An Experimental Comparison of Appearance Models and Geometric Class Models Represented by Invariants

J. Mundy, A. Liu
General Electric Corporate
Research and Development
Schenectady, NY, USA

N. Pillow, A. Zisserman, S. Utcke, S.Abdallah
Oxford University,
Oxford, UK

S. Nayer
Dept. of Computer Science
Columbia University
NY, USA

C. Rothwell
INRIA
Sophia Antipolis, France

Abstract

This paper describes an experimental investigation of the recognition performance of two approaches to the representation of objects for recognition. The first representation, generally known as appearance modelling, describes an object by a set of images. The image set is acquired for a range of views and illumination conditions which are expected to be encountered in subsequent recognition. This image database provides a description of the object. Recognition is carried out by constructing an eigenvector space to compute efficiently the distance between a new image and any image in the database. The second representation is a geometric description based on the projected boundary of an object. General object classes such as planar objects, surfaces of revolution and repeated structures support the construction of invariant descriptions and invariant index functions for recognition.

In this paper we present a preliminary investigation of the relative performance of the two approaches. Two objects, a planar object and a rotationally symmetric object are modelled using both approaches. In the experiments, each object is intentionally occluded by an unmodelled distractor for a range of viewpoints. The resulting images are submitted to two separate recognition systems. Appearance-based recognition is carried out by SLAM and recognition of invariant geometric classes by MORSE/LEWIS.

1 Introduction

Over the last few years, there has been increasing interest in object recognition based on a set of images of an object. This representation of an object is called an *appearance* model. Over the same period, a number of recognition systems have been implemented which employ a geometric description of an object. These geometric descriptions are based on general object classes, such as planar structures or surfaces of revolution. Each approach has strengths and limitations which can complement each other to form a more competent overall recognition system.

In this paper, we present some initial results of experiments to characterize the performance of implemented recognition systems for each approach. It is emphasized that these results are very preliminary, but do represent an attempt directly to compare the two recognition processes under identical conditions. This sort of comparison is badly needed to advance our understanding of object recognition, but it is usually difficult to obtain such data due to the state of implementation of most research software. The results of this investigation highlight a number of strengths and weaknesses of each approach and consequently the type of shape representation that is appropriate for model based object recognition.

The appearance (SLAM) and geometry based systems (LEWIS/MORSE) are introduced in section 1.1. The focus of this paper is to provide the results of a first set of experiments using these systems on the same set of images and test objects. We chose to focus on the performance of each system with respect to camera viewpoint and in the presence of occlusion. The performance of recognition under varying illumination is certainly also of great interest but was not considered here.

At this stage, we haven't provided the same overall library of objects to each system. Ultimately, a rigorous comparison should be based on installing identical object instances in the model library of each system. For the geometric invariant recognition systems there are 8 planar objects and 21 Surfaces of Revolution (SOR's). For SLAM, the object library consists of four objects: two objects we have selected for testing, one SOR and one planar object, and two other objects. Even this limited test has produced a number of interesting results and the experiments establish many issues for further investigation.

Before considering the experimental results, details of each recognition approach are reviewed in sections 2 and 3. The two systems are compared on the same test images in section 4.

1.1 Background to recognition systems

1.1.1 Appearance-Based — SLAM

We are fortunate in the experiments reported here, since the Software Library for Appearance Modelling (SLAM) system has been developed to be portable and has been disseminated to a large number of research groups. The SLAM software provides tools for acquiring new appearance models and for setting up a recognition program using the resulting image library. SLAM is implemented in C++ and is provided with a detailed set of manual pages and a set of example images for testing the use of the libraries.

1.1.2 Geometry Based — LEWIS/MORSE

Two systems for object recognition based on geometric invariants are also available. These systems have been implemented in a number of versions and are now reasonably mature and with acquisition tools to support the insertion of new object models. The first system, called LEWIS [13], was originally developed by Charlie Rothwell for his thesis at Oxford. A second generation of LEWIS was implemented in C++ by Bill Hoffman and Chien-ming Huang at GE-CRD for applications to aerial reconnaissance. This C++ version was re-implemented and improved by Rothwell to provide a standard for comparison with new recognition approaches using invariants.

A second system for the recognition of surfaces of revolution (SORs), MORSE, has been implemented using the same basic C++ libraries as LEWIS by Nic Pillow and Sven Utcke at Oxford. This system is a re-implementation and improvement of the original SOR recognition system by Jane Liu and Sven Utcke. Both the LEWIS and SOR system have tools for acquiring an object model from an image by selecting a set of features corresponding to the outline boundary of the object. Each system uses an identical representation for the model library and for indexing to retrieve candidate model hypotheses. In the experiments here, we are using the LEWIS and SOR systems in tandem to provide an overall recognition for both planar and SOR object classes in a given image.

2 Appearance models for recognition

The representation of an object by its image appearance is an empirical model which is defined for the range of viewing conditions under which the object is to be recognized[10]. The assumption is that if the intensity distribution in a new image is *near* a stored image of some object then the image contains the object. The appearance model makes no commitment to constraints which might be embodied in a class such as object shape or surface texture. Therefore a recognition system based on appearance can acquire descriptions of any type of object. The only requirement is that a sufficient number of views of the object have to be acquired to provide an image close in appearance to any image in which the object is to be recognized.

An appearance model can be generalized by interpolating between the acquired views. A new view of an object can be generated by assuming that an object induces a manifold in the space of image pixel measurements. For example, suppose that a series of images of an object are collected for a sample of two dimensional translations on the ground plane. It is assumed that views of the object at other translation positions can be closely approximated by interpolating the image intensities of images in the neighborhood of a given position.

A full appearance model requires six degrees of freedom to account for object pose variation and two parameters to account for illumination direction. Then a sample of images is collected for an object to sample this eight-dimensional manifold in the high dimensional space of image pixels. In the current experiments, an object is represented by a 128×128 pixel array so the appearance manifold is embedded in a space of 16384 dimensions.

The efficiency of computing distances in such a high dimensional space can be vastly

improved by the use of principal components analysis [3] (PCA). The use of PCA and appearance models has been used quite effectively for face recognition [1, 2, 6, 11]. PCA captures the variation in image data by projecting the image on to a low dimensional subspace of eign-vectors which are constructed from the covariance matrix of the image intensities. Since image intensities are typically correlated, it possible to accurately represent a full 16384 vector in 10 or 20 dimensional eigenvector space.

The eigenvectors are computed by accumulating the covariance matrix, C_{ij} of image vectors over the database of stored object appearances. That is,

$$C_{ij} = \sum_k (I_i(k) - \mu_i)(I_j(k) - \mu_j)$$

where $I_i(k)$ is the intensity of pixel i for image k in the database. μ_i is the mean value of pixel i over the set of stored images. The eigenvectors of C_{ij} define a linear transformation of the image data onto a vector subspace. An eigenvector can be dropped if its contribution to the overall variance of the image data is small. For images with smooth patches of relatively small image variation, the variance over the data base can be captured with only 10 or 20 of the original 16384 dimensions.

The SLAM software used in these experiments carries out the following specific steps.

1. Image normalization. A bounding box is constructed around the object by thresholding the object from the background. A standard image resolution, e.g. 128×128 is used to sample the bounding box and the intensity distribution is histogram equalized to provide a standard intensity range.

2. Each resulting image vector is added to the image data base with a label corresponding to the object class.

3. A covariance matrix is computed for the database and a suitable vector subspace defined which accounts for the variance in the database. In the current experiments, 10 eigenvectors define the subspace.

4. The manifolds for each object class are approximated with a spline. In the current experiments, the appearance variation is limited to one dimension of object rotation, so the manifold is a spline curve.

5. The fitted spline curve or surface is sampled to provide an adequate coverage of the required view conditions. For example, if for a rotation interval of 5^o the manifold can be sampled to 1^o to provide detailed coverage.

6. To recognize an object in a new image instance, the image vector is dotted with each of the eigenvectors of the database to construct eigenvalue coordinates. The closest stored manifold sample point in the database is located by a binary search strategy. The object is then assigned the class label of that closest manifold.

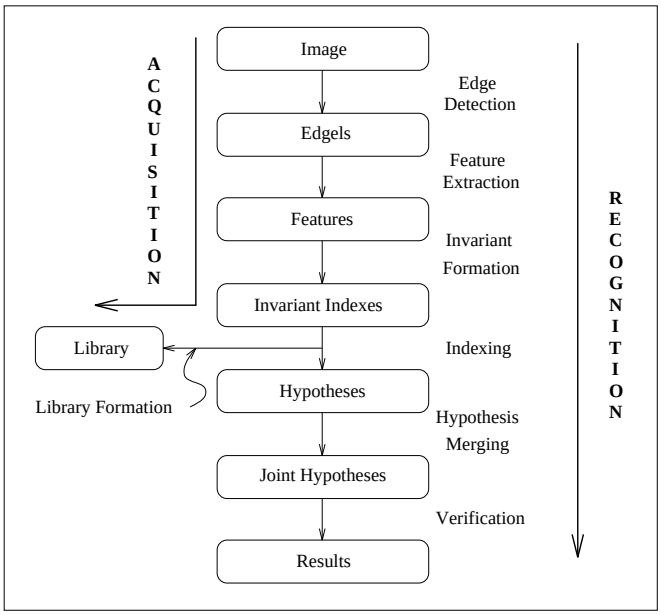


Figure 1: The recognition system has a single grey scale image as input and the outputs are verified hypotheses with associated confidence values. Many of the processes are shared by the acquisition and the recognition paths. The recognition system is similar to previous systems [5] in all but the indexing and hypothesis merging stages.

3 Geometry-based recognition

There are two recognition systems that are used. The first, LEWIS, is for planar object recognition. This system uses indices based on plane projective invariants. The second, MORSE, is for 3D curved object recognition.

3.1 LEWIS

Here we summarise the main features of the LEWIS planar object recognition system. The shape representation for the planar objects is *projective*, rather than Euclidean, i.e. the equivalence class of each object is a projective transformation. A projective representation of shape 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 plane projective 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 stages of recognition are shown in figure 1. In more detail:

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. These are applicable to image curves that are "algebraic" (lines, conics). For non-convex smooth

curve segments canonical frame invariants [9] 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 ones such as 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.

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.

Once sets of grouped features, $\boldsymbol{f}$, 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 $\boldsymbol{M}(\boldsymbol{f})$. 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.

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.

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*.

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.

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.

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 being 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.

A detailed description of this system appears in [13].

3.2 MORSE

This is a recognition system for 3D objects. The system is organised around a number of geometrically defined object *classes*. The classes include surfaces of revolution, canal surfaces (pipes) and polyhedra. These three classes cover a large number of manufactured objects. A class provides two functions. First, it provides a grouping relationship in the image. The geometric class defined in 3D *induces* relationships in the image which must hold between points on the image outline (the perspective projection of the object). The resulting image constraints enable both identification and grouping of image features belonging to objects of that class. Second, the class also supports the computation of 3D invariant descriptions including symmetry axes, canonical coordinate frames and projective signatures. Both grouping and invariant formation are viewpoint invariant, and proceed with no information on object pose.

Recognition consists of two major processes. The first is recognising an object as belonging to one of the classes. This is achieved by grouping or feature organization directed by the constraints provided by a particular geometric class. The second stage is identification which is achieved by using invariant indices derived for specific instances of the class. This differs from LEWIS where recognition is targetted directly at particular objects, not first at a class of objects.

For example, if a scene contains several SORs, the grouper first identifies curve pairings that could have arisen from a SOR (they satisfy a conjugate symmetry constraint, see section 3.2.1), and then groups those arising from the same SOR. Such identification and

grouping is possible because the 2D image curve which results from imaging the 3D class is tightly constrained. In turn these constraints can be used during grouping to test the validity of the class assumption. Subsequently, the organized SOR boundaries can be used to provide invariant indices constructed from portions of the boundary (see section 3.2.3).

Indexing the model library via the invariants generates recognition hypotheses for particular *models* of that class. For example, if an SOR profile (image outline) had been grouped the SOR invariants might index particular models, e.g. vase #2 or bottle #4, and recognition hypotheses for these models would then be verified by projecting the specific model boundary into the scene.

In the test examples of this paper the 3D object is of the SOR class. So in the following we describe in more detail the grouping and invariant index formation for this class. Full details of the MORSE system are given in [15].

3.2.1 SOR grouping

Here we describe in some detail the grouping mechanism for a SOR. For each SOR in image, the outcome of the grouping is an identified axis and pairs of bitangents (one half of each pair on either side of the axis), and repaired profile contours around the object, as illustrated in figure 15.

The profile of a surface of revolution can be separated into two 'sides' by the projected symmetry axis. The two sides are tightly constrained — they are related by a particular four degree of freedom plane projective transformation — a planar harmonic homology [8]. This relationship is exact. The transformation is represented by a non-singular 3×3 matrix $\mathtt{T}$, where $\mathtt{T}^2 = \mathtt{I}$ [14]. Two pairs of point correspondences determine $\mathtt{T}$. This transformation is fundamental to processing profiles of SORs: the line of fixed points of $\mathtt{T}$ is the imaged symmetry axis; $\mathtt{T}$ provides point to point correspondence between the sides of the profile; this disambiguates the matching of bitangents used to form the invariants; and finally, $\mathtt{T}$ can be used to *repair* missing profile portions, filling in gaps by transforming over points from the other side of the profile.

Based on these properties, grouping for this class can be carried out by associating curves which are projectively equivalent, and then testing if the transformation between two projectively related curves is a harmonic homology. If it is not then the associated curves can be ruled out as members of this class. This is simply tested by checking if $\mathtt{T}^2 = \mathtt{I}$.

Provided the field of view is not too large, $\mathtt{T}$ can be approximated by a three degree of freedom affine transformation (an involution) [7]. This is an excellent approximation.

3.2.2 Grouping Mechanism

Grouping involves determining which curves could have arisen from a surface of revolution, and repairing, where possible, missing curve segments due to occlusion and the usual problems of segmentation. In particular profile curves are often broken into a number of portions. Since the image may well contain several surfaces of revolution the grouping stage also partitions the profile curves into those arising from different surfaces. Grouping proceeds in three stages:

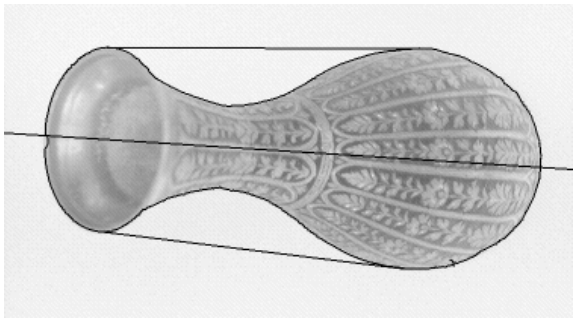


Figure 2: A grouped SOR. The original image is shown with the superposition of the profile, bitangent pair — just one in this case — and the computed axis.

(i) Associate conjugate curve fragments The two sides of the profile are affine related, and consequently corresponding profile curves have the same affine invariants. Associating conjugate (corresponding) curve fragments is then a matter of matching curve segments with the same affine invariants. Here area of profile concavities is used as the (relative) affine invariant.

(ii) Select profile fragment pairings A number of these putative matches can then be eliminated because the affine transformation between the curves is not an involution.

The transformation is determined using a type of *affine snake*. This is computed in two stages: first, an approximate solution is determined by matching a number of distinguished points (such as bitangent contact points); second, the approximate solution is used to transform a number of sample points from one profile side to the other. The squared distance between the transformed sample points and the other profile side is then minimized numerically over the three parameters of the transformation. Effectively this uses one side of the profile to define a snake, and then determines the affine transformation (constrained to have the above properties) which most closely aligns it with the other side.

(iii) Group profile fragments It then remains to group curve fragments which may have arisen from the same profile curve. Grouping is based on the similarity of the three parameters of the affine transformation (i.e. the symmetry axis and correspondence direction). Using these parameters the paired SOR concavities are partitioned into sets, and the associated profile curve fragments are joined using existing local edgel chain topology and smooth curve continuation.

Further details are given in [15].

3.2.3 SOR invariants

It has been shown previously [7] that intersections of corresponding pairs of bitangents on an SOR profile are projections of points on the axis of the object. Constructing the set of such bitangent intersection points and computing their cross-ratio provided a means of object identification. However that required finding four distinct bitangent pairs which limits the model SORs to ones with a suitably rich geometry.

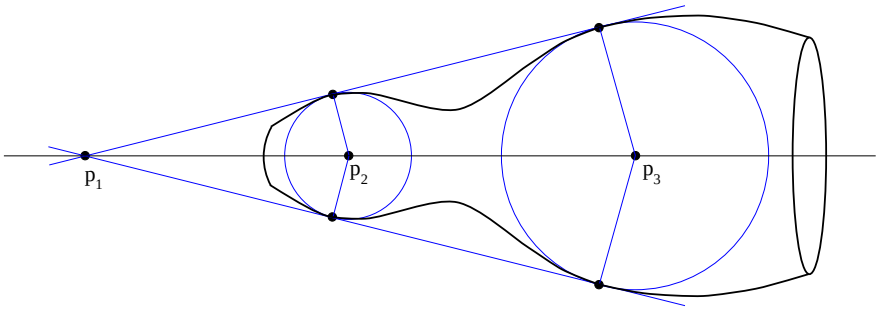


Figure 3: The construction of p_2 and p_3, two of the distinguished points on the axis for computing the cross-ratio.

The invariant used here is again a cross-ratio of distinguished points on the SOR axis, but these four points are determined from a single bitangent pair (e.g. from a single concavity) in an image. This invariant is a *quasi-invariant* — it is invariant to an excellent approximation under perspective imaging. Its construction requires the image aspect ratio to be correct.

The distinguished points are constructed as follows: the first, p_1, is the intersection of the bitangent lines as before. The 2nd and 3rd points are constructed as in figure 3. Each pair of distinguished profile points — in this case bitangent points — lie on the projection of a single cross-section of the SOR: a circle of points on its surface. A sphere which is tangent to the SOR at this circle of points will have its centre at a particular point on the axis.

Under orthographic projection, the projection of this sphere is a circle, which is tangent to the profile at the bitangent contact points and whose centre lies on the axis at the projection of the centre of the sphere's [12]. For perspective projection, a sphere does not project to a circle, but under reasonable imaging conditions the distortion is slight, and the construction, although only quasi-invariant, holds well.

The last cross-ratio point, p_4, is constructed as shown in figure 4. The cast tangent rays are the image of a virtual cone, with p_1 being the image of its vertex. The cone is tangent to the SOR at a circular cross-section, and in the image, then, the cast tangent points are images of the two points on this cross-section that form part of the SOR profile. A distinguished point on the axis is generated as above as the centre of a circle bitangent to these points.

The limitation of this invariant construction is that the cast tangent point within the concavity must be visible in the image, ie not self-occluded due to viewpoint. But as can be seen from figure 15, the technique works over a large part of the viewing sphere, and as we get an invariant from each bitangent pair (concavity), it is much more robust with respect to (self-) occlusion of concavities than previously published work (e.g. [7]).

4 Experiments

Two objects were selected for the experiments, one SOR and one planar shape (a floppy diskette).

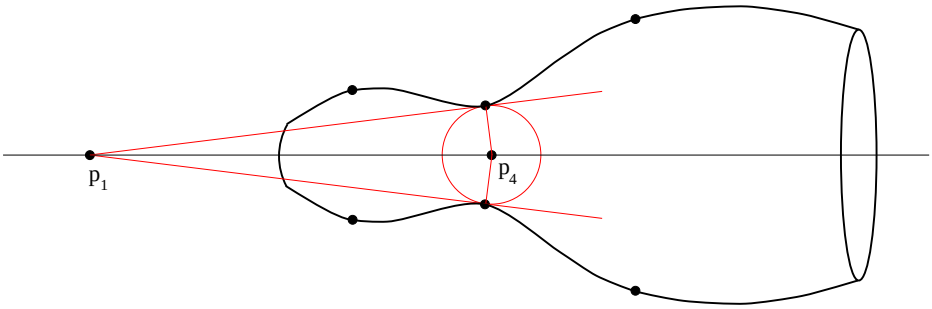


Figure 4: The construction of the 4th distinguished axis point from a single SOR bitangent pair. On each side of the profile, a ray is cast from p_1 so as to be tangent to the interior of the concavity; the point of tangency is distinguished on the profile. As before, the corresponding axis point is found as the centre of the circle bitangent to the profile at this pair of tangent points.

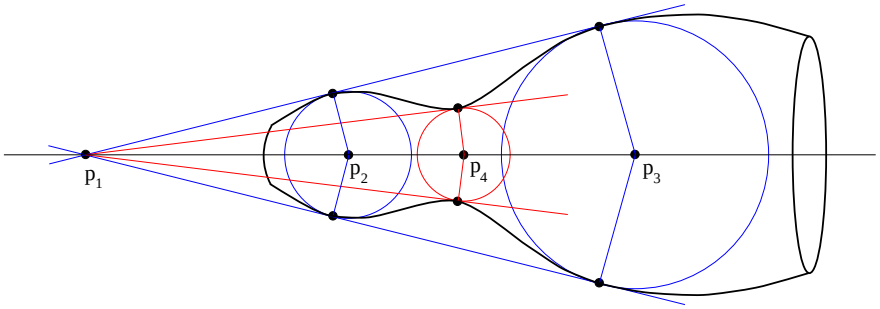


Figure 5: The full construction used to generate the cross-ratio points.

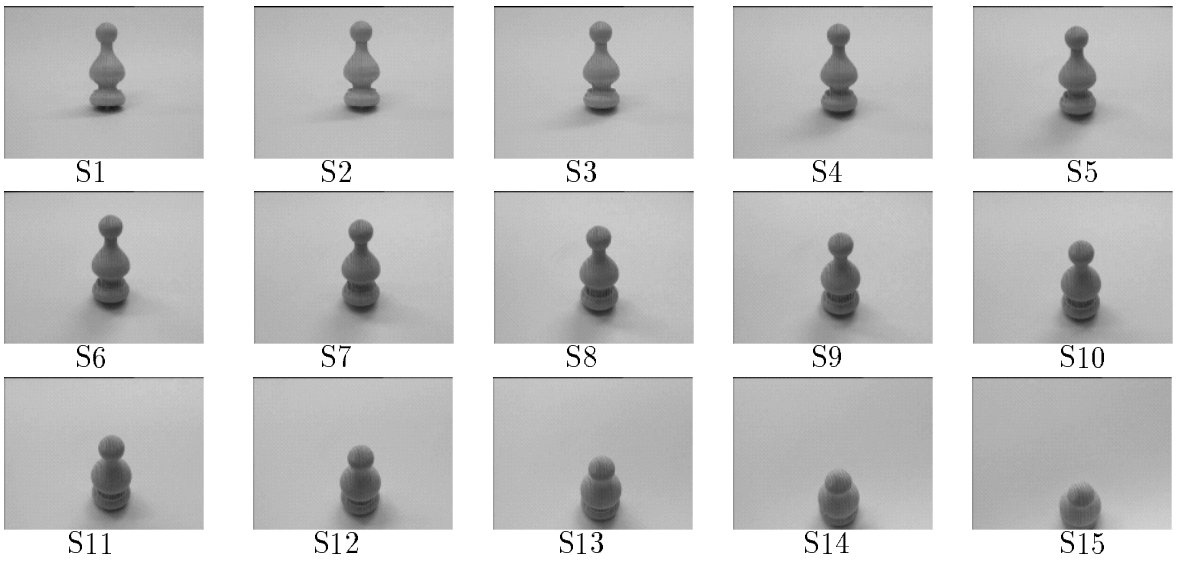


Figure 6: The acquisition images of the SOR.

Acquisition Images There are two sets of acquisition images. For the SOR, the slant of the symmetry axis was varied in 5^o steps, keeping the SOR at the horizontal centre of the image. The 15 acquisition images are shown in figure 6. The planar object was imaged in an oblique view and the orientation about the normal to the object's plane is varied in 10^o steps to acquire its appearance model. The 36 acquisition images are shown in figure 7. The illumination in each case was from a diffuse source and maintained constant over the image collection.

Test Images There are three sets of test images. Two of the sets are of one of the objects together with an unmodelled distractor object, placed to provide increasing occlusion. One set is for the SOR and distractor, the other for the disk and distractor. In each case there are 9 images corresponding to three degrees of occlusion, varying from no occlusion, slight occlusion to heavy occlusion; and for each degree of occlusion three viewpoints are imaged. The images are shown in figures 8 and 9 for the SOR and disk respectively.

The third set of test images contains both the SOR and disk. These are shown in figure 10.

In all the following experiments the systems are not optimised and are used completely automatically, i.e. the same parameters are used on all images.

4.1 Appearance-based system

Each set of acquisition images was normalized as described in section 2. Then the full set of images was used to construct the library covariance matrix and eigenspace. The manifolds for each class are shown in Figure 11. The manifold is displayed by selecting the three largest eigenvalues to form a three dimensional subspace. In the case of these experiments, the manifolds are curves since only one parameter is varied in generating the appearance model.

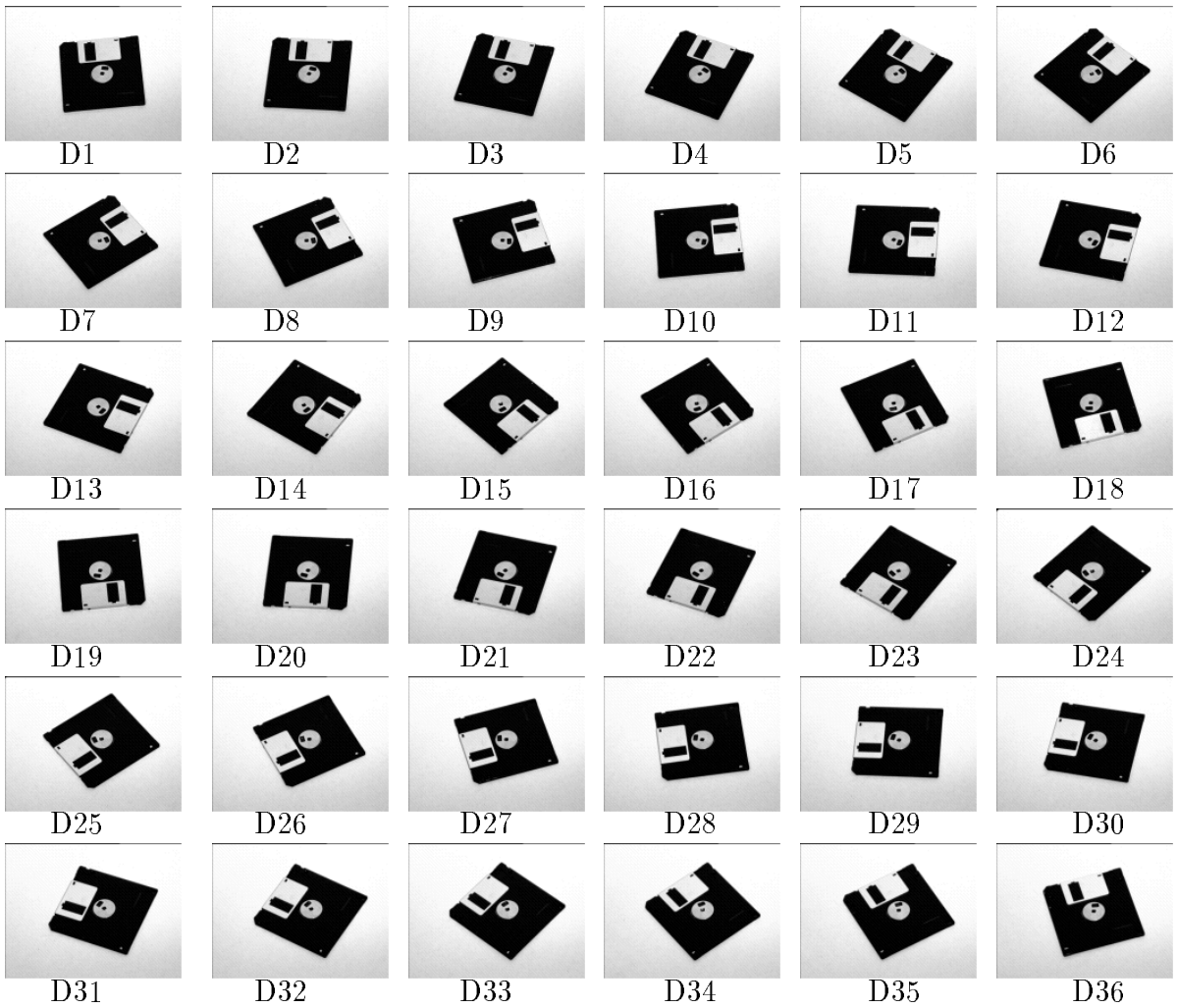


Figure 7: The acquisition images for the planar object.

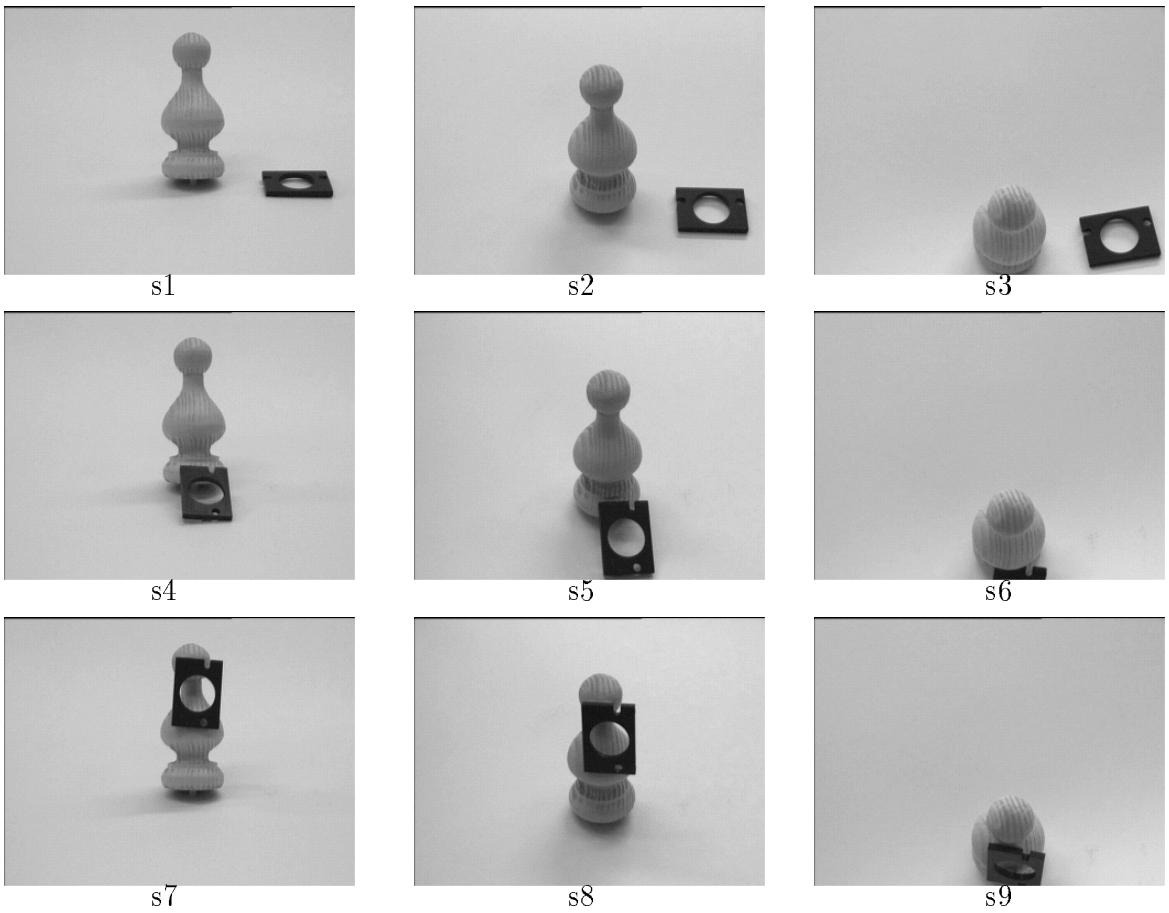


Figure 8: The SOR test images with clutter.

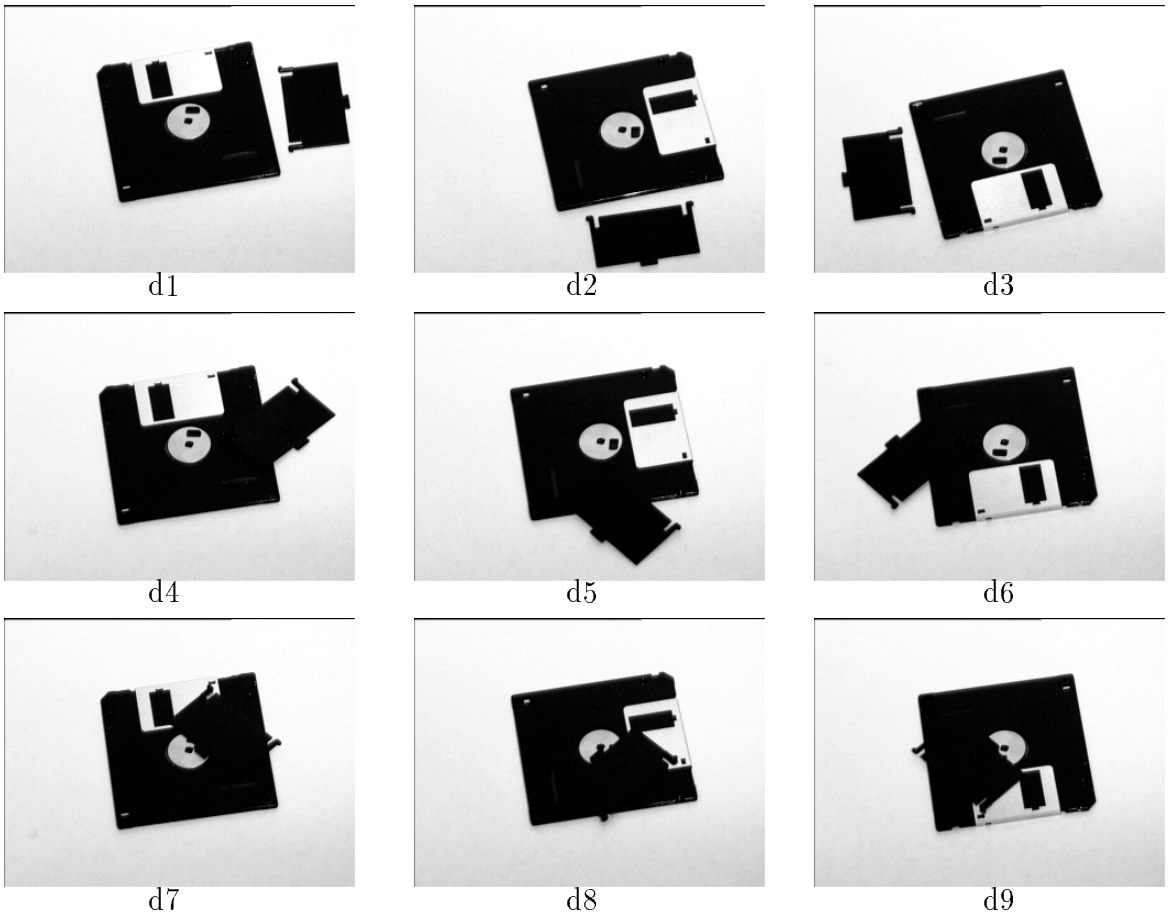


Figure 9: The planar test images with clutter.

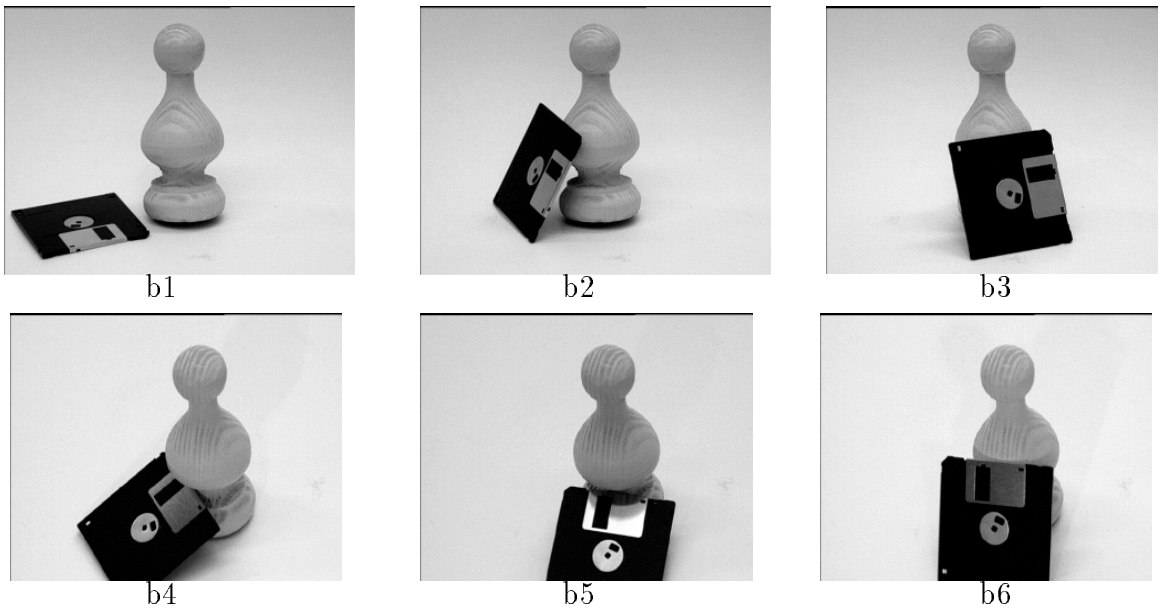
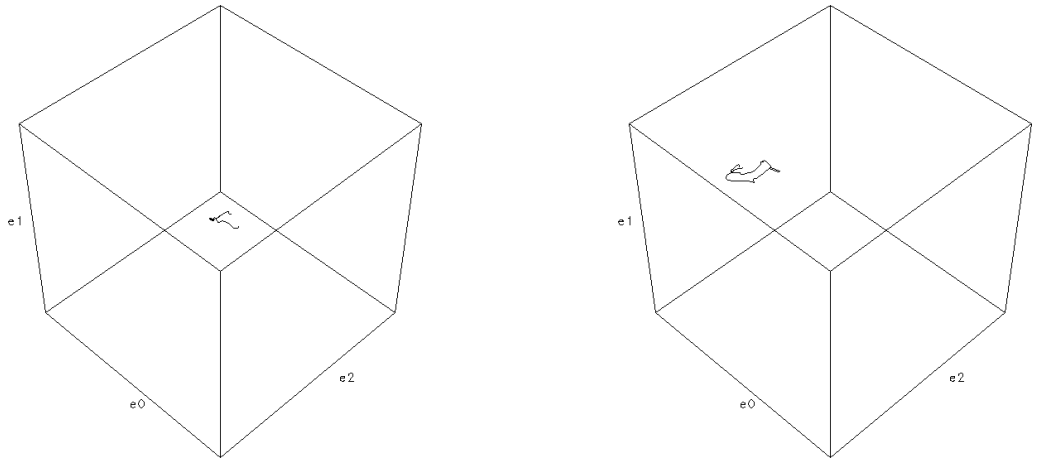


Figure 10: Test images of the SOR and disk.



a: SOR b: disk

Figure 11: The one dimensional manifolds for the two objects in the appearance modelbase: (a) SOR, (b) disk. The manifolds are parameterized by the rotation applied during image acquisition.

The actual eigenspace used for recognition has ten dimensions. That is, the set of eigenvectors with the ten largest eignvalues provides a good approximation to each of the images in the image library.

Recognition proceeds by normalizing a new image and then decomposing the image vector into the ten eigenvectors defined for the modelbase. The resulting ten eigenvalues define a point in eigenspace. The closest stored point, within a tolerance threshold, is retrieved and associated with the stored label to classify the object. For the recognition experiments there are *four* objects in the appearance model library: the SOR and disk, and two other objects provided with the SLAM library: an aspirin case and a car.

Recognition Results Matching is against 4 objects. All 9 SOR test images were correctly recognised, but only 6 of the 9 disk test images were correctly recognised. Results for the disk images are given in table 1.

Only one of the 6 test images containing both the SOR and the disk were correctly recognized — the closest image was an SOR image, the other 5 had closest images corresponding to aspirin and car images. These results are in table 2. The second closest image were also incorrect for 5 of the 6 tests. Again only b5 had as second closest image an SOR image. Clearly performance is poor when images contain two objects from the model base.

4.2 Geometry-based system

4.2.1 LEWIS

The aim here is to recognize the disk from the test images. There are 7 planar objects in the LEWIS library; these are shown in figure 12. Recognition proceeds against all the models in this library together with the test disk (i.e. 8 objects in total); the disk model

Image Id	SLAM-Classified Object	Nearest Distance
d1	ASPIRIN	0.420
d2	ASPIRIN	0.197
d3	ASPIRIN	0.278
d4	disk	0.032
d5	disk	0.453
d6	disk	0.154
d7	disk	0.929
d8	disk	0.253
d9	disk	0.168

Table 1: Appearance-based classifications for images in figure 9.

Image Id	SLAM-Classified Object	Nearest Distance
b1	CAR	1.022
b2	CAR	0.914
b3	ASPIRIN	0.129
b4	ASPIRIN	1.118
b5	SOR	0.165
b6	ASPIRIN	1.191

Table 2: Appearance-based classifications for images of the SOR and disk in figure 10.

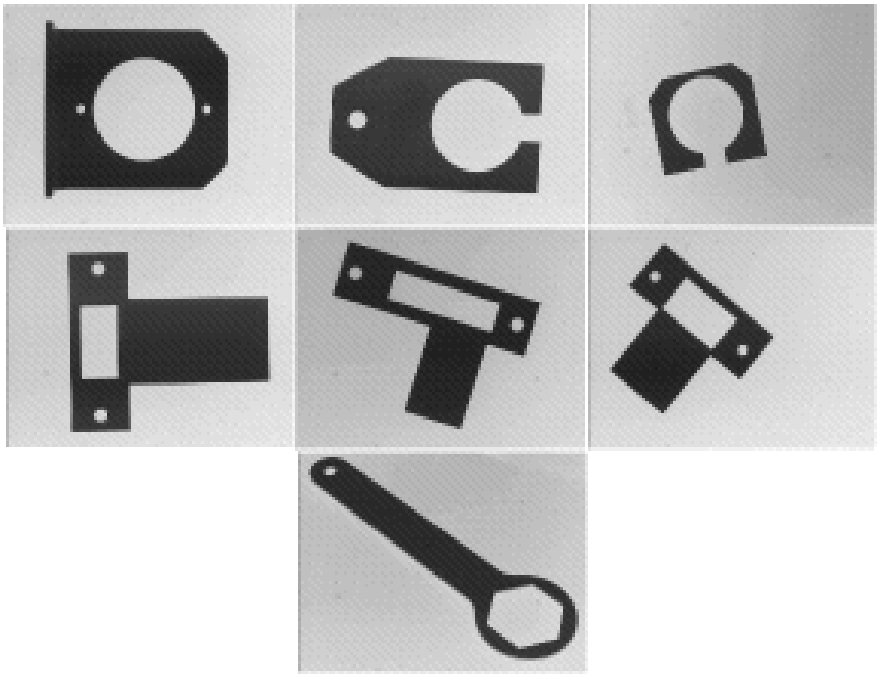


Figure 12: The images contained in the Lewis model-base.

is acquired from three images of the disk. The recognized disks from both image sets are shown in figure 13.

Disk test images Of the nine images (figure 9), the disk was correctly recognized in seven with no false positives. The two failures, b3 and b5, were caused by relative translation of the disk itself and the disk case: all the invariants are based on sets of two lines and one conic, the latter always being extracted from the central circular component of the disk which can move relative to the case. This movement causes the invariants to deviate from their model values.

SOR and disk test images Just two of the six, b2 and b3, were correctly recognized; however, there are no false positive. The false negatives are again caused by non-rigidity of the disk. In fact, disk hypotheses are generated for many of these images, but the back-projection computed is not sufficiently accurate to permit verification.

4.2.2 MORSE

The aim here is to recognize the SOR from the test images. There are 20 SORs in the SOR library; these are shown in figure 14. Recognition proceeds against all the models in this library together with the test SOR (i.e. 21 objects in total).

SOR test images Five of the nine images are grouped successfully, and in all of these cases the correct SOR is recognised — i.e. where grouping succeeds there are no false positives or negatives. Grouping fails for two reasons: first, in images s3, s6 and s9, the

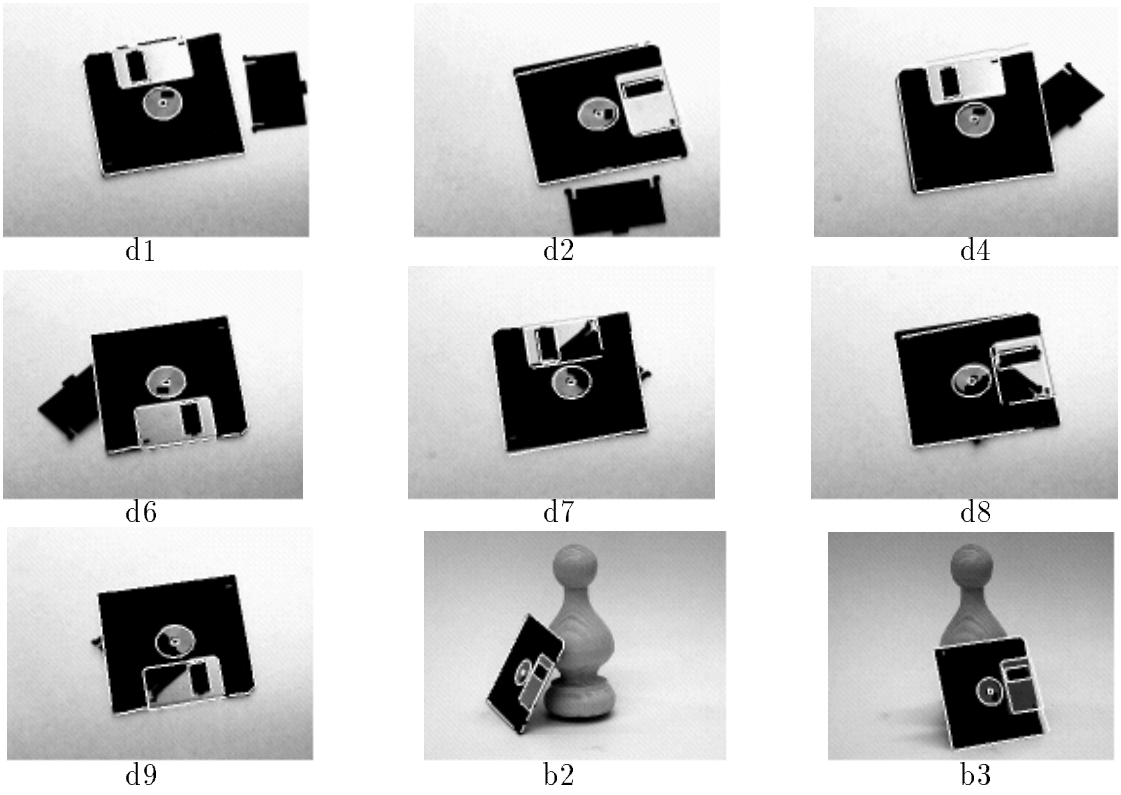


Figure 13: The recognized disk images.

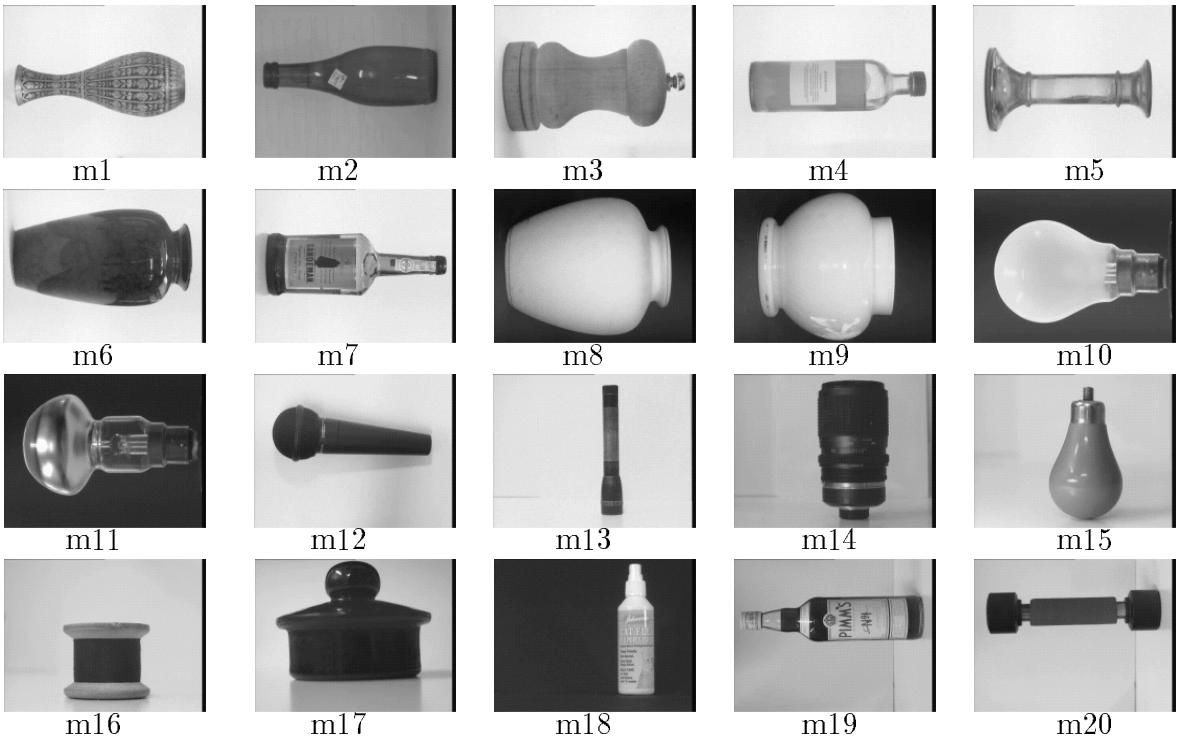


Figure 14: The SOR model-base.

SOR is essentially viewed from above and the "sides of the SOR", necessary to drive the grouping and measure the characteristic geometry, are not visible; secondly, in s8 the bitangent points are occluded, and grouping can not begin. The recognised SORs are shown in figure 15.

SOR and disk test images Three of the six, b1, b4 and b5, are recognised. In all three cases where recognition fails this is due to a grouping failure because too much of the outline is occluded for grouping to begin. Both these tests illustrate the need for more flexible grouping; see figure 15.

Acquisition images Since only one image is required for acquiring the SOR model, recognition can also be tested on all but one of the SOR acquisition images. In this case the first image in figure 6 is used to acquire the model for the SOR. Grouping succeeds in all but the last 3 (most slanted) of the 14 remaining acquisition images. Invariants are formed from those images for which grouping is possible. The values measured for the invariants are given in table 3. Based on these invariants:

One invariant alone With an error bound of 3% on invariant values, invariants (multiple) are always generated for the correct model, but also 3 other models — m7, m13 and m18 in figure 14 — are at times incorrectly hypothesized. These would be false positives if one invariant alone was used.

Two invariants If two invariants corresponding to the same object are required to generate a recognition hypothesis then all the false positives are eliminated (the bound can be increased as far as 5% without introducing new false positives). However, this procedure generates a false negative: for s7 — i.e. a recognition hypothesis is *not* generated for this image. An error bound of 7% generates recognition hypotheses for all images, but again reintroduces false positives, for objects m1 and m7.

Recognition does not involve simply generating recognition hypotheses but also verification of these hypotheses. In all cases the false positives are removed by verification.

4.3 Comparison

The fundamental limitation of the SOR recognition at present is grouping. In particular the reliance on seeing two or more distinguished points (bitangent points) on the outline. There is not, at present, a problem with false positives, even though the model bases contain 21 and 8 models respectively for the SOR and LEWIS systems.

For the appearance-based recognition system there are only 4 models in the library. Of course, the classification of nearby points would be more subject to error as the number of stored manifolds is increased. Also, the test images have been carefully controlled to reduce translations relative to the acquisition and avoid strongly differing illumination conditions.

Image	Invariant Values			
S1	1.33	1.97	2.10	3.26
S2	1.37	2.07	2.10	3.22
S3	1.34	1.97	2.09	3.21
S4	1.31	1.97	2.00	3.20
S5		1.90	2.08	3.27
S6	1.42	1.90	1.99	3.13
				3.48
S7	1.32	1.95	2.05	3.16
S8			2.04	3.12
S9	1.37	1.94		3.00
S10	1.44	2.02	2.96	3.09
S11			2.06	3.24
S12			2.10	3.48
s1	1.36	2.06	2.11	3.26
s2	1.36	2.06	2.08	3.13
s4	1.36	2.05		
s5	1.32	2.00	2.02	
s7			2.10	
b1	1.32		2.11	3.44
	1.47		2.23	3.57
b4	1.33	1.91		3.03
b5		1.90		

Table 3: The invariants produced automatically from the grouped SORs in figures 6, 8 and 10. Note that S1 is the model against which SOR hypotheses are compared. The corresponding recognition images are shown in figure 15.

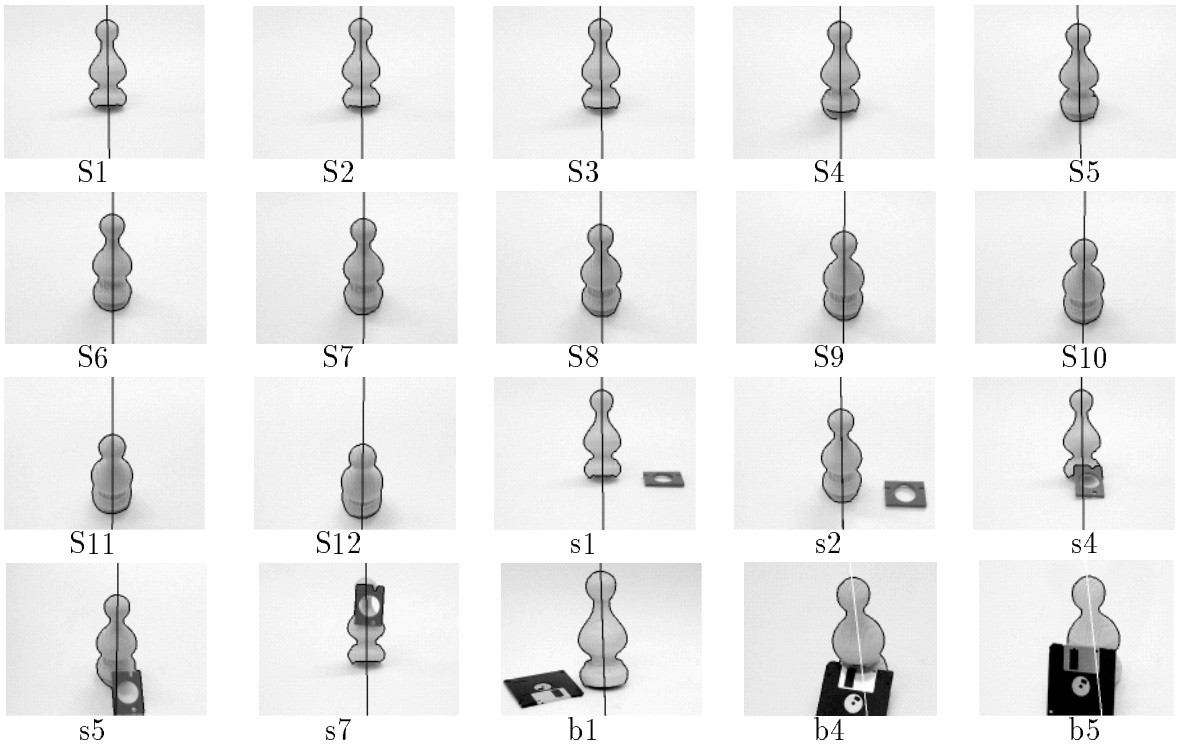


Figure 15: The recognized SOR images. The invariant values for these are given in table 3.

5 Summary and proposal for further comparisons

5.1 General observations

Appearance models have the great advantage of making a minimal commitment to the properties of an object's class. That is, one does not have to have a formal description of the constraints peculiar to an object. Indeed, there are many objects for which there does not exist a formal, mechanisable description outside of its appearance. For example, we currently do not have a theory for the description of even simple curved shapes except for a limited set of geometric classes, such as planarity or symmetry.

On the other hand, it is difficult to see how an appearance model can be generalized to incorporate objects which should be considered to be in the same class. Incidental variations in appearance, such as surface albedo or texture of otherwise identical objects must be treated as completely separate object instances.

Another issue for appearance models is that they are dependent on scene context. The approach depends on some form of figure-ground segmentation and subsequent image normalization. In the presence of occlusion, mutual illumination, and mutual shadows, this segmentation and normalization is difficult to achieve. Geometric models based on boundaries extracted from image discontinuities provide a much more robust description in the presence of these effects. For example, it can be shown that effects of mutual illumination and shadowing lead to complex and unpredictable patterns of intensity in real scene [4]. The only reliable invariants to illumination and mutual object placement are intensity discontinuities. Therefore, it can be expected that geometric boundary

descriptions will be much more invariant than normalized intensity patterns.

It can be argued that these intensity invariants would be eventually captured through a proper statistical analysis of the eigenvectors of the library of appearance models. Eigenvectors which contribute the most mutual information to the object classes should be essentially high-pass filters which emphasize object boundaries. However, it is likely to be more effective to start with an object description which is known to be invariant to irrelevant image variations. Then the information in recognition features is focussed on the salient differences which distinguish objects for recognition.

Geometric class models also provide constraints which support figure-ground segmentation based on class constraints. For example, in the case of rotational symmetry, the involution between corresponding points on the boundary provides a grouping mechanism which can extract SOR features in the presence of complex textured backgrounds and significant occlusion.

Still, for many objects, geometric shape is not the most significant attribute in deciding the class. The key mode of description must be based on appearance attributes such as surface properties or texture markings. Moreover, there are many attributes of shape which do not seem to provide tight constraints for segmentation and grouping, e.g., natural shapes which only have approximate symmetry. For such objects no significant advantage arises from a geometric description except for invariance to illumination.

5.2 Future investigation

This comparison of representations for recognition has raised many significant issues for further investigation. It is now clear to us that this sort of study is essential for rapid progress in object representation. The areas where various representations complement each other is a fertile direction for new research.

At the most abstract level, we have seen that appearance modelling is largely empirical while the geometric invariance model originates from a theoretical understanding of image formation and perspective projection. The contest is then based on the completeness of appearance model data acquisition vs the applicability of a geometric representation of actual shapes in the world. It is impossible to acquire a fully complete appearance model for all variations which could occur and at the same time, there are many aspects of real objects which we still don't have an adequate symbolic representation.

Clearly, the best approach is to combine the two representations. Chris Taylor [6] has made some moves in this direction by using a template to correct for geometry (faces) and then eigenfunctions after geometric and intensity normalisation.

In general, well-founded segmentation and grouping mechanisms should be used to derive object descriptions. Then statistical classifiers can account for the aspects of object appearance which cannot be modelled theoretically. Currently, the geometric invariant recognition systems use statistics only in defining tolerance on invariant values. Additional features can be added, such as surface markings, which can only be described by an appearance model, using geometric segmentation and grouping to isolate specific object surfaces. Such regions are typically defined for a specific object, rather than a class, and therefore are best used during verification.

Some of the significant questions which were raised by this initial study are:

1. What is the effect of a larger modelbase? It is expected that the effective separation of objects in eigenspace and in invariant space will be reduced as the number of library objects is increased.

2. What is the effect of more degrees of freedom on the appearance model manifold? In the current experiments, we only varied one rotational degree of freedom. It might be expected that SLAM's recognition tolerance will be reduced as the dimension of the manifold increases. The invariant representation is not affected by object pose or internal camera parameters which would all have to be included in an appearance model.

3. How severe is the figure-ground isolation problem? The current experimental setup does not provide very challenging background textures. It may be that the power of geometric grouping constraints is essential to object isolation.

4. What is the effect of mutual illumination and shadowing? We did not explore the case where one object shadows another or where light is mutually reflected onto adjacent surfaces. It is not certain how these effects will affect the tolerance of the appearance eigenspace.

5. How will a statistically optimum set of eigenvectors compare to the principal components currently used in SLAM? It is not necessarily the case that the eigenvectors which rapidly converge to a good approximation of the image intensity also provide maximum separation of the object manifolds.

Acknowledgements

We are grateful to Sameer A. Nene for his technical assistance with SLAM. Financial support was provided by several agencies: ESPRIT Project 6448 'VIVA'; the UK EPSRC; and General Electric.

References

[1] Beymer D., "Face Recognition Under Varying Pose", *Proc CVPR*, 756-761, 1994.

[2] Craw I. and Cameron P., "Paramterizing images for recognition and reconstruction", *Proc BMVC*, 367-370, 1991.

[3] Duda, R.O. and Hart, P.E., *Pattern Classification and Scene Analysis,* Wiley, 1973.

[4] Forsyth, D. and Zisserman, A. "Reflections on shading", *PAMI*, 13, 7, 671-679, 1991.

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

[6] Lanitis A., Taylor C.J., Cootes T.F., "A Unified Approach to Coding and Interpreting Face Images", *Proc ICCV*, 368-373, 1995.

[7] 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', *Proc. CVPR*, 1993.

[8] Mukherjee D.P., Zisserman A. and Brady J.M., 'Shape from symmetry—detecting and exploiting symmetry in affine images', *Phil. Trans. R. Soc. Lond. A*, 351, 77-106, 1995.

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

[10] Murase H. and Nayar S.K. 'Visual Learning and Recognition of 3-D Objects from Appearance'. *IJCV*, 14(1), 1995.

[11] Pentland A., Moghaddam B. and Starner T., "View-based and modular eigen-spaces for face recognition", *Proc CVPR*, 84-91, 1994.

[12] Pillow N., Utcke S. and Zisserman A. 'Viewpoint-Invariant Representation of Generalized Cylinders Using the Symmetry Set'. *Image and Vision Computing*, 13(5), 1995.

[13] Rothwell C.A. *Object Recognition through Invariant Indexing.* OUP, 1995.

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

[15] Zisserman A., Forsyth D., Mundy J., Rothwell C., Liu J. and Pillow N., '3D Object Recognition using Invariance', *Oxford University Engineering Report*, OUEL 2027, 1994. To appear *AI Journal.*