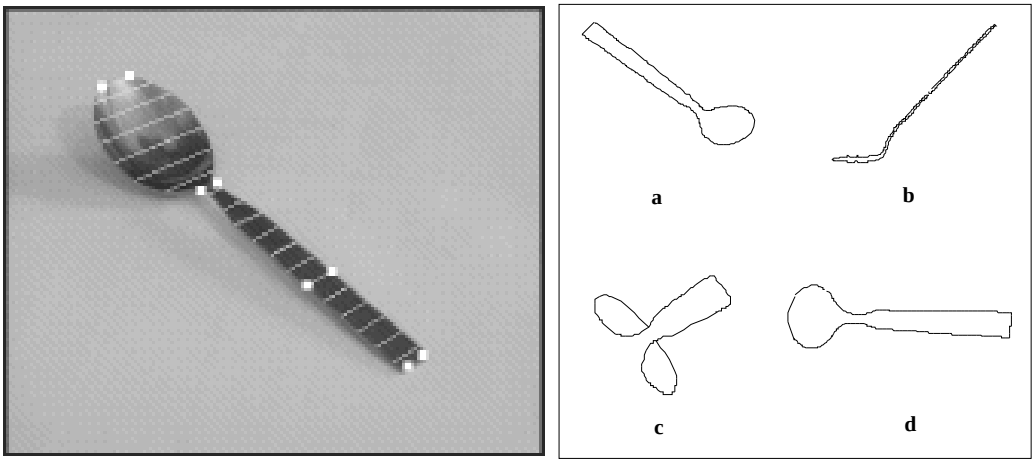


Figure 12: A single view of an object with a plane of bilateral symmetry, such as a teaspoon, is sufficient to allow a full 3D projective reconstruction. Only two pairs of distinguished points are needed for the approach, these are recovered from surface markings and can be used to determine the epipolar structure of the image. Using this epipolar structure an arbitrary number of correspondences can be produced. Four different views are shown of the 3D reconstruction computed from the image. The construction works very well: note the planarity of the handle recovered in (b), and the full 3D shape in all of the images.

The reasoning outlined above can be applied to objects with more than one bilateral symmetry [18] and to objects projectively equivalent to ones with bilateral symmetry.

3.4.2 Translational Repetition

In this case the structures, $\mathcal{S}$ and $\mathcal{S}'$, are related by a 3D translation. As described above, the structure of $\mathcal{S}$ can be recovered up to an affine ambiguity, from a single image of the duplicate structure.

Measuring Invariants Affine invariants are computed from the perspective image in three stages. First, structure is recovered up to a projective ambiguity using uncalibrated stereo. Second, the plane at infinity [64] is determined in this projective coordinate frame as follows: a line on $\mathcal{S}$ is parallel to its counterpart on $\mathcal{S}'$, so the intersection of corresponding lines is a point $\mathbf{P}$ on the plane at infinity. The image, $\mathbf{p}$, of $\mathbf{P}$ is computed by intersecting the imaged corresponding lines. Since $\mathbf{p}$ lies on both lines its 3D position, $\mathbf{P}$, can be determined by stereo. Three such points determine the plane at infinity. Third, the structure is projectively transformed such that the plane at infinity has the standard form $X_4 = 0$. The structure is then known up to an affine ambiguity, and affine invariants measured from this structure have the same value as invariants measured on the 3D Euclidean structure $\mathcal{S}$.

Correspondence and Grouping As in the bilateral symmetry case, lines joining corresponding 3D points are parallel. The image correspondence is again auto-epipolar (all corresponding lines intersect in a single epipole).

As an example, invariants are calculated from the images of two translated polyhedral structures shown in figure 13. The translation between the duplicated structure differs in each case demonstrating that the invariants are associated with the structure itself, i.e., $\mathcal{S}$. Affine invariants are computed for the 3D vertex positions which are computed using the epipolar geometry of the translated copies. The values of the invariants are given in table 3. The differences between the invariants computed from each image are small, even though the translation vector between the speakers and the viewpoint varies significantly. In many cases of such repeated structures, the object copies are rigidly connected, but this example illustrates that the affine invariants are independent of the translation vector as well.

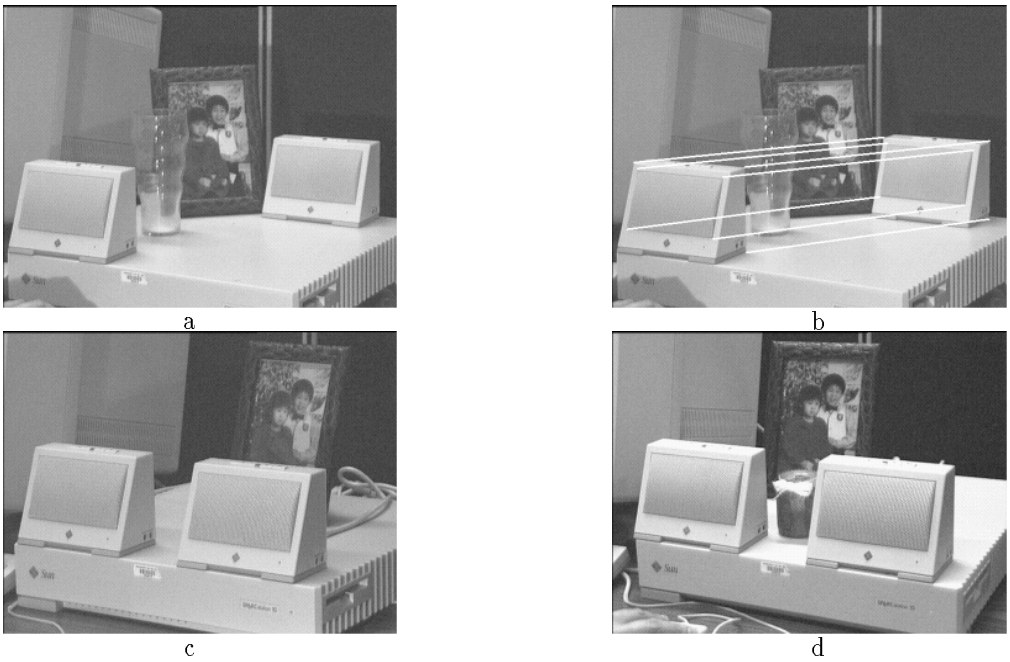


Figure 13: One object (a speaker) repeated under translation. The epipolar correspondence lines for image (a) are shown in (b). The translation vector is different for images (a) and (c) and the same between (c) and (d). Affine invariants computed from these images are compared in table 3.

image a	image c	image d
-0.2249	-0.2324	-0.2317
-0.0642	-0.0685	-0.0626
1.2833	1.2979	1.2849

Table 3: Comparison of 3D affine invariants computed for the speaker from figure 13. The invariant is the 3D position of one corner of the speaker in an affine frame defined by four other points on the speaker. The values are fairly stable, even though the images have different translation vectors and viewpoints.

3.4.3 Other Repeated Structures

The notion of structure repetition under a transformation is extensible to more general situations. It is not necessary that the copies be Euclidean equivalent and repeated under translation. The copy transformation can be a full 3D projective transformation whilst still preserving an epipolar correspondence within the image. The 3D reconstruction of the object geometry is then known only up to a 3D projective transformation of space.

In the case that there are three or more Euclidean equivalent structures, the geometry can be recovered up to a 3D similarity. This follows from the equivalence of this case to three views of a single object taken with an identical camera, where it has been demonstrated that structure can be recovered up to a 3D similarity [17].

It is also interesting to speculate about *approximately-repeated* structures. Suppose that the structure is not repeated according to a rigid 3D transformation but is only an approximation to such a transformation. This approximate repetition occurs in natural objects such as animals and vegetation. It seems that the invariants which one can compute from an idealised form of the approximate repetition will not be very far from an invariant description of the actual structure. For example in a bunch of grapes, it can be assumed that each grape is copy of the other under an affine transformation or perhaps even a scaled Euclidean transformation. Another example is texture which can be thought of as a statistical repeated

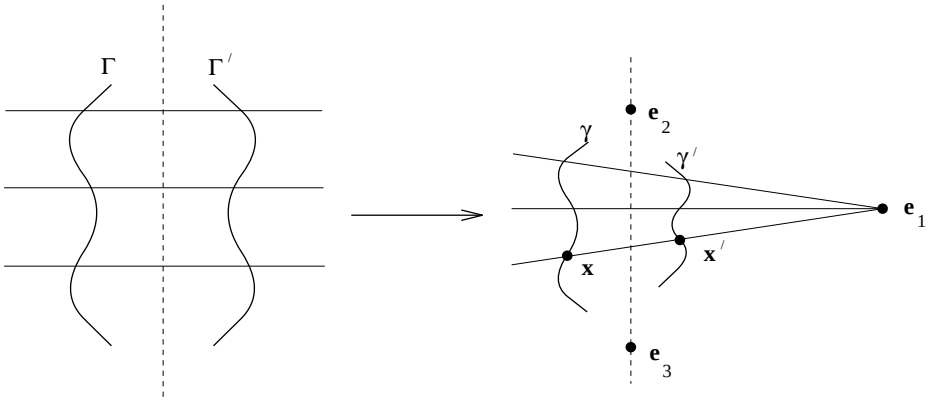


Figure 14: The profile of a surface of revolution is projectively equivalent to two curves with bilateral symmetry. Under a projective transformation parallel correspondences (left) converge to a vanishing point (right). Corresponding points $\mathbf{x} \leftrightarrow \mathbf{x}'$ are related in this case by a particular plane projective transformation, $\mathtt{T}$, called a planar harmonic homology. The transformation has a line of fixed points, the image of the axis of symmetry, which result from two of the eigenvalues of $\mathtt{T}$ being equal. There is also a fixed point, $\mathbf{e}_1$, not on the line, called the centre of the homology which defines correspondences between symmetrical points on each side of the contour. That is, corresponding point pairs and the centre of the homology are collinear. The cross-ratio of $\mathbf{e}_1$, the corresponding points $\mathbf{x}, \mathbf{x}'$, and the intersection of their join with the axis, is harmonic. (The line of fixed points is $\mathbf{e}_2 \times \mathbf{e}_3$, where $\mathbf{e}_2$ and $\mathbf{e}_3$ are the eigenvectors with equal eigenvalues. The third eigenvector, $\mathbf{e}_1$, is distinct and non-zero, and is the centre for a pencil of fixed lines.)

structure.

3.5 Rotational Symmetry

Surfaces of revolution have had considerable attention, though generally with calibrated cameras [13], or as a special case of a generalised cylinder [53, 72].

Profile Geometry The image curves forming the two "sides" of the profile are related by a plane projective transformation, $\mathtt{T}$, with the property that $\mathtt{T}^2 = \mathtt{I}$. Such a projective transformation is called a *planar harmonic homology* [64]. It arises in this case because the image transformation is a conjugate reflection (whose conjugating element is a projective transformation). To see this, construct the plane containing the axis of the surface and the optical centre. The surface then has a mirror symmetry in this plane, as does the cone of rays through the optical centre and tangent to the surface. This cone yields the profile when it is intersected with the image plane. Clearly, the contour generators are, in general, space curves, related by a mirror symmetry in space. If the image plane is perpendicular to the plane of symmetry, then the profile has a mirror symmetry; but the profile for any other image plane is within a projective transformation of the perpendicular plane.

In this case, there is no epipolar geometry defined, since reflection in the symmetry plane does not move the optical centre. However, a correspondence relation still exists and is generated by the planar homology between the opposing sides. A planar *harmonic* homology (see figure 14) is a special case of a planar homology (see figure 20b) for which the characteristic invariant of the homology is harmonic [64]. For planar homologies there is a fixed point which is the centre of a pencil of fixed lines which define correspondence pairs. That is, corresponding points lie on the same line of the pencil, in the same manner as the epipolar geometry of translated cameras.

Measuring Invariants The intersections of "corresponding" profile bitangents lie on the projection of the object's axis. The image intersection points are projections of the intersection points between planes bitangent to the surface and the 3D object axis. This point is viewpoint independent. This is shown schematically in figure 15. Four such points are sufficient to measure a cross-ratio (the points are

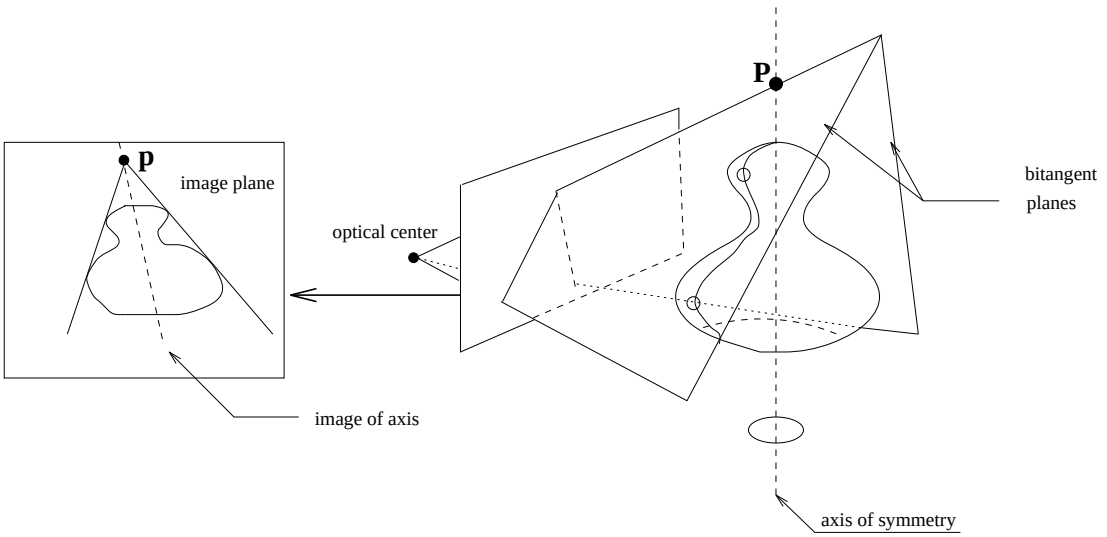


Figure 15: A rotationally symmetric object, and the planes bitangent to the object and passing through the optical centre, are shown. It is clear from the figure that the intersection of these planes is a line, also passing through the optical centre. Each plane appears as a line in the image: the intersection of the planes appears as a point, $\mathbf{p}$, which is the image of the point, $\mathbf{P}$, at which the bitangent planes intersect the axis of symmetry.

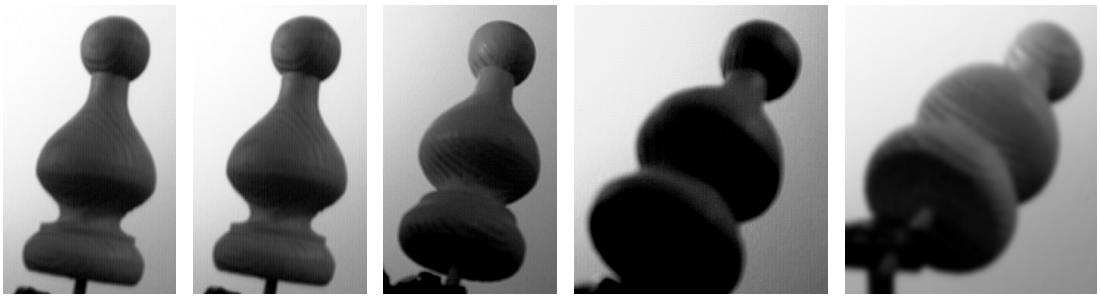


Figure 16: Five perspective images of a surface of revolution at different inclinations. The invariant values are given in table 4.

collinear in space, all lying on the axis of rotational symmetry). In this manner a projective invariant (the cross-ratio) is associated with the surface [20, 35, 45].

The construction extends to straight homogeneous generalised cylinders (SHGCs) [3, 35]. Again, the intersections of corresponding profile bitangents correspond to a viewpoint-invariant 3D point, and are collinear, so cross-ratios can be formed. Figure 16 shows images of a surface of revolution with the calculated invariants given in table 4.

Correspondence and Grouping As described above, the profile of a rotationally symmetric surface can be separated into two "sides", which are related by a planar harmonic homology, $\mathtt{T}$. There are a number of consequences of this result:

1. The two sides of the profile can be grouped by associating curves which are projectively equivalent. For example, by matching projectively equivalent concavity curves. This correspondence can be achieved automatically by the planar recognition system described in section 2.

2. If the projective transformation between two projectively related curves is not a harmonic homology, then the grouped curves can be ruled out as arising from a surface of revolution. This is simply tested by checking if $\mathtt{T}^2 = \mathtt{I}$.

angle	cross-ratio	length ratio
45.0	0.486187	1.40862
40.0	0.490561	1.98153
35.0	0.486796	2.14017
25.0	0.486640	2.38409
15.0	0.486260	2.70539
0.0	0.494849	4.13687

Table 4: Stability of invariants for a surface of revolution. The invariants are computed from measured points. The angle is the inclination of the axis of the lamp-base to the camera plane (figure 16). Typical affine (length ratio) and projective (cross-ratio) invariants are shown. Note that the value of the affine invariant changes at extreme angles, whereas to two significant figures the projective invariant remains stable to the second decimal place.

3. Under real imaging conditions the transformation `T` relating the two sides of a profile will be close to affine. This quasi-invariant condition can be used in two ways: first, lines joining corresponding points on the two sides of the profile will be almost parallel. Second, relative (not scalar) affine invariants can be used to match concavity curves [43].

4. `T` provides point to point correspondence between the sides of the profile, this can be used to disambiguate bitangent matches. This correspondence can be used to *repair* missing profile portions, filling in gaps by transforming over points from the other side of the profile.

5. The projected axis can be determined directly from the projectivity as a line of fixed points of the homology [64].

To illustrate the power of this grouping constraint, figure 17 shows an image with many surfaces of revolution of various types and sizes. The matched concavities are partitioned into sets, and the profile curves corresponding to each set are grouped. The entire process is automatic and relies only on the properties of the homology between symmetrical portions of the profile.

3.6 Canal Surfaces

A canal surface is the parallel surface of a space curve. It is the locus of points which are a fixed perpendicular distance from the curve. Equivalently it can be generated as the envelope of a sphere swept with the centre on the curve. Common examples are pipes or tubes such as occur in plumbing. In the following we consider canal surfaces for which the generating curve, α, is *planar*. For such surfaces we have:

> Under general viewing conditions, an inflection in the generating curve gives rise to two inflections in the profile, one on either "side". The tangents on the contour generator at the pre-images of the profile inflections, and the tangent at the generating curve inflection, are parallel.

The consequence of this is that tangents at the paired profile inflections intersect in the vanishing point of the generating curve inflection tangent. This vanishing point lies on the vanishing line of the plane of the canal surface generating curve. This is illustrated in figure 18a. Note that a straight line is simply a degenerate inflection, so invariants can be obtained from a piecewise linear generating curve.

Computing Invariants The canal surface is the envelope of spheres, and the canal profile the envelope of sphere profiles [61]. Under affine imaging conditions, provided the image has the correct aspect ratio (scaled orthographic projection), the sphere profile is a circle, and the sphere centre projects

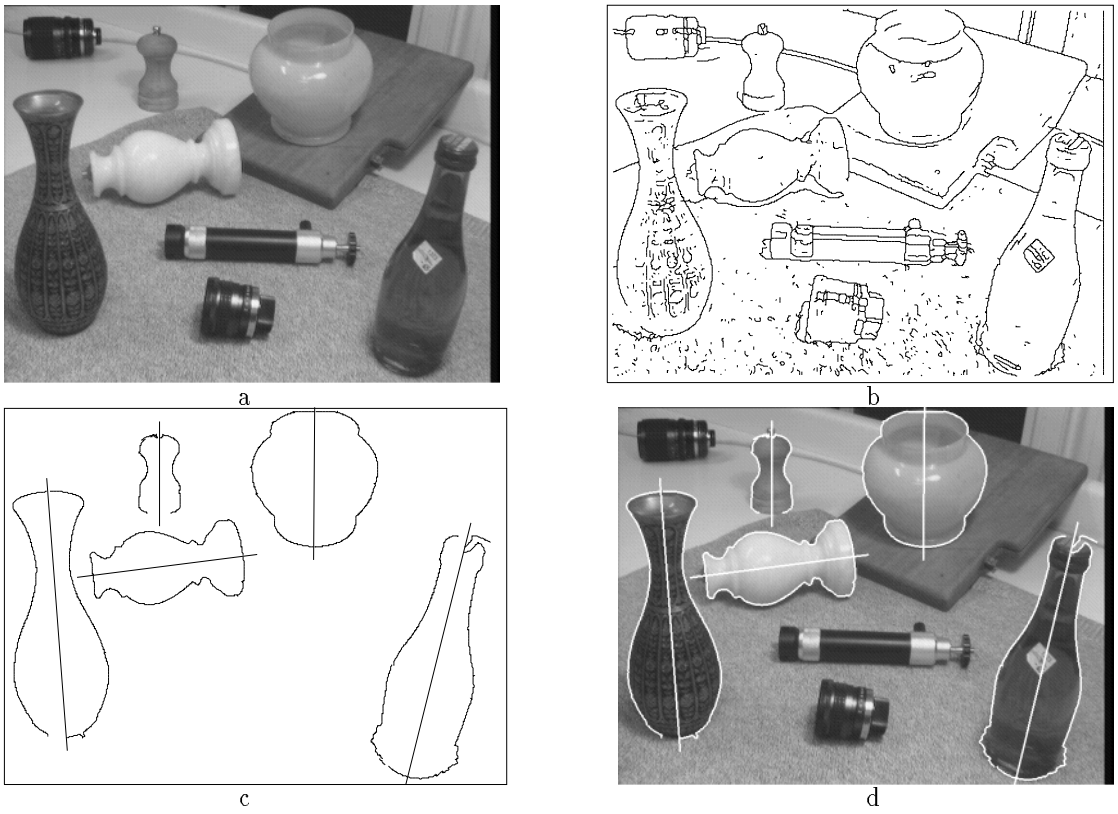


Figure 17: (a) Original image containing several surfaces of revolution. (b) The linked edges computed from (a). (c) Extracted surface of revolution profiles with axes computed automatically using grouping constraints based on a harmonic homology. (d) Extracted surface of revolution profiles and axes superimposed on the original image.

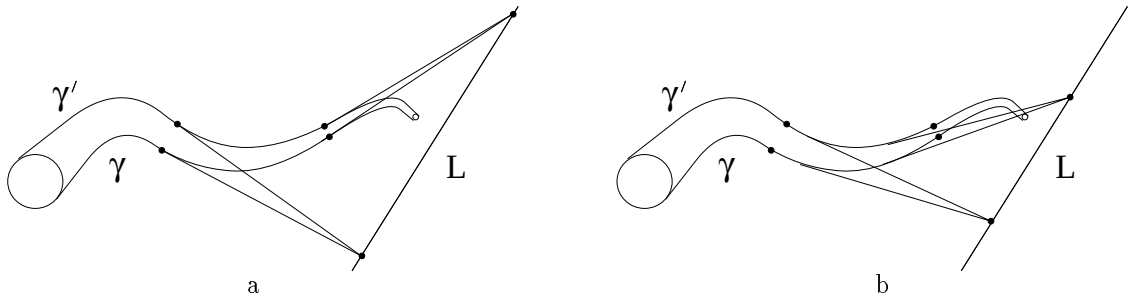


Figure 18: For a canal surface with a planar axis, (a) inflections in the profile occur in pairs for each inflection of the axis. The intersection of a pair of inflection tangents determines the vanishing point of the tangent line at the axis inflection. Two such vanishing points determine the vanishing line, $\mathbf{l}_\infty$, of the plane of the axis; (b) corresponding profile tangents (profile points arising from the same surface circular cross-section) also intersect on $\mathbf{l}_\infty$. Their intersection point is the vanishing point of the corresponding axis tangent line.

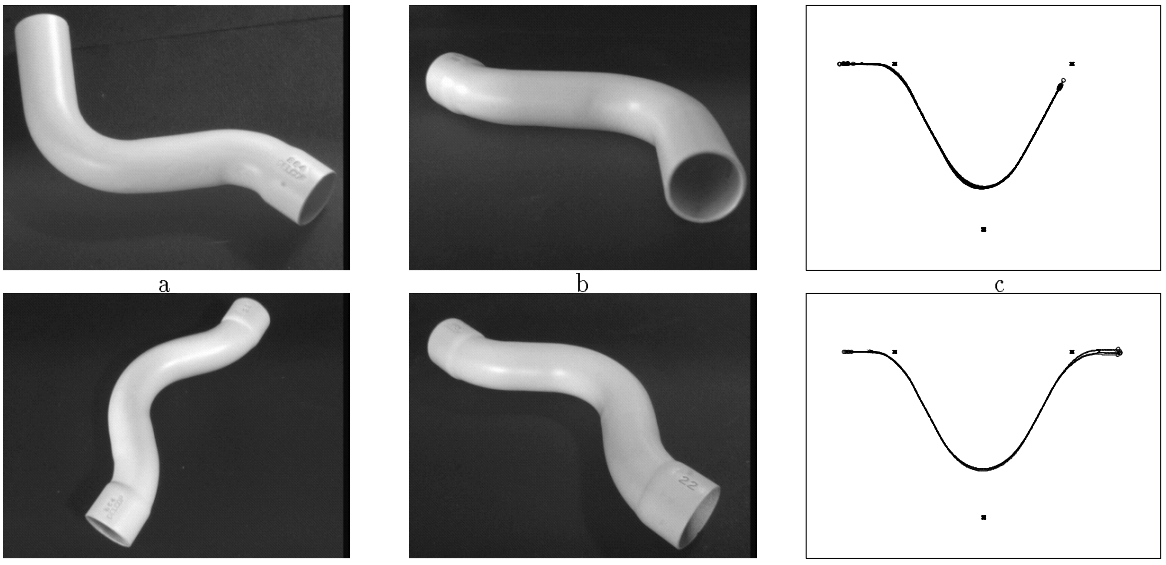


Figure 19: Affine normalisation of canal surface symmetry sets. Each row shows two images of the same pipe, and the symmetry set from these views and others transformed to an affine canonical frame. The canonical frames (c) and (f) contain symmetry sets generated from thirteen and five images respectively. At least half of each set show significant perspective distortions. Note the variation in pipe width in the middle column due to perspective. Affine (as opposed to projective) normalisation can be achieved because the vanishing line of the plane of the generating curve is known (It is computed using the construction of figure 18a). The canonical frame curves are clearly very stable against variation in viewing position. Moreover, different pipes can be distinguished based solely on this affine representation [51]. The slight instability present towards the ends of the canonical frame curves are due to errors in the extracted symmetry set which occur where the pipe radius changes.

to the circle centre. The circle centre can be recovered from the *symmetry set*[5] of the canal profile, which is thus projectively related to the generating curve, α. This relation is exact under affine imaging conditions, and is an extremely good approximation under perspective with a realistic field of view — another example of a quasi-invariant. Consequently, invariants computed for the symmetry set are invariants of the generating curve. For example, invariants can be computed from measurements on the symmetry set curve in a canonical frame in a similar manner to the 'footprints' of Lamdan *et al.* [34]. Figure 19 shows examples of such curves.

Correspondence and Grouping As in the case of bilateral symmetry, the constraint of a canal surface with a planar generating curve establishes a planar projective constraint in the image. In this case two vanishing points determine the vanishing line, $\mathbf{l}_\infty$, of the plane containing the generating curve. Subsequently, inflections on the profile can be paired by the intersections of their tangents on this vanishing line. Furthermore, it can be shown that this intersection constraint holds for *all* corresponding profile points, i.e.,

> Corresponding profile tangents (i.e., points whose pre-image is on the same circular cross-section) intersect on $\mathbf{l}_\infty$.

See figure 18b.

[5] The symmetry set is the locus of centres of circles bitangent to a plane curve. It is studied in detail by Giblin & Brassett [23].

Under affine imaging conditions the two sides of the profile are parallel curves of the symmetry set (the projection of the generating curve). This follows directly from the profile curves being the envelope of constant-radius circles swept along the symmetry set.

3.7 Polyhedra

Recovering the structure of polyhedral objects from a single view has been widely explored, with the most detailed study appearing in [67]. In this work, Sugihara shows that the incidence equations between polyhedral vertices and faces, observed in the image, lead to a linear system of equations in the coefficients of the polyhedron's faces and image observations.

The equations in this system are incidence equations for vertices of the polyhedron incident on plane faces. In particular, given vertex $\mathbf{V}_i = (X_i, Y_i, Z_i)$ lying on face $\mathbf{F}_j = (A_j, B_j, C_j, 1)$, it must be the case that

$$A_j X_i + B_j Y_i + C_j Z_i + 1 = 0.$$

Assume that the camera image plane is the plane $Z = 1$ and the focal point is at $(0, 0, 0)$; these assumptions can be accounted for by the geometric ambiguity in the reconstruction. Then $\mathbf{V}_i$ projects to image point $(u_i, v_i) = (X_i/Z_i, Y_i/Z_i)$. If vertex $\mathbf{V}_i$ also lies on face $\mathbf{F}_k$, we can divide the incidence equations by Z_i and subtract to eliminate $1/Z_i$, obtaining:

$$(A_j - A_k)u_i + (B_j - B_k)v_i + (C_j - C_k) = 0$$

where u_i and v_i are known, and the coefficients of the planes are unknowns. This system of equations always has at least a three dimensional family of solutions, corresponding to a polyhedron where all faces are the same plane (since a plane figure has three degrees of freedom in 3D space). If the family of solutions is four dimensional, then a generic element of the family is a system of planes that is projectively equivalent to the faces of the original polyhedron. This case holds when the reconstruction of the polyhedron can not be made impossible by a small shift of the vertices, that is, is "position free" in the terminology of Sugihara, and many or most of the visible faces have at least four vertices per face. Although this is by no means a generic polyhedron, it is a useful case because many human artifacts satisfy these constraints. Given the added assumption that vertices are trihedral, it is possible to reconstruct faces for which only two edges are visible; thus, on viewing a cube, all six faces can be recovered. This leads to a novel formulation of the aspect graph idea, where substantively fewer aspects are necessary for effective representation. The case where the polyhedron consists only of triangular faces is equivalent to an unconstrained set of points. That is, a set of points can always be triangulated to form a polyhedron. As in the case of a general point set (section 3), vertex positions are unconstrained by the image view and no invariants can be constructed.

Computing Invariants Assuming an uncalibrated camera, it can be shown [59, 60] that for polyhedra that lead to a system of equations having a four dimensional solution space (such as cubes) any solution of this system is projectively equivalent to the original (Euclidean) polyhedron. Consequently, projective invariants of the solution are the same as those measured on the original polyhedron.

Correspondence and Grouping Approaches to grouping and correspondence for this class are well established from the decade or so of blocks world vision research. The main basis for grouping is topological, where one seeks to construct a complete polyhedral structure with consistent incident relations between vertices, edges and faces.

For particular sub-classes of polyhedra, and for particular aspects, further constraints are available. For example, a cube has three major directions which define a triple of vanishing points in the image. All edges aligned with a major direction must pass through the same vanishing point. Similar incidence constraints apply to any polyhedra projectively equivalent to a cube. Constraints of this type can be used to extract a polyhedral wireframe from a polyhedral silhouette, inferring internal boundaries.

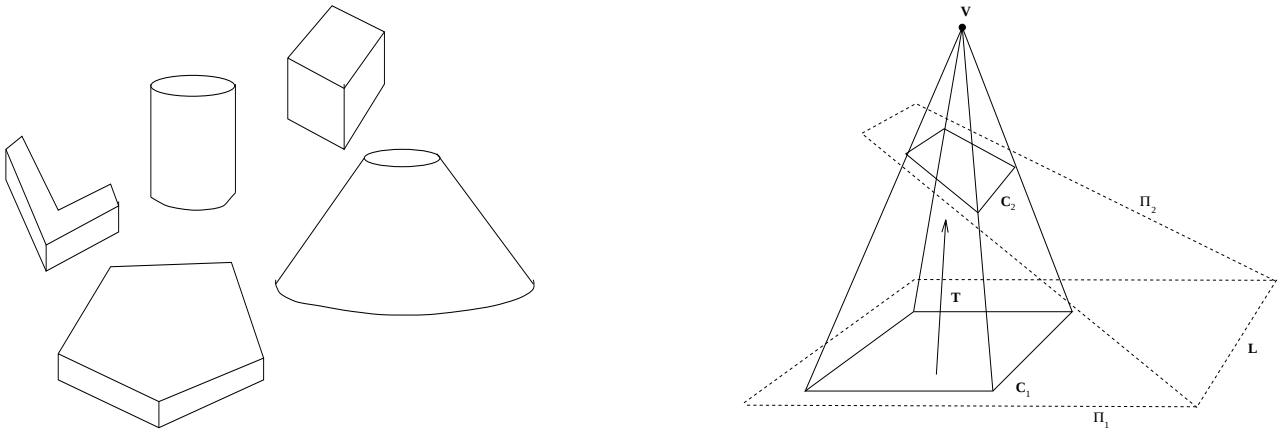


Figure 20: An extruded surface is a section cut from a general cone by two planes. (a) A range of examples of extruded surfaces; note that for most examples, the vertex is at infinity. (b) The top and base image curves, C_1 and C_2, of an extruded surface are related by a particular projective transformation $\mathtt{T}$, called a planar homology. Corresponding points lie on lines through $\mathbf{V}$, which is the fixed point of the transformation (the centre or vertex). The line $\mathbf{L}$, which is the image of the intersection of the two planes that cut the "cone", is a line of fixed points of the transformation (the axis).

3.8 Extruded Surfaces

An extruded surface is a special case of a generalised cylinder, formed by a section cut from a general cone by two planes (see figure 20a) in such a way that the section of surface does not include the vertex of the cone[21]. This is the projective generalisation of a surface formed by a system of parallel lines, with plane ends (such a surface can be extruded from a nozzle). Extruded surfaces, and surfaces made up from extruded components, are extremely common — examples include most tin cans, boxes, books, and many plastic bottles.

Outline Geometry The base and top curve are perspectively related in 3D, and thus related in the image by a projective transformation, $\mathtt{T}$. This transformation is a *planar homology* [64]. It has five degrees of freedom: the vertex (2 dof), axis (2 dof) and the cross ratio defined by the vertex, a pair of corresponding points, and the intersection of the line joining these points with the axis (1 dof). The cross-ratio is the same for all points related by the homology[6]. As in the case of a planar harmonic homology (figure 14) a planar homology has a line of fixed points and a fixed point not on this line:

1. The homology vertex is the projection of the 3D cone vertex. It is the fixed point of $\mathtt{T}$.
2. The homology axis is the projection of the line of intersection of the top and base planes. It is the line of fixed points of $\mathtt{T}$.

The profile curves of an extruded surface are a pair of lines which intersect at the image vertex.

Computing Invariants The projective geometry of an extruded surface is completely defined by three elements: a plane cross-section; a cone vertex, not on the plane; and, a line in the plane. The plane cross-section and vertex together define the cone. The line is the axis of the pencil of planes which intersect the cone to generate the top and base curves. These elements can be recovered from an image of the surface since the cross-section of the cone is determined up to a projective transformation from the imaged base curve or top curve, and the line is the line of fixed points of the projective transformation relating top and base image curves.

[6] In the case of a harmonic homology, the cross-ratio is harmonic, i.e., known, so there are only four remaining degrees of freedom. The sides of the profile of a surface of revolution are related by a harmonic homology, as described in section 3.5.

In essence, the invariants of an extruded surface are those of the plane cross-section plus an extra line in the plane, obtained from intersection of the base and top planes. Thus extra invariants are available over the plane cross-section alone. For example, in figure 20 a five line invariant can be computed from the image, although the top curve only contains four lines. In the case that the top and base planes are parallel, affine invariants of the curve can be measured from a perspective image.

Correspondence and Grouping As described above, for an extruded surface the top and base image curves are related by a planar homology, $\mathtt{T}$. Grouping proceeds by finding curves which are projectively related. The class assumption can then be tested immediately since the projective transformation must be a homology if the curves are from an extruded surface (for example, two of the eigenvalues will be equal). The homology then defines the vertex, axis, and correspondence for the surface, which is used for further grouping. Additionally, since the surface is ruled, and all rulings pass through the vertex, the intersection of line segments in the profile determines the imaged vertex. Similarly, all corresponding point pairs on C and C' (figure 20b) define a pencil of lines which pass through the vertex, and corresponding tangents intersect on the line of fixed points, $\mathbf{l}$.

3.9 Algebraic Surfaces

Algebraic surfaces are surfaces for which a single polynomial vanishes: examples include spheres ($x^2 + y^2 + z^2 - 1 = 0$) and ellipsoids which are both degree two surfaces (quadrics), and a wide range of popular surfaces in modelling such as rational bicubic patches. Smooth quadrics are all projectively equivalent (just as all conics are projectively equivalent) so that there are no projective invariants of the surface to recover from images. Although a single quadric does not have any projective invariants, two or more quadrics do. Similarly, if the surface has degree 3 or greater, there are projective invariants to recover from images. In theory these invariants can be recovered from the surface profile alone [22], though this has not been implemented in practice.

4 An architecture for a 3D recognition system

We have demonstrated that a large vocabulary of 3D invariants can be derived from the geometric constraints associated with object class definitions, e.g., that of a surface of revolution. In general, these curve, surface or volume class constraints enable the construction of invariants, and permit at least partial reconstruction of the 3D structure from a single perspective view. Class constraints also provide image feature grouping mechanisms and associated indexing machinery.

The work to date, however, has focused on the derivation of invariants, structure recovery, and grouping for single object classes. Experimental validation has been restricted to isolated objects of a given class against an uncluttered background. An important next step is to integrate the approaches which have been developed into a unified 3D object recognition system. It is only in the context of a full system that the effectiveness of a class-based invariant representation for recognition can be convincingly demonstrated.

4.1 Fundamental Principles

Object recognition should be based on 3D geometric descriptions, both of objects and of the relationships between objects. To date, systems have largely ignored these relationships; as we show below, requiring *consistency* in inter-object relationships yields substantial information. In the architecture we describe, this information is encapsulated in an internal database, known as the *scene*. The scene provides a working reconstruction against which hypotheses can be checked to provide immediate detection of a false recognition hypothesis. For example, if two objects are hypothesised in such a way that one must be wholly occluded by the other, then at least one of the object hypotheses must be wrong.

Central to the architecture is efficient management of *control* of each level. Even for relatively small images, vast numbers of hypotheses for feature correspondences and model interpretations can be constructed. It is impossible to explore all avenues of interpretation, so some basis must be established for

scheduling feature combination, hypothesis generation and verification of hypotheses. The priority of scheduling should be based on a tradeoff between the cost and the benefit of a computation.

Finally, *class* pervades the architecture, influencing segmentation, grouping, indexing, and hypothesis confirmation.

4.1.1 Class

The idea that objects should be organised in a taxonomy and classified before proceeding to recognition is a natural and well-accepted principle. The problem with this philosophy is that many ontological distinctions are not manifested in observable properties, for example, the difference between a hollow container and a solid block. Our geometric approach to object classification is based directly on visible features; its main strength is that it is not vested in abstract, philosophical differences, so much as in image observable distinctions. Object class has its most important effects in considering feature grouping and the structure of the modelbase.

Class drives grouping, as opposed to the usual "heuristics" that are used to associate image features. Each object class defines a grouping mechanism based on its image invariant relation. For this reason, object class is typically settled at an early stage in the grouping process, and identity emerges only after modelbase access. For example, there is no point in grouping lines into faces, as required for polyhedral class grouping, to recognise a rotationally symmetric object. A rotational symmetry hypothesis requires image curves related by a planar harmonic homology. The projective matching of these curves can be carried out by computing and matching projective invariants of the curves. This is an application of planar object recognition techniques within a single image.

Class determines the access functions and partitioning of the modelbase. The modelbase itself is a collection of facts about objects and their properties. These facts must be organised in such a way as to allow easy retrieval; a hashing mechanism is appropriate. By the time the modelbase is accessed, the object group will contain a strong implicit hypothesis about object class — for example, a pair of concavity-curves cannot be passed to the polyhedral hashing mechanism. In fact, the modelbase can be viewed as a rather conventional database, organised to answer certain queries very efficiently.

4.1.2 Consistency

Consistency tests arise from computing and representing relationships between objects. To date, there have been few "hard" geometric consistency tests for inter-object relations. In fact, strong geometric tests emerge from the observation that objects share the same Euclidean frame and the same camera. These tests make it possible to recover the Euclidean identity of objects even if the calibration of the camera is initially unknown.

Suppose that models are Euclidean (i.e., the relation between the model and object is an Euclidean transformation, as opposed to the projective transformation of the 2D recognition system), and recognition hypotheses have been formed for a number of objects. Even though the camera is uncalibrated, the Euclidean consistency of the recognition hypotheses can be tested by a comparison of the set of ray cones from the optical centre to each object.

The cones are determined from $\mathtt{P}$, the rank three, 3×4 projection matrix of equation (2). Given a hypothesised object, $\mathtt{P}$ is determined from the known 3D Euclidean geometry and the image features by standard resectioning (as in, for example [56]). Partitioning $\mathtt{P}$ as $\mathtt{P} = [\mathtt{M}|-\mathtt{M}\mathbf{t}]$ [29], then $\mathbf{t}$ is the optical centre which is the null space of $\mathtt{P}$. A cone of rays from the optical centre to other Euclidean objects in the scene can then be constructed.

If the hypotheses for each object are correct, the ray cones for each object should be Euclidean equivalent. That is, there will be a rotation about the optical centre which superimposes the cones for each object from each hypothesis. Thus, inconsistent hypotheses can be detected by the failure of this test and the goal is to build up the largest pairwise consistent set of object hypotheses. An example of hypothesis labelling and reconstruction is shown in figure 21.

Another consistency test involves decomposing the matrix $\mathtt{M}$ (above) as $\mathtt{M} = \mathtt{KR}$ by QR decomposition [25], where $\mathtt{R}$ is a rotation matrix, and $\mathtt{K}$ an upper triangular matrix containing the intrinsic parameters of the camera. Each hypothesis must agree on the camera intrinsic parameters and inconsistent hypotheses can be detected from differing decompositions for $\mathtt{K}$.

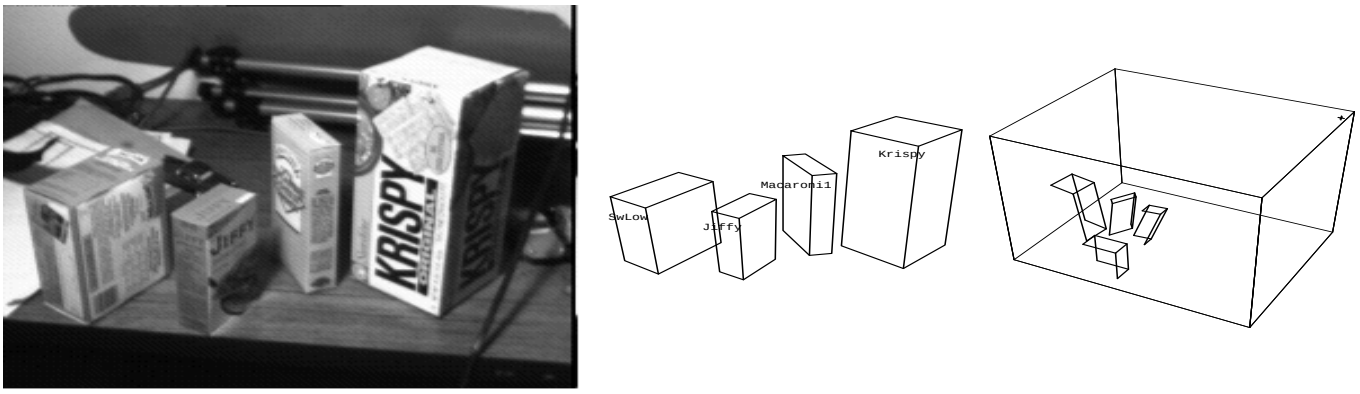


Figure 21: (a) A scene containing polyhedra all of which are projectively equivalent, but Euclidean inequivalent. (b) A labelling of the scene. (c) A Euclidean reconstruction of the polyhedral world shown in (a). The optical centre is the marked point in the top right-hand corner of the figure.

Given an image containing a large number of known objects, once the first few have been recognised and used to construct a consistent world frame, this frame can be accepted and used to prune additional hypotheses in a depth-first search for consistency. At this point, rather than searching for consistent groups of object hypotheses, individual hypotheses can be tested against the established frame with little risk of error. Furthermore, if this frame is accepted, then it can be used to condition grouping and indexing activities. The Euclidean reconstruction of the world forms the scene database.

4.2 The Architecture

Representation is organised into a number of layers as illustrated in figure 22. These stages of representation are not very different from other recognition architectures, however the three principles of *class*, *global consistency* and *control* provide a unifying theme.

Segmentation and grouping The key to successful recognition is efficient and robust feature segmentation and grouping. There are four levels of image feature representation and grouping:

Level I: **Pixel-level features** are defined with respect to an image coordinate system and reflect the quantised nature of pixel coordinates. Typically, features will be produced using an edge operator with sub-pixel accuracy, and the resulting edgels linked into a network reflecting the topology of the image boundaries.

Level II: **Geometric features** curves from level I are described in terms of geometric primitives, where appropriate. For example, algebraic curves such as line segments and conics, smooth curves, and concavities defined by bitangents.

Level III: **Generic Grouped features** This level of grouping is applied to all features produced at level II. The output is a number of groupings and databases which are used by the class-based groupers described below. Generic grouping includes: near incidence (jumping small gaps, completing corners and junctions); collinearity; marking bitangent and other distinguished points; finding sets of parallel line segments; affine or projective equivalence of curve segments (e.g., concavities). These relations can be viewed as queries to a spatially organised data base. For example, typical queries might be: "what other lines are parallel to a given line and above a certain length in the region of interest?", or "what other lines are collinear with the given line over the entire image?". In the current design there is no attempt at enforcing "backwards compatibility". For example, if a grouper at level III hypothesises that two curves should be joined there is no attempt to correct the level II representation. Ultimately, it may be important to ensure such consistency between levels.

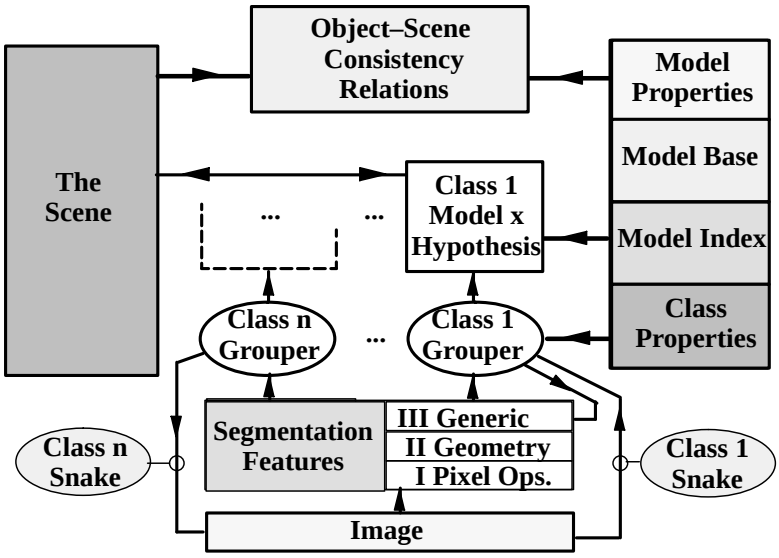


Figure 22: The proposed architecture for object recognition organised around geometric *classes* with associated grouping and indexing methods.

Level IV: **Class-based grouping** Each class has an associated "class-based grouper" that interrogates the level III groupings and databases, and attempts to form groups appropriate for its class. The grouping mechanism is based on the image invariant-relation as described in section 3 for each class. A good example is given by the rotationally symmetric class which defines a grouping constraint in terms of the harmonic homology between the corresponding sides of the profile (section 3.5).

In addition to grouping, such constraints can be used to *repair* missing portions of the outline due to occlusion or poor contrast. For example, for a surface of revolution, a "snake" or deformable template can be defined by one side and applied to the other under the transformation of the homology. The transformation between both sides can be iterated to improve the geometric correspondence of both sides. Such class-based snakes can also augment the initial edgel extraction process. Figure 23 shows an example of repair and augmentation, where a polyhedral class snake recovers poorly defined interior edges from the exterior polyhedra outline, again based on the class constraints.

Indexing and Hypothesis Combination The groups defined by each class also define the indexing function used to retrieve specific objects from the modelbase. For example, for a canal surface the indexes are computed from the symmetry set of the profile, for a surface of revolution from distinguished points on the axis.

Indexing is handled by a series of hash-tables, one per class, that take the invariants of a set of grouped features and associate with them models in the modelbase. For complex objects, there may be many feature groups that index to the object, leading to a situation where a single instance could generate many recognition hypotheses. In the planar recognition system, this problem is handled by merging consistent object hypotheses into joint hypotheses.

Forming joint hypotheses (cliques), is fairly successful for small numbers of feature groups, but for more complex objects, there are potentially quite substantial combinatorial problems. However, the

principle that feature groups belonging to the same object should accrete into a more complex feature grouping, is a good one. This accretion can be implemented in a more general fashion as follows: if a feature group results in a successful indexing attempt to a relatively small number of models, it leaves a record of that attempt in an image-scene relational data structure. When another feature group indexes to the same object or list of objects, and is within some grouping horizon of the first group, the two feature groups can be associated in a larger feature grouping, based on their correspondence to the same object structure. To make this record, the system forms a collection of keys out of the image feature position and each possible object model in turn, and stores a unique identifier for the image feature group in an image-scene database using these keys. The storage mechanism is such that, if the database is queried using a model identity and feature position, it will return any image features that indexed that model and are "near" (for some horizon) the original feature group. Note that other forms of image-scene information could be used in addition to Euclidean distance in the image; for example, an indexing hypothesis might be associated with a pose or frame hypothesis.

Verification In the planar system, two stages of verification were used: plausibility of the projective transformation taking the object from the model to image frame, and image support measured by the proportion of the back-projected model perimeter that lines up with image features. Such verification, based only on object outline, can fail through accidental correspondences with texture (for example, oriented markings such as wood grain).

To avoid this problem, verification is augmented in a number of ways. First, surface markings and surface texture will be stored for each object in the model library. During verification, the internal surface properties of an object can be compared with the properties actually observed in the interior of a model hypothesis. Second, the reliability of verification will be improved by scene consistency analysis. For example, if one object is deemed to be behind another with respect to a given camera viewpoint, then it would be inconsistent to declare a large portion of confirmed boundary for the occluded object. More generically, the "score" for a hypothesis is improved if, when portions of the perimeter cannot be matched, there is independent evidence of an occlusion occurring. For example, one piece of evidence for occlusion is that aligned "T" junctions occur at each end of the occlusion.

The modelbase The modelbase will be organised around object class. For each class there will be appropriate hash-tables for indexing, and a data base of models. For example, canal surfaces and surfaces of revolution will have separate indexing tables, containing respectively affine and projective invariants, and separate model libraries.

In the 2D recognition system (section 2.3) the modelbase typically acted as a passive repository that contained geometric models, and was indexed to identify an object. The modelbase can be more powerful than this. It can also store aggregated statistics derived from all the models in each class library. These can be used to improve efficiency. For example, suppose the maximum number of undulations of any surface of revolution in the library is stored. Then if a putative profile is returned by the grouper which has more undulations than this, there is no point computing invariants or indexing. Similarly, if there are only trihedral vertices for any polyhedra, then there is no need for the polyhedral grouper to attempt to group or index with four concurrent lines. Exploiting the modelbase in this manner can greatly strengthen the performance of the system.

Objects which do not correspond to a single volumetric primitive, i.e., *composite* [72] objects, will have multiple representations: each representation covering a possible image segmentation and grouping. For example, a mug might be represented, in the composite 3D structure class, as a surface of revolution together with a canal surface (the handle). Equally, the handle could be represented as a digital plane curve, and the mug body as a canal surface or extruded surface. All such representations will be included in the modelbase. It is only through recognition that the common concept "mug" is achieved.

The scene An additional source of constraints and parameters is the 3D scene, which can also be viewed as a database which reflects the current configuration of the world and cameras. It provides a representation of all the information currently available about the common Euclidean frame in which objects reside. This 3D spatial layout can be used at a number of stages, for example, to determine occlusion relations amongst model hypotheses, and for camera viewpoint consistency.

4.3 Model acquisition

Typically, models will be acquired from multiple views of objects. The fact that such models can serve as sufficient representations for recognition is a major advantage of the invariance approach to recognition. Our goal is to provide a model acquisition tool which permits additions to the model library to be as simple as providing four or five unoccluded views of the object. This goal was achieved in the planar object recognition system, where only one or two unoccluded views were required to construct a model.

Of course, the problem is more difficult for 3D objects, but the partitioning into classes, based on geometry, will permit the efficient grouping and correspondence construction required for model description. Since the object classes consist of 3D volumetric primitives, we expect that only a small number of views will be required for most objects, and that these views will be defined by the extraction of a sufficient set of stable features over a wide range of viewpoints. This contrasts with the construction of aspect graphs based on topology [65], which define a large number of aspects, or distinct views, which cannot be reliably distinguished from image feature groups. That is, fine topological properties of an image feature group are unlikely to be reliably recovered from image segmentation and grouping. The integrity of the object boundary topology is a secondary result achieved after an initial recognition hypothesis has occurred.

Euclidean information in object models is required for scene consistency techniques to work. One approach is to derive the Euclidean properties from self-calibrated camera views. Three or more general views with a single camera are sufficient to derive internal camera parameters [17, 30], and a 3D scaled Euclidean reconstruction for point sets.

For manufactured objects, Computer Aided Design (CAD) models can be used to provide a Euclidean description. However, it should be noted that it is often the case that CAD models used for part design do not necessarily correspond exactly to the manufactured version of the part. The description obtained from imagery is more "realistic" and incorporates many details which are not practical to include in a CAD model, such as filets and attachment hardware. Conversely, certain CAD features may be irrelevant in practice since they are not manifested visually in any image.

On the other hand, it is important to develop the idea of Platonic generalisation of a model description. For example, if an image curve is sufficiently straight, it can be interpreted as an instance of an "ideal" line even though its manifestation as an image feature is never perfectly straight. Similarly, a pair of profile curves may match closely enough to be considered the outline of a rotationally symmetric object, even though they are not perfect instances of such a projection. The benefit of constructing a Platonic ideal description is that over a large set of views and feature reconstructions, the ideal description represents the natural *mean* over the set of reconstructions. Also, the Platonic description is in accordance with the formal mathematical constraints used to drive the grouping and indexing process.

4.4 MORSE

These ideas are being incorporated into a system, called MORSE, whose implementation is currently underway. MORSE is named after the detective character originated by Colin Dexter, who is able to ferret out truth given apparently unpromising evidence. Our earlier system for 2D object recognition is called LEWIS, the name of Morse's less capable assistant. MORSE will provide an environment for research on object representation for recognition, by providing a context in which issues such as the distinctiveness of representations, the usefulness of feature groups, and the significance of consistency, can be addressed. The system is being implemented in C++ using a class hierarchy based on the Image Understanding Environment (IUE)[7].

The current state of progress of MORSE is illustrated in figure 23. Class-based groupers have been implemented for surfaces of revolution (section 3.5), canal surfaces (section 3.6), and polyhedra (section 3.7). In each case the grouping is based solely on the constraints on the structure of the profiles for each class. Profiles for each class are extracted completely automatically. As is demonstrated in the figure, recognition proceeds by first recognising an object as belonging to one of the classes (for example a surface of revolution). Subsequently the object will be identified (for example as a particular vase).

[7] The IUE is an ARPA funded project to produce an object-oriented programming environment for vision research. A central object hierarchy in the IUE is the *spatial-object* which incorporates many of the descriptive requirements described in the previous sections. The IUE also has an extensive set of classes for object and image transformations which are a central issue in MORSE.

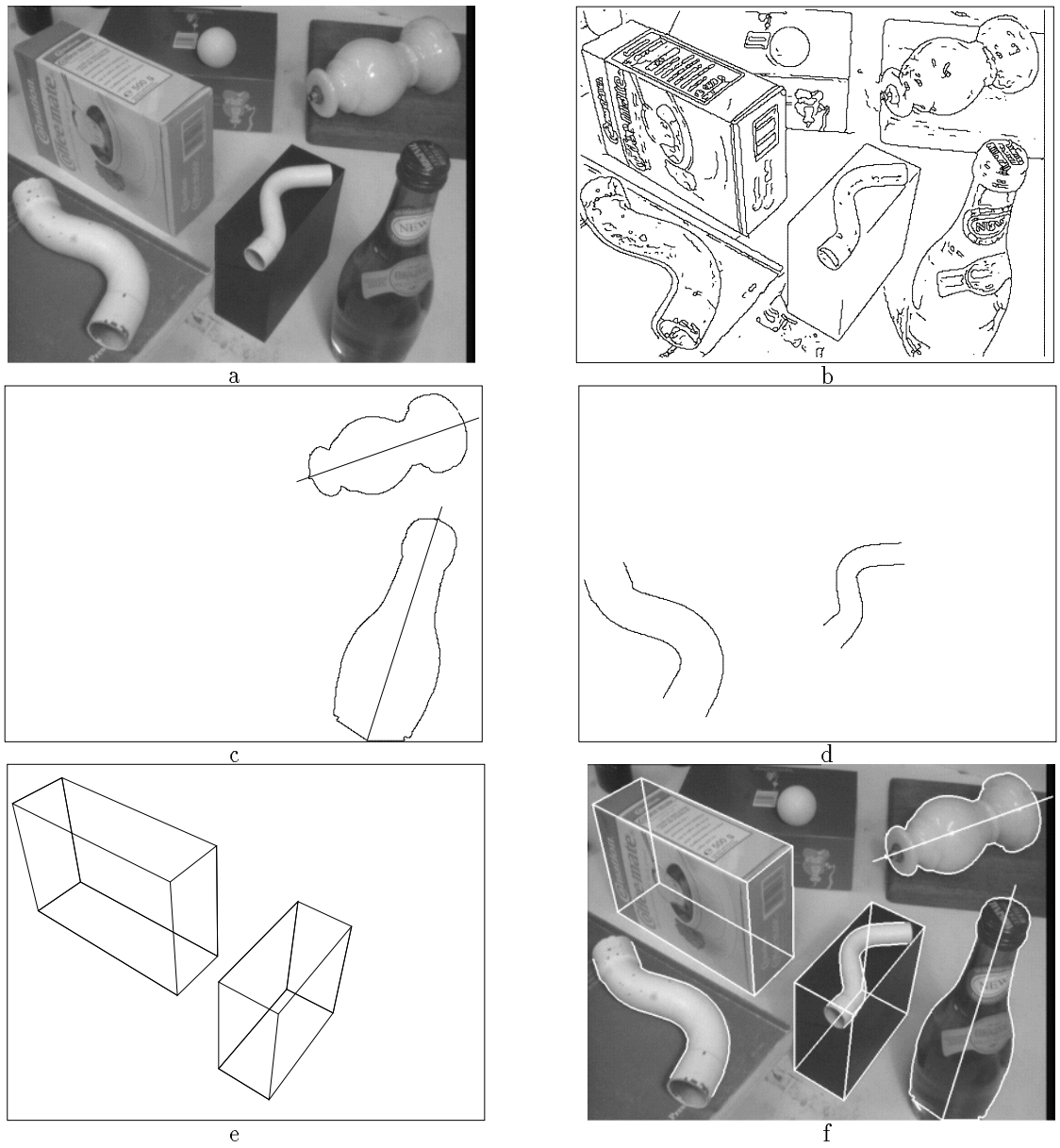


Figure 23: (a) Original image containing two surfaces of revolution, two canal surfaces, and two polyhedra. (b) The linked edges computed from (a). Profiles are extracted and grouped automatically from these linked edges by the class-based groupers. (c) Extracted surface of revolution profiles with axes. Note that gaps in the edgel chains have been repaired in the recovered profile. (d) Extracted canal profiles. (e) Extracted polyhedra outlines. (f) Extracted profiles superimposed on original image. All the correct instances of a class have been grouped, and no false instances grouped.

5 Discussion

5.1 Critique of the invariance approach

To conclude, it is useful to clarify many of the points just presented by responding to a number of major criticisms which can be made of the invariance approach to recognition. It will be instructive to employ these critical points as a benchmark of the progress in recognition which can be attributed to the invariant framework.

1. **The extreme nature of projective ambiguity:** Invariance concentrates on projective representations. In practice, perspective distortions in images are small and so can be ignored. Furthermore, the projective equivalence class is too large — a sphere and an ellipsoid are in the same class, as are a cube and a truncated pyramid. Thus, the recognition system can not distinguish between them.

   Response: First, we have demonstrated with the planar recognition system that a projective representation *is* sufficient for many practical examples. Second, although it is almost always the case that only projective structure can be recovered in a single uncalibrated image of an object, this does not mean that the recognition system is bound to projective ambiguities. For example, for certain classes, affine or similarity invariants can be measured in a perspective image, e.g., a structure repeated by translation (section 3.4). Euclidean consistency, section 4.1.2, can be used to reduce ambiguity from projective to similarity for an image of multiple objects.

2. **The exclusiveness of geometry:** Invariance at present concentrates on geometry to the exclusion of other important object properties that should be used in a recognition system such as: colour, texture (e.g., wood grain), surface markings (e.g., pictures or lettering on a can), and surface properties (e.g., metal vs dielectric).

   Response: Geometry very largely dominates object descriptions in the system sketched above. There is some way to go before colour, texture, surface markings or surface properties can stand on an equal footing with geometric information. These are, at present, measured in images relatively unreliably compared to geometry. Nevertheless, such cues fit into the proposed architecture. For example surface markings and texture can be used as additional invariant indexes (see section 5.2), and could certainly be used as additional measures during verification.

3. **The lack of abstract classification:** Invariance does not address the problem of classification, only identification. In a typical model-based system, the class to which an object belongs can be determined only by recognising it as a specific object. For example, an unknown object might be identified as a "1991 Red Mazda 323 Hatchback", which is a member of the class "car". This class membership is determined by subsequent reasoning: it cannot be directly identified as a car, as distinct from a fish, despite the differences between the two classes.

   Response: Abstract classification in its broadest sense presents severe conceptual and philosophical problems, which we have carefully avoided addressing. Until it is possible to address these problems concretely, by, for example, stating exactly what a program that distinguished between a general fish and a general bicycle would do, they will be difficult to solve. However, the architecture proposed contains a first step in this direction, by distinguishing between classes of object on the basis of the techniques required to construct representations from images. In particular, if a group of edge segments is classified as, for example, the profile of a rotationally symmetric object, techniques exist for confirming that classification (in this case, by determining that there is a projective equivalence, $\mathtt{T}$, on the profile, such that $\mathtt{T}^2 = \mathtt{I}$).

4. **The rigidity of exact geometry:** Geometry is not the appropriate language to represent objects such as clothes, plants and animals, which can articulate and deform. A deformable template, or even non-geometric descriptions, such as a set of colour histograms, may be much more effective in representing such objects.

Response: It is not yet clear what representations one would want to extract for objects that have no clearly defined geometry (for example, what aspects distinguish one shirt from another?). As a result, exact geometry is probably going to dominate recognition for some time to come. However, there is clearly a need for a hybrid of geometric invariance and statistics for certain classes of deformation; invariance would allow for change of viewpoint, while statistics would cover the deformation.

5. **Complex objects:** Invariants might be suitable for representing and recognising "simple" parts, such as surfaces of revolution or quadrics ("geons") but do not yet cover assembling these shapes into complex objects, such as telephones or aeroplanes.

   Response: Some of the 3D classes, for example, surfaces of revolution or canal surfaces, are "simple" — essentially, little more than plane curves. This does not affect their usefulness in representing a large number of real objects. Other classes of objects are more genuinely 3D objects — for example, repeated structures or polyhedra.

   Objects consisting of, for example, a hierarchy of parts, are not explicitly addressed in this approach. However, the seeds of a solution are present in the use of geometric relations between feature groups (intra-object invariants), such as those shown in figure 3, in forming joint-recognition hypotheses. One could advance the system architecture above to do the same thing, recognising parts individually and then using projective, or Euclidean information about object pose that results from the consistency checking, to determine whether components lie in such a way as to make up a composite object.

   Concerning segmentation, it may be the case that the grouping relations defined for each class provide a natural means of segmenting a complex outline into primitive volumetric parts. For example, when the harmonic homology on the profile ceases to apply this would indicate that the surface of revolution part had finished.

   Of course, there will be objects, e.g., potatoes, that can not be represented by a combination of the classes described here. However, such generic shapes are currently difficult to distinguish with any representation under the distortions of perspective imaging. Our view is that it is better to proceed with a set of classes which can support reliable recognition and establish a benchmark of performance for future systems to build on.

5.2 Avenues of future research

Indexing allows fast recognition of objects drawn from a diverse collection of classes: a range of specific techniques for recovering the projective invariants necessary for indexing various object classes has been displayed and demonstrated. These ideas have been integrated to produce a recognition system architecture that should be capable of handling large, diverse modelbases, and that addresses many of the concerns recently raised about indexing in recognition systems.

However, object recognition is not yet "solved." There are a range of avenues of research that promise exciting developments; we indicate a few topics most interesting to us:

- **The role of quasi-invariants:** Using quasi-invariants for indexing is a problem, because of the cost incurred if the "wrong" object is indexed by a quasi-invariant applied outside its domain of stability. Avoiding this requires complex hypothesis combination to ensure the "right" object is indexed. The benefit of quasi-invariants is the use of simpler feature groups at the start of the recognition process. However, simple feature groups are often not very discriminating. Instead, we propose that quasi-invariants should be used to *schedule* grouping. The quasi-invariants identified in this paper (e.g., the relation between sides of a surface of revolution profile), can be used to schedule promising groups for further growth.

- **Learning invariants:** invariant indexes are a good goal for a learning algorithm; an ideal algorithm would, given a large model base, determine by some offline process of generating views, the functions and image features most useful in indexing models effectively. Alternatively, invariants could be

extracted from a large number of real images of an object taken from varying viewpoints. The advantage of the latter is that the invariant descriptors would only involve features that could be reliably measured in images.

- **The use of texture and surface markings:** Clearly, texture and surface markings have a part to play in verification. However, surface markings, together with the profile of certain surface classes, can be used to generate further projective invariants. For example, by facilitating the backprojection of the markings onto the surface for that class. These marking invariants can augment indexes based on the object profile alone. For example, without surface markings, quadrics are projectively equivalent, but four points (markings) on a quadric surface have two projective invariants in space, which can be recovered from a single image.

- **Extensions to Grouping Computation:** In recent experiments with control for grouping features for rotationally symmetric objects and repeated structures, the idea of *synchrony* in edgel curve and line segment linking has emerged. For example in exploring the topological links along the occluding boundary of a rotationally symmetric object, it should be possible to use the constraints of the planar homology to control the linking sequence. In a complex scene with many possible edgel chain connections, these constraints will considerably reduce the number of feasible paths generated for symmetrical association. Once a single concavity is determined, the rest of the boundary can be recovered by a synchronised edge following algorithm. As new parts of the boundary are confirmed, the homology transform parameters can be iteratively refined.

  The same type of strategy can be followed for any geometric class based on symmetry or structural repetition. The constraints inherent in these classes can be extended right down to the edgel linking stage. Such an approach is currently being implemented.

Acknowledgements

We are grateful for discussions with Santanu Chaudhury, Peter Giblin, Richard Hartley, Jitendra Malik, Yael Moses, Sven Utcke and Luc Van Gool. Ellen Walker of RPI provided support and guidance for Jane Liu. Martin Armstrong and Paul Beardsley computed the affine invariants for the translational repeated structure. Financial support was provided by several agencies: ESPRIT Project 6448 'VIVA'; a NSF Young Investigator Award with matching funds from GE, Tektronix, Rockwell International and Eugene Rikel; NSF grant no. IRI-9209729; US Air Force Office of Scientific Research grant no. AFOSR-91-0361; General Electric; and The Newton Institute, Cambridge, under SERC grant GRG59981.

References

[1] Ayache, N. and Faugeras, O.D. "HYPER: A New Approach for the Recognition and Positioning of Two-Dimensional Objects," *IEEE Trans. PAMI,* Vol. 8, No. 1, p.44-54, January 1986.

[2] Ayache, N. and Faugeras, O.D. "Building a Consistent 3D Representation of a Mobile Robot Environment by Combining Multiple Stereo Views," *Proc. IJCAI*, p.808-810, 1987.

[3] Binford, T.O. "Inferring Surfaces from Images," *Artificial Intelligence,* Vol. 17, p.205-244, 1981.

[4] Binford, T.O. and Levitt, T.S. "Quasi-Invariants: Theory and Explanation," *Proc. DARPA IU Workshop*, p.819-829, 1993.

[5] Binford, T.O., Kapur D. and Mundy J.L., "The Relation between invariants and quasi-invariants", *Proc. Asian Conference in Computer Vision*, 1993.

[6] Bolles, R.C. and Horaud, R. "3DPO: A Three-Dimensional Part Orientation System," *Three Dimensional Vision,* Kanade, T. editor, Kluwer Academic Publishers, p.399-450, 1987.

[7] Borgefors, G. "Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm," *IEEE Trans. PAMI,* Vol. 10, No. 6, p.849-865, November 1988.

[8] Brooks, R.A. "Model-Based Three-Dimensional Interpretations of Two-Dimensional Images," *PAMI-5*, No. 2, March 1983.

[9] Burns, J.B., Weiss, R.S. and Riseman, E.M. "The Non-existence of General-case View-Invariants," in [45].

[10] Cass, T.A. "Polynomial-Time Object Recognition in the Presence of Clutter, Occlusion, and Uncertainty," *Proc. ECCV*, LNCS 588, Springer-Verlag, p.834-842, 1992.

[11] Clemens, D.T. and Jacobs, D.W. "Model Group Indexing for Recognition," *IEEE Trans. PAMI,* Vol. 13, No. 10, p.1007-1017, October 1991.

[12] Cox, I.J., Rehg, J.M. and Hingorani, S. "A Bayesian Multiple Hypothesis Approach to Contour Grouping," *Proc. ECCV*, LNCS 588, Springer-Verlag, p.72-77, 1992.

[13] Dhome, M., Lapreste, J.T., Rives, G. and Richetin, M. "Spatial Localization of Modelled Objects of Revolution in Monocular Perspective Vision," *Proc. ECCV*, LNCS 427, Springer-Verlag, p.475-485, 1990.

[14] Ettinger, G.J. "Large Hierarchical Object Recognition Using Libraries of Parameterized Model Sub-Parts," *Proc. CVPR88*, p.32-41, 1988.

[15] Faugeras, O.D. and Hebert, M. "The Representation, Recognition, and Locating of 3-D Objects," *International Journal of Robotics Research,* Vol. 5, No. 3, p.27-52, 1986.

[16] Faugeras, O.D., "What can be Seen in Three Dimensions with an Uncalibrated Stereo Rig?" *Proc. ECCV*, LNCS 588, Springer-Verlag, p.563-578, 1992.

[17] Faugeras, O.D., Luong, Q.T. and Maybank, S.J. "Camera Self-Calibration: Theory and Experiments", *Proc. ECCV*, LNCS 588, Springer-Verlag, 1992.

[18] Fawcett R., Zisserman A. and Brady J.M., "Extracting Structure from an Affine View of a 3D Point Set with One or Two Bilateral Symmetries", *Image and Vision Computing*, 12, 9, 615-622, 1994.

[19] Fisher, R.B. *From Surfaces to Objects: Computer Vision and Three Dimensional Scene Analysis,* John Wiley and Sons, 1989.

[20] Forsyth, D.A., Mundy, J.L., Zisserman, A.P. and Rothwell, C.A. "Recognising Curved Surfaces from their Outlines," *Proc. ECCV*, LNCS 588, Springer-Verlag, p.639-648, 1992.

[21] Forsyth, D.A. and Rothwell, C.A. "Representations of 3D objects that incorporate surface markings", in [46].

[22] Forsyth, D.A., "Recognizing Algebraic Surfaces from their Outlines," To appear *International J. of Computer Vision*, 1995.

[23] Giblin P.J. and Brassett S.A., "Local Symmetry of Plane Curves", *American Mathematical Monthly*, vol. 92, no. 10, pp. 689–707, 1985.

[24] Goad, C. "Special Purpose Automatic Programming for 3D Model-Based Vision," *Proc. DARPA IU Workshop*, p.371-381, 1983.

[25] Golub, G.H. and Van Loan, C.F., *Matrix Computations*, John Hopkins University Press, 1983.

[26] Gordon, G.,"Shape from Symmetry", *Proc. SPIE Conf. Intelligent Robots and Computer Vision VIII*, Philadelphia, 1989.

[27] Grimson, W.E.L. and Lozano-Pérez, T. "Localizing Overlapping Parts by Searching the Interpretation Tree," *IEEE Trans. PAMI,* Vol. 9, No. 4, p.469-482, July 1987.

[28] Grimson, W.E.L. *Object Recognition by Computer, The Role of Geometric Constraints,* MIT Press, 1990.

[29] Hartley, R.I., Gupta, R. and Chang, T. "Stereo from Uncalibrated Cameras," *Proc. CVPR92*, p.761-764, 1992.

[30] Hartley, R.I., "Euclidean Reconstruction from Uncalibrated Views", in [46].

[31] Huttenlocher, D.P. and Ullman, S. "Object Recognition Using Alignment," *Proc. ICCV1*, p.102-111, 1987.

[32] Koenderink, J.J., "What does the Occluding Contour tell us about Solid Shape?" *Perception*, Vol. 13, 1984.

[33] Kriegman, D.J. and Ponce, J. "On Recognizing and Positioning Curved 3-D Objects from Image Contours," *IEEE Trans. PAMI,* Vol. 12, No. 12, p.1127-1137, December 1990.

[34] Lamdan, Y., Schwartz, J.T. and Wolfson, H.J. "Object Recognition by Affine Invariant Matching," *Proc. CVPR88*, p.335-344, 1988.

[35] Liu J.S., Mundy J.L., Forsyth D.A., Zisserman A. and Rothwell C.A., "Efficient Recognition of Rotationally Symmetric Surfaces and Straight Homogeneous Generalized Cylinders", *CVPR*, 1993.

[36] Liu J.S., Mundy J.L. and Walker E.L., "Recognizing Arbitrary Objects from Multiple Projections," *Proc. Asian Conference in Computer Vision*, 1993.

[37] Lowe, D.G. *Perceptual Organization and Visual Recognition,* Kluwer Academic Publishers, 1985.

[38] Lowe, D.G. "The Viewpoint Consistency Constraint," *International Journal of Computer Vision,* Vol. 1, No. 1, p.57-72, 1987.

[39] Mitsumoto H., Tamura S., Okazaki K., Kajimi N. and Fukui Y., "3D Reconstruction Using Mirror Images Based on a Plane Symmetry Recovery Method", *PAMI*, 14, 9, 941-945, 1992.

[40] Mohr, R., Morin, L. and Grosso, E., "Relative Positioning with Uncalibrated Cameras,", in [45].

[41] Moons, T., Van Gool, L., Van Diest, M. and Pauwels, E. "Affine structure from perspective images pairs", in [46].

[42] Moses, Y. and Ullman, S. "Limitations of Non Model-Based Recognition Systems," *Proc. ECCV*, LNCS 588, Springer-Verlag, p.820-828, 1992.

[43] Mukherjee D.P., Zisserman A. and Brady J.M., 'Shape from symmetry—detecting and exploiting symmetry in affine images'. To appear, *Proc. Royal Soc.*, 1995.

[44] Mundy, J.L. and Heller, A.J. "The Evolution and Testing of a Model-Based Object Recognition System," *Proc. ICCV3*, p.268-282, 1990.

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

[46] Mundy J.L., Zisserman A., and Forsyth D.A., editors, *Applications of Invariance in Computer Vision*, LNCS 825, Springer-Verlag, 1994.

[47] Mundy, J.L. and Zisserman, A. "Repeated structures: image correspondence constraints and 3D structure recovery," in [46].

[48] Murray, D.W. "Model-Based Recognition using 3D Structure from Motion," *Image and Vision Computing,* Vol. 5, p.85-90, 1987.

[49] Murray, D.W. and Cook, D.B. "Using the Orientation of Fragmentary 3D Edge Segments for Polyhedral Object Recognition," *IJCV*, No. 2, p.153-169, 1988.

[50] Nielsen, L. "Automated Guidance of Vehicles using Vision and Projective Invariant Marking," *Automatica,* Vol. 24, p.135-148, 1988.

[51] Pillow N., Utcke S. and Zisserman A. "Viewpoint-Invariant Representation of Generalized Cylinders Using the Symmetry Set", To appear *Image and Vision Computing*, 1995.

[52] Pollard, S.B, Pridmore, T.P, Porrill, J., Mayhew, J.E.W. and Frisby, J.P. "Geometrical Modeling from Multiple Stereo Views," *International Journal of Robotics Research,* Vol. 8, No. 4, p.132-138, 1989.

[53] Ponce, J. "Invariant properties of straight homogeneous generalised cylinders," *IEEE PAMI*, **11**, 9, 951-965, 1989.

[54] Reid, I. "Recognising Parameterized Models from Range Data," D.Phil. Thesis, Department of Engineering Science, University of Oxford, Oxford, 1991.

[55] Reiss, T.H., *Recognizing Planar Objects Using Invariant Image Features*, LNCS 676, Springer Verlag, 1993.

[56] Roberts, L.G. "Machine Perception of Three-Dimensional Solids," In *Optical and Electro-optical Information Processing,* editor Tippett, J. *et al.*, MIT Press, 1965,

[57] Rothwell, C.A., Zisserman, A., Forsyth, D.A. and Mundy, J.L., "Canonical Frames for Planar Object Recognition" *Proc. ECCV*, LNCS 588, Springer-Verlag, p.757-772, 1992.

[58] Rothwell, C.A., Zisserman, A., Mundy, J.L. and Forsyth, D.A. "Efficient Model Library Access by Projectively Invariant Indexing Functions", *Proc. CVPR92*, p.109-114, 1992.

[59] Rothwell, C.A., Forsyth, D.A., Zisserman, A. and Mundy, J.L. "Extracting Projective Structure from Single Perspective Views of 3D Point Sets", *ICCV*, 573-582, 1993.

[60] Rothwell, C.A. *Recognition Using Projective Invariance*, D.Phil. Thesis, Department of Engineering Science, University of Oxford, 1993. To appear OUP.

[61] Rycroft J.E. *A Geometrical Investigation into the Projections of Surfaces and Space Curves*, Ph. D. Thesis, University of Liverpool, 1992.

[62] Sha'ashua, A. and Ullman, S. "Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network," *Proc. ICCV2*, p.321-327, 1988.

[63] Sparr, G., "Notes on geometric invariants in vision", *Lund Research Report*, 1993.

[64] Springer, C.E., *Geometry and Analysis of Projective Spaces,* Freeman, 1964.

[65] Stewman, J. and Bowyer, K.W. "Creating the Perspective Projection Aspect Graph of Polyhedral Objects," *Proc. ICCV2*, 1988.

[66] Stockman, G. "Object Recognition and Localization via Pose Clustering," *Computer Vision Graphics and Image Processing,* Vol. 40, p.361-387, 1987.

[67] Sugihara, K. *Machine interpretation of Line Drawings,* MIT Press, 1986.

[68] Thompson, D.W. and Mundy, J.L. "Three-Dimensional Model Matching from an Unconstrained Viewpoint," *Proc. ICRA*, p.208-220, 1987.

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

[70] Van Gool, L. Kempenaers, P. and Oosterlinck, A. "Recognition and Semi-Differential Invariants," *Proc. CVPR91*, p.454-460, 1991.

[71] Weiss, I. "Geometric Invariants and Object Recognition," *IJCV*, 10, 3, 1993.

[72] Zerroug M. and Nevatia R., "Using invariance and quasi-invariance for the segmentation and recovery of curved objects", in [46].

[73] Zisserman A., Forsyth D.A., Mundy J.L. and Rothwell C. "Recognizing General Curved Objects Efficiently", in [45].