

# An Experimental Comparison of Appearance and Geometric Model Based Recognition

J. Mundy<sup>1</sup>, A. Liu<sup>1</sup>, N. Pillow<sup>2</sup>, A. Zisserman<sup>2</sup>, S. Abdallah<sup>2</sup>, S. Utcke<sup>2</sup>,  
S. Nayar<sup>3</sup> and C. Rothwell<sup>4</sup>

<sup>1</sup> General Electric Corporate Research and Development, Schenectady, NY, USA

<sup>2</sup> Robotics Research Group, University of Oxford, Oxford, UK

<sup>3</sup> Dept. of Computer Science, Columbia University, NY, USA

<sup>4</sup> 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 an 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 Lewis/Morse.

## 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 (SORs). 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 preliminary, but do represent an attempt to

directly compare the two recognition processes under identical conditions, i.e. using the same set of test images and model libraries. 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.

This sort of comparison of recognition methodologies 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 reviewed in sections 2 and 3 respectively. The two systems are compared on the same test images in section 4.

## 2 Appearance Models for Recognition — SLAM

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 [11, 14]. The assumption is that if the intensity pattern 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. The major assumptions are that the object can be segmented from the scene and is not (significantly) occluded.

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 neighbourhood 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  $128 \times 128$  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, 15]. PCA captures the variation in image data by projecting

the image onto a low dimensional subspace of eigenvectors which are constructed from the covariance matrix of the image intensities. Since image intensities are typically correlated, it is possible accurately to 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;  $\mu_i$  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 database 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.  $128 \times 128$  is used to sample the bounding box. The intensity is normalized by subtracting the average pixel intensity, and then converted to a unit vector in the 16384 dimensional space.
2. Each resulting image vector is added to the image database with a label corresponding to the object class.
3. A suitable vector subspace is constructed which accounts for the variance in the database [10]. 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, for a rotation interval of  $5^\circ$ , the manifold can be sampled to  $1^\circ$  to provide detailed coverage.
6. To recognize an object in a new image instance, the image is segmented and normalized (as described above) and then projected onto the eigenspace. 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.

### 3 Geometry-Based Recognition — Lewis/Morse

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. In the experiments here, we are using the Lewis and Morse systems in tandem to provide an overall recognition for both planar and SOR object classes in a given image.

### 3.1 Lewis

Here we summarize the main features of the Lewis planar object recognition system. The system is built upon plane projective invariants of the object outline. These invariants have the same value in any perspective image of the object. 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 Fig.1. In more detail:

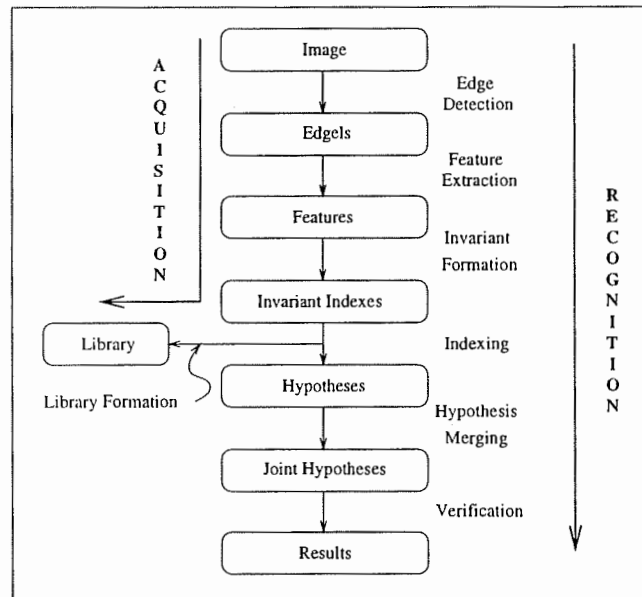


Fig. 1. The recognition system has a single grey scale image as input and the outputs are verified hypotheses with associated confidence values. Many of the processes are shared by the acquisition and the recognition paths.

**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). More details of these invariants are given in [9].

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,  $\mathbf{f}$ , 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  $M(\mathbf{f})$ . The invariant vector formed by the above process represents a point in the multidimensional invariant space. The space is quantized 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.

Back-projection and subsequent searching involves the entire model boundary, not just the features used to form the invariant. Projected model edgels must lie close to image edgels with similar orientation (within 5 pixels and  $15^\circ$ ). 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.** A model can be acquired directly from a single image. No special orientations or knowledge of the camera calibration 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 back-projection 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 [17].

### 3.2 Morse

This is a recognition system for 3D objects. The system is organized 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 establishing 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 targeted 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 an SOR (they satisfy a particular transformation, see section 3.2), and then groups those arising from the same SOR. Such identification and grouping is possible because the 2D image curve

which  
strain  
Subs-  
indic

h  
these  
had  
#2  
verit  
1  
follo  
for t

SOI  
ated  
cons  
jecti  
exac  
T<sup>2</sup> =  
is fr  
ima  
side  
the  
in g

atir  
atic  
it i:  
Th

pai  
out

SC  
res  
po  
ter  
ide  
lin

SC  
e.  
-  
o  
h

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).

Indexing the model library via the invariants generates recognition hypotheses for particular *models* of that class. For example, if an SOR 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 onto the image.

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 [18].

**SOR Grouping.** The imaged outline 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  $3 \times 3$  matrix  $T$ , where  $T^2 = I$ . Two pairs of point correspondences determine  $T$ . This transformation is fundamental to grouping outlines of SORs: the line of fixed points of  $T$  is the imaged symmetry axis;  $T$  provides point to point correspondence between the sides of the outline; this disambiguates the matching of bitangents used to form the invariants; and finally,  $T$  can be used to *repair* missing outline portions, filling in gaps by transforming over points from the other side of the outline.

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 planar 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  $T^2 = I$ .

For each SOR in an image, the outcome of the grouping is an identified axis, pairs of bitangents (one half of each pair on either side of the axis), and repaired outline contours around the object; as illustrated in Fig.12.

**SOR Invariants.** It has been shown previously [7] that intersections of corresponding pairs of bitangents on an SOR's imaged outline 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.

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 construction of the invariant is illustrated schematically in Fig.2.

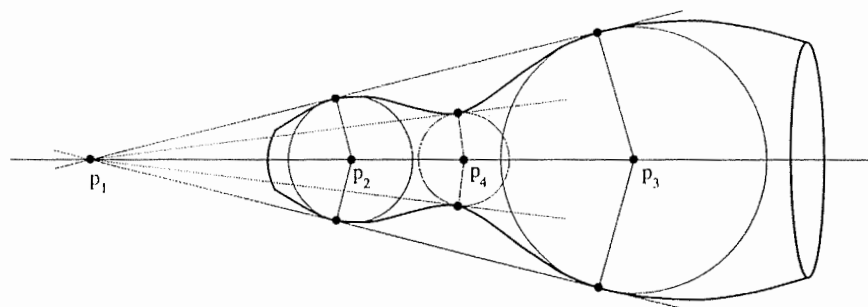


Fig. 2. The construction used to generate the SOR quasi-invariant. The invariant is the cross-ratio of four distinguished axis points. The first,  $p_1$ , is the intersection of the concavity's bitangent lines;  $p_2$  and  $p_3$  are the centres of circles tangent to the outline at each of the bitangent points; by a similar circle construction,  $p_4$  comes from tangent lines cast from  $p_1$  to the interior of the concavity.

As in Lewis, the invariants are used to index into a model-base to generate hypotheses, which are then merged and finally verified using back-projection.

#### 4 Experiments

Four objects are used for the experiments, three SORs and one planar shape (a floppy disk). The four objects are shown in Fig. 3.

**Acquisition Images** A set of acquisition images of each object is required for SLAM in order to sample the pose space. For the SORs, the slant of the symmetry axis was varied in  $5^\circ$  steps, keeping the SOR at the horizontal centre of the image. The 15 acquisition images for SOR1 are shown in Fig. 4. Similar sets are used for SOR2 and SOR3. The planar object was imaged in an oblique view and the orientation about the normal to the object's plane is varied in  $10^\circ$  steps to acquire its appearance model. The 36 acquisition images are shown in Fig. 5. The illumination in all cases was from a diffuse source and maintained constant over the image collection.

**Test Images** There are three sets of test images:

1. **SOR1 + distractor:** 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. An unmodelled distractor is included to allow partial occlusion of the SOR. The images are shown in Fig. 6.
2. **Disk + distractor:** 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. An unmodelled



it is the  
he con-  
at each  
es cast

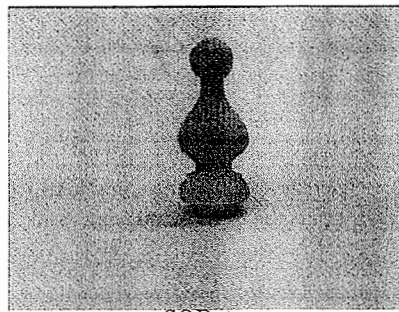
nerate  
ion.

pe (a

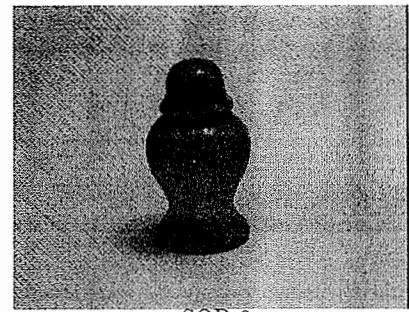
d for  
sym-  
re of  
sets  
view  
teps  
ig.5.  
tant

rees  
on;  
led  
are.

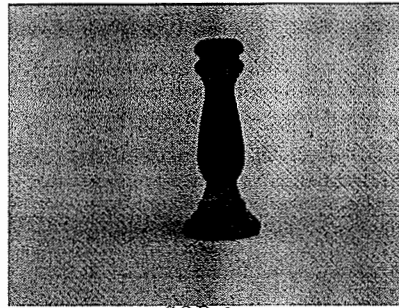
ees  
on;  
led



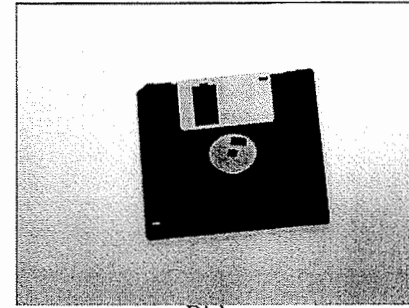
SOR 1



SOR 2



SOR 3



Disk

Fig. 3. The four objects used in the experiments.

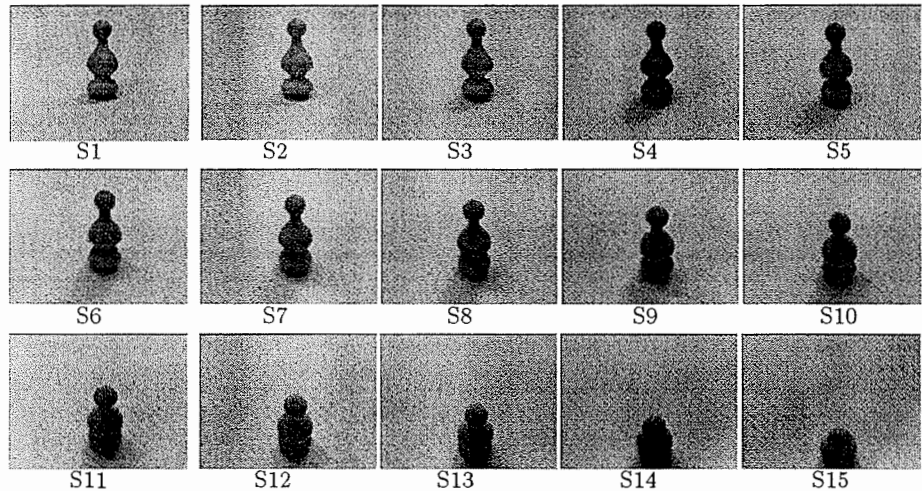


Fig. 4. The acquisition images of the SOR.

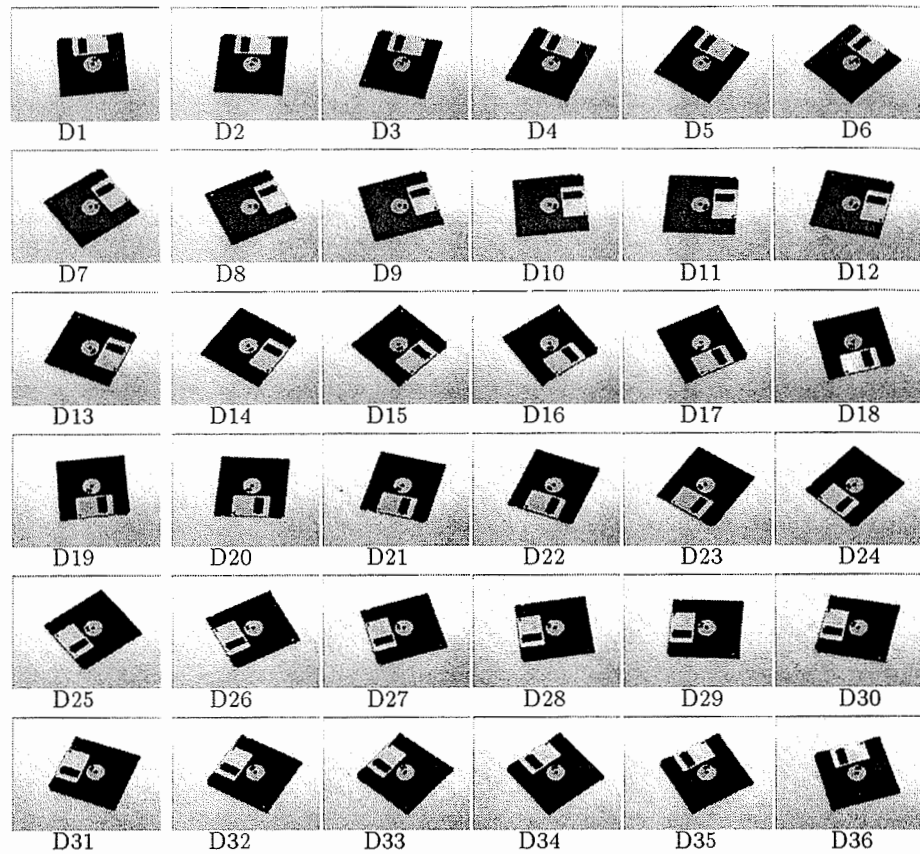


Fig. 5. The acquisition images for the planar object.

distractor is included to allow partial occlusion of the disk. The images are shown in Fig. 7.

3. **SOR1 + disk:** There are 6 images corresponding to two viewpoints for the SOR, with the disk varying in pose, causing different occlusions for both the SOR and disk. The images are shown in Fig. 8.

#### 4.1 Appearance-Based System

Each set of acquisition images is normalized as described in section 2. In the experiments below recognition is tested against two different model libraries. Each model library is constructed by including all the acquisition images (i.e. in this case, one set for each object that is modelled), and constructing the PCA eigenspace.

The two model libraries are: first, only the three SORs of Fig. 3; and second, all four objects of Fig 3 (i.e. now including the disk) together with 20 other

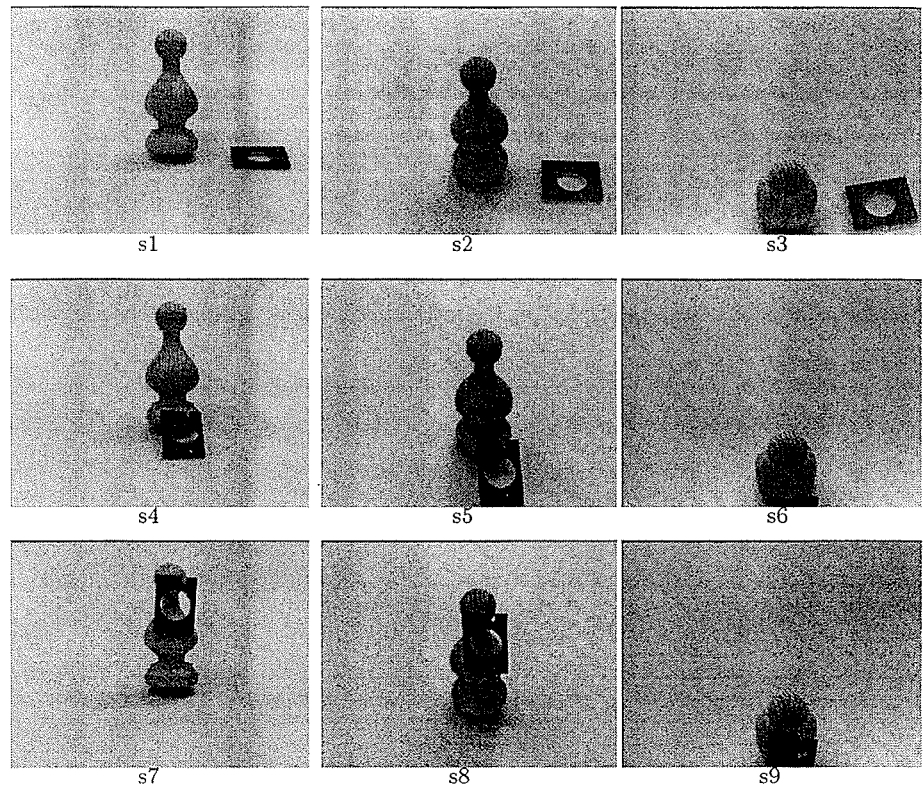


Fig. 6. The SOR test images with distractor.

objects shown in Fig.9. The first model library allows SLAM and Morse to be compared on the same set of objects.

The manifold representing each object can be displayed as a three dimensional subspace by projecting the acquisition images onto the first three principal components. In the case of these experiments, the manifolds are curves since only one parameter is varied in generating the appearance model. A typical example is shown in Fig.16.

The actual eigenspace used for recognition has ten dimensions, i.e. recognition proceeds by normalizing a new image and then projecting it onto the sub-space defined by ten eigenvectors. The closest stored point, within a tolerance threshold, is retrieved and associated with the stored label to classify the object.

#### 4.2 Geometry-Based System

In the Lewis and Morse systems only one image is needed in principle to acquire a model. However, to reduce the effects of measurement error and to determine

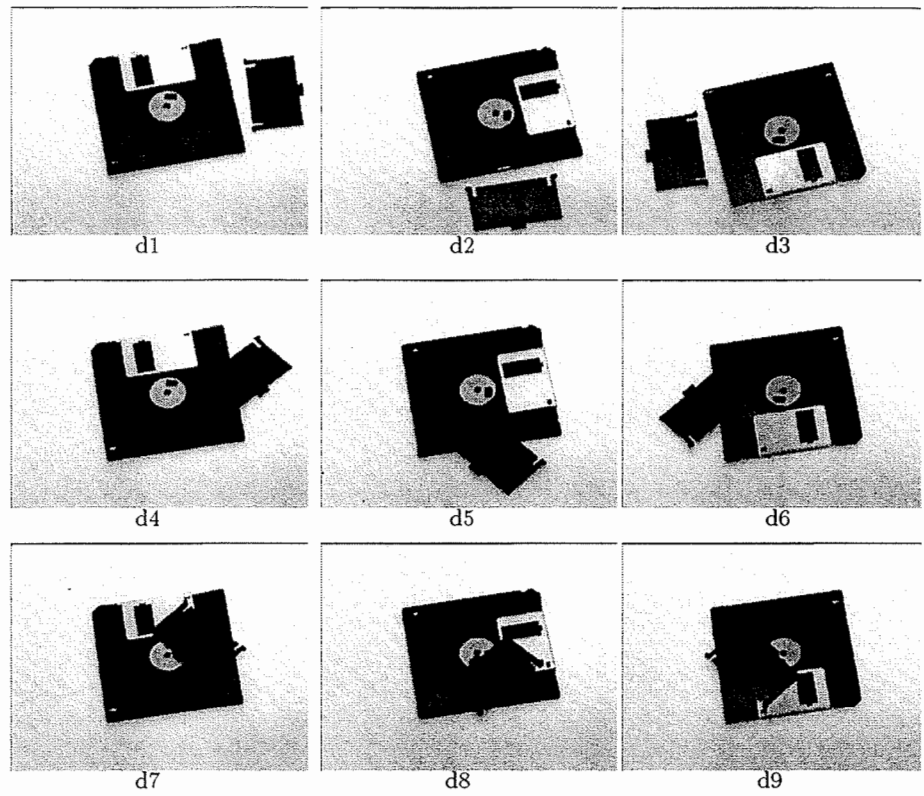


Fig. 7. The disk test images with distractor.

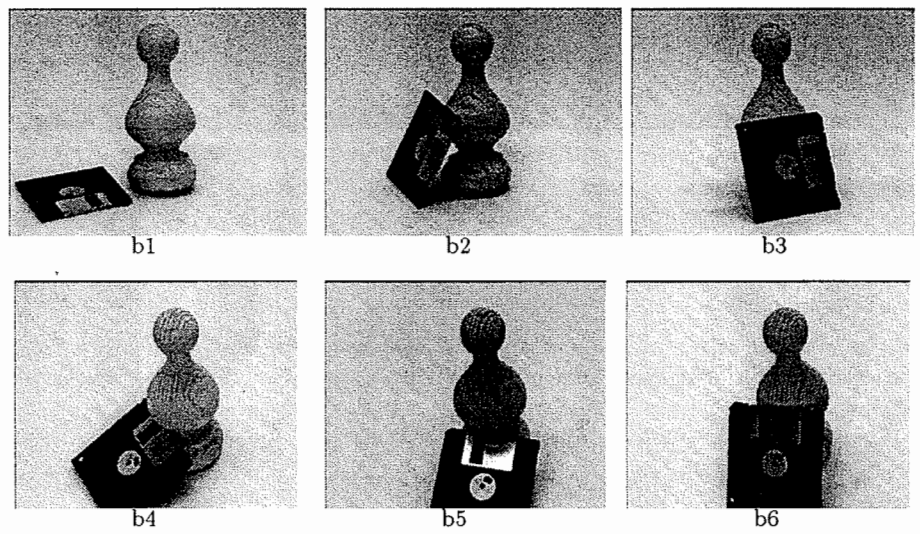


Fig. 8. Test images containing both SOR1 and the disk.

UNIVERSITY OF TORONTO LIBRARY



Fig. 9. The objects which, with those of Fig. 3, form the larger of the model-bases used by SLAM.

variances, models are generally constructed by averaging information from a number of images. This is particularly important in the case of the disk for the Lewis system, because of non-rigidity of the disk (see below).

For the Morse system two model libraries are used. First, a library consisting of only the three SORs of Fig. 3. Second, a library consisting of these three SORs, together with 21 others. The complete set of 24 SORs is shown in Fig. 10.

For the Lewis system there are 8 planar objects in the model library; these are the disk of Fig. 3, together with the 7 objects shown in Fig. 11.

In all the following experiments the same parameters are used for all images.

#### 4.3 Results I — Identical Model Libraries — 3 SORs only

This test is not applicable to Lewis, since it cannot represent SORs. Only SLAM and Morse are applied here, each having only the 3 SORs (SOR1, SOR2, SOR3) in their model library.

##### Test images: SOR1 + distractor (Fig. 6)

**SLAM:** SOR1 is correctly recognized in all nine images.

**Morse:** Five of the nine images are grouped successfully, and in all of these cases the SOR1 is correctly recognized — i.e. where grouping succeeds there are no false positives or negatives. Grouping fails for two reasons: first, in

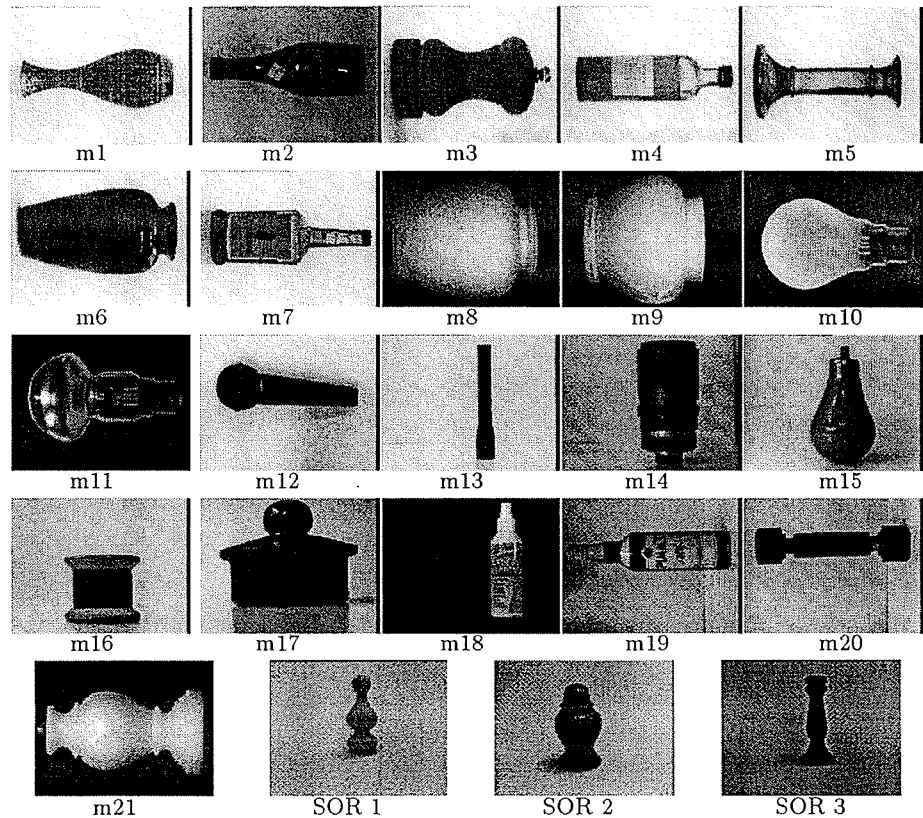


Fig. 10. The entire SOR model-base.

images s3, s6 and s9, the 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 recognized SORs are shown in Fig.12.

#### Test images: SOR1 + disk (Fig.8)

**SLAM:** SOR1 is recognized in two of the six images, b1 and b5. In the other four images an incorrect SOR (SOR2 or SOR3) is recognized. The results are listed in table 1.

**Morse:** Three of the six, b1, b5 and b6, are recognized. In the three cases where recognition fails this is due to a grouping failure because too much of the outline is occluded for grouping to begin. There are no false positives. The recognized SORs are shown in Fig.13.

The occlusion of one object by another causes two problems for SLAM: first, it makes segmentation more difficult; second, it perturbs the position of the

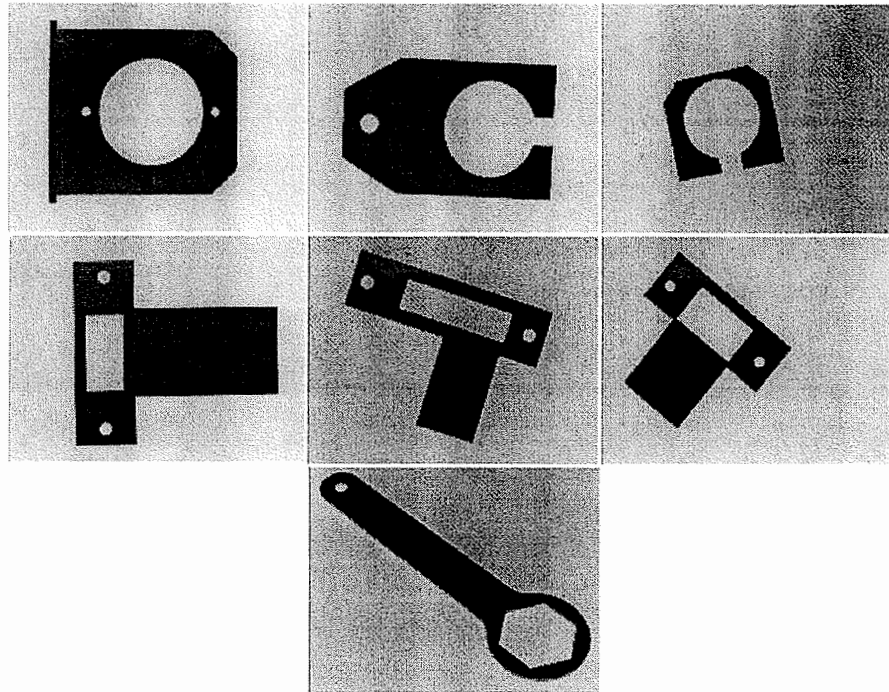


Fig. 11. Images of the objects contained in the Lewis model library.

projection of the image into the eigenspace [12]. It is unclear which is applicable here.

**Summary** The conclusion from this test is that the geometric system, Morse, is fail safe — i.e. there are no false positives, but grouping is its weakness at present. It correctly recognizes SOR1 in 8 of the 15 test images which contain it (7 false negatives, no false positives). The appearance system, SLAM, can have a problem distinguishing between the 3 SORs. It correctly recognizes SOR1 in 11 of the 15 images. However, there are 4 false positives where the wrong SOR is recognized (in total: 4 false negatives, 4 false positives).

#### 4.4 Results II — Variation with Model Libraries

Here the number of objects in each model library is increased to investigate how this affects performance. For SLAM, 24 objects are used (the three test SORs, the disk, and twenty others). For Morse, 24 SORs are used (i.e. the three test SORs and 21 others), and for Lewis 8 planar objects are included (the disk and 7 others).

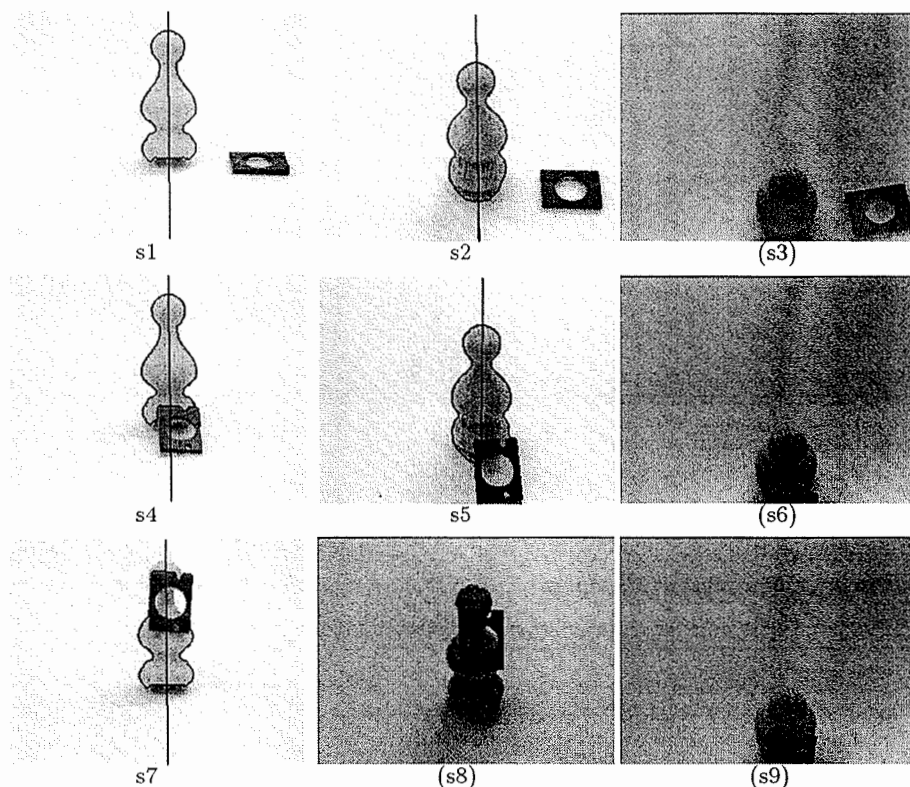


Fig. 12. The SOR test images recognized by Morse. An axis is super-imposed on each that is correctly identified; those with labels in brackets failed.

| Image | SLAM-Classified Object | Nearest Distance / $10^7$ |
|-------|------------------------|---------------------------|
| b1    | SOR1                   | 1.13                      |
| b2    | SOR2                   | 1.02                      |
| b3    | SOR3                   | 1.20                      |
| b4    | SOR2                   | 0.73                      |
| b5    | SOR1                   | 1.68                      |
| b6    | SOR2                   | 1.31                      |

Table 1. SLAM recognition results for images in Fig.8 (SOR1 and disk) against a model library of SOR1, SOR2 and SOR3.

Fig. 1  
model  
identif

Test  
SLA  
Mor:  
n  
fi  
a  
w  
b  
Lewi  
n  
fa

Test  
SLA  
Mor:  
o  
p  
Lewi  
d

Test  
SLA  
a

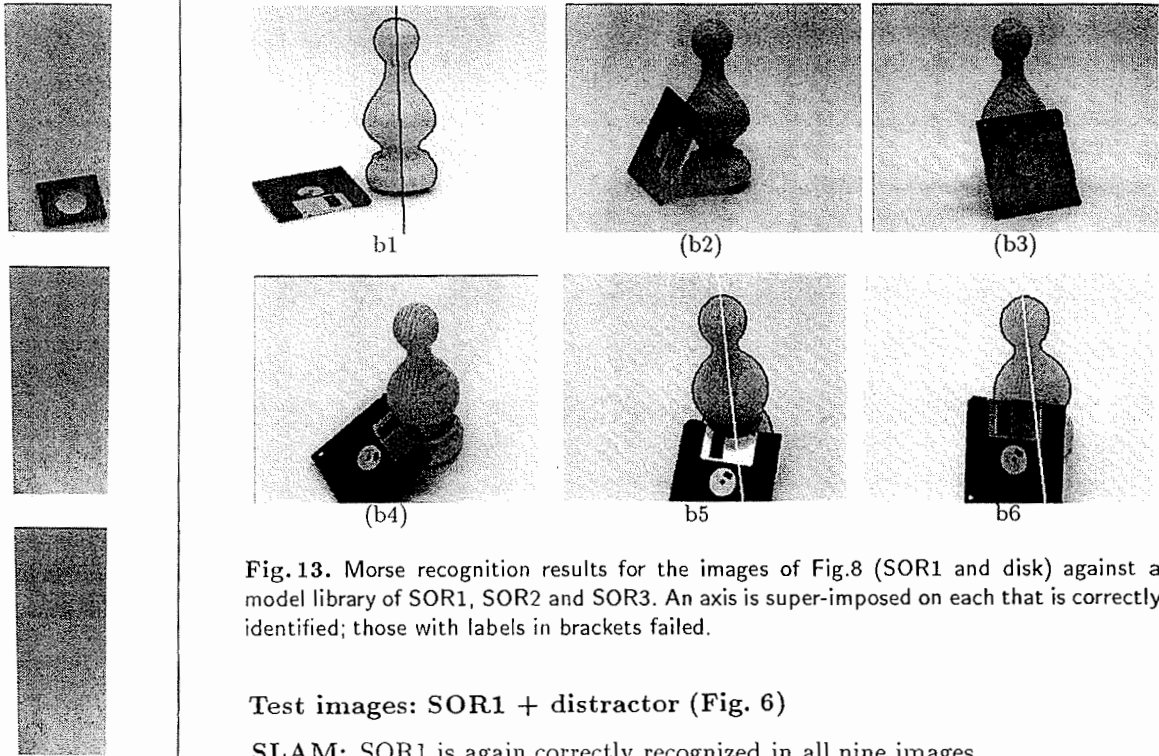


Fig. 13. Morse recognition results for the images of Fig. 8 (SOR1 and disk) against a model library of SOR1, SOR2 and SOR3. An axis is super-imposed on each that is correctly identified; those with labels in brackets failed.

#### Test images: SOR1 + distractor (Fig. 6)

**SLAM:** SOR1 is again correctly recognized in all nine images.

**Morse:** The performance of the system did not change when an additional 21 models were added to the library, i.e. SOR1 is again correctly recognized in five of the nine images as shown in Fig. 12. The recognition was successful in all cases where the SOR could be grouped. Although recognition hypotheses were generated for other SORs in a few cases, these were always eliminated by the verification stage so there were no false positives.

**Lewis:** None of the objects in the Lewis model-base appear in these images, so no models should be recognized. This is indeed the case, i.e. there are no false positives.

#### Test images: Disk + distractor (Fig. 7)

**SLAM:** The disk was correctly recognized in all cases.

**Morse:** Neither of the objects in these images is in the Morse library (SORs only), and no SORs were recognized in these images i.e. there were no false positives.

**Lewis:** The disk was correctly recognized in all nine images. The recognized disks are shown in Fig. 14.

#### Test images: SOR1 + disk (Fig. 8)

**SLAM:** In all cases the nearest manifold was SOR2, i.e. recognition failed on all images.

ach that

ainst a

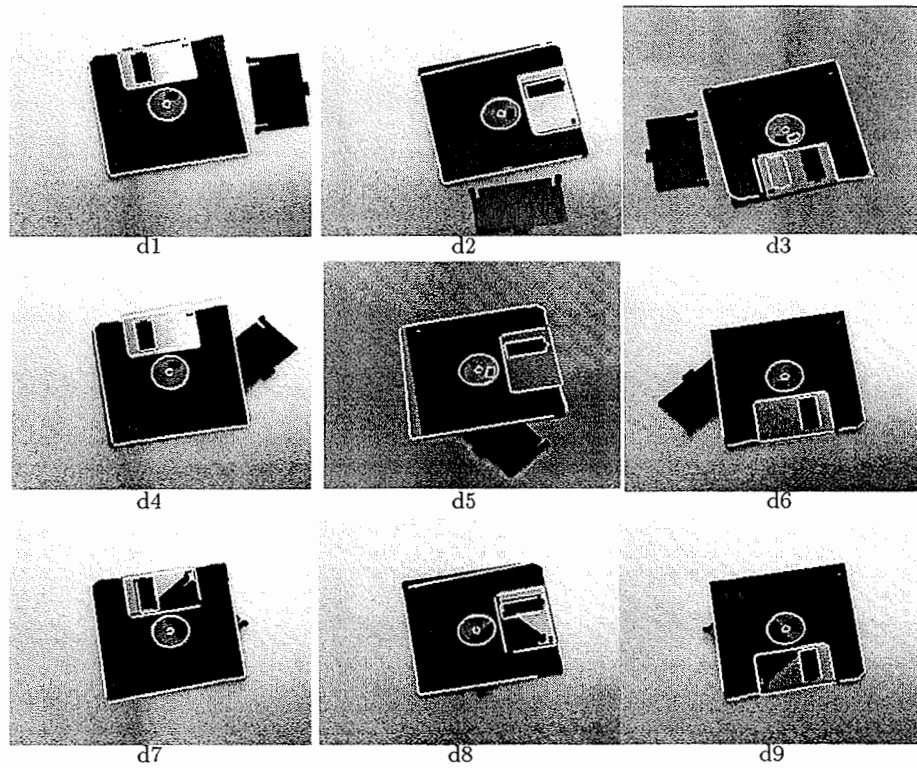


Fig. 14. The disk + distractor images all successfully recognized by Lewis. The images show the model outline back-projected onto the image.

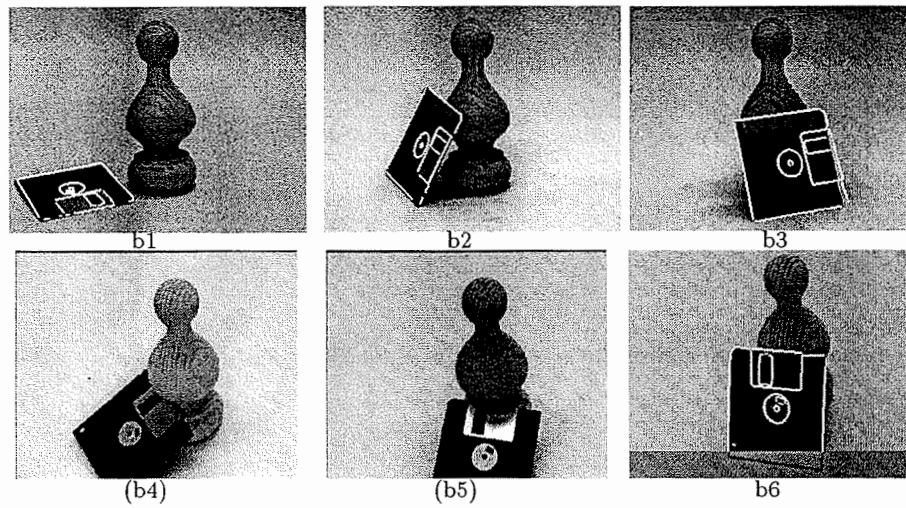


Fig. 15. The disks recognized by Lewis in the disk + SOR images. Where successful, the back-projected outline is displayed; those that failed have labels in brackets.

Morse:  
els.  
posi  
Lewis:  
and  
Fig.

The  
disk its  
from tw  
disk, an  
non-ri

1. The  
of t  
acq  
and  
high  
the
2. The  
ima  
veri

Summ:  
employ  
fication  
systems  
ance of  
the case  
in these  
explana

## 5 Su

### 5.1 G

It is clea  
one ano  
formal c  
approac  
Current  
except f  
(pipes)

On 1  
generali  
geometr

**Morse:** Again, the performance does not change on adding additional models. Three of the six, b1, b4 and b5, are recognized, and there are no false positives. Results are shown in Fig.13.

**Lewis:** The disk was correctly recognized in four of the six images, b1, b2, b3 and b6, and there are no false positives. The recognized disks are shown in Fig.15.

The two Lewis failures, b4 and b5, were caused by a relative translation of the disk itself and the disk case: The invariant used for these two images is computed from two lines and one conic; the conic is the central circular component of the disk, and this can move relative to the case. There are two consequences of this non-rigidity:

1. The invariants deviate from a fixed value. However, the value and variance of this invariant is obtained by averaging measurements from a number of acquisition images in Fig.5 (where there is relative motion between the disk and case between images). Consequently, the invariant will have a relatively high variance and measured values should lie in the predicted range. This is the case: disk *hypotheses* are generated for b4 and b5.
2. The back-projection of the model outline will be displaced from its veridical image position. The back projection is used for verification, and it is the verification stage which prevents recognition in these two cases.

**Summary** Again, both the geometric systems are fail-safe. This is due to the employment of a verification stage. The appearance system does not have a verification test, and does produce false positives. The performance of the geometric systems is not affected by adding additional objects to the library. The performance of the appearance system is only affected by additional library models in the case of the SOR1 and disk test images. However, an SOR is still recognized in these images, even though this is not always the correct SOR. A possible explanation for this is shown in Fig.16.

## 5 Summary and Proposal for Further Comparisons

### 5.1 General Observations

It is clear that both approaches have strengths and drawbacks that complement one another. Appearance models have the great advantage of not requiring a formal description of the constraints peculiar to an object. In the geometric approach this description is required to facilitate pose invariant recognition. Currently, there is not a theory for the description of even simple curved shapes, except for a limited set of geometric classes such as SORs and canal surfaces (pipes) [16], and certainly not for arbitrary shapes.

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 geometric class. Incidental variations in appearance, such as surface albedo or

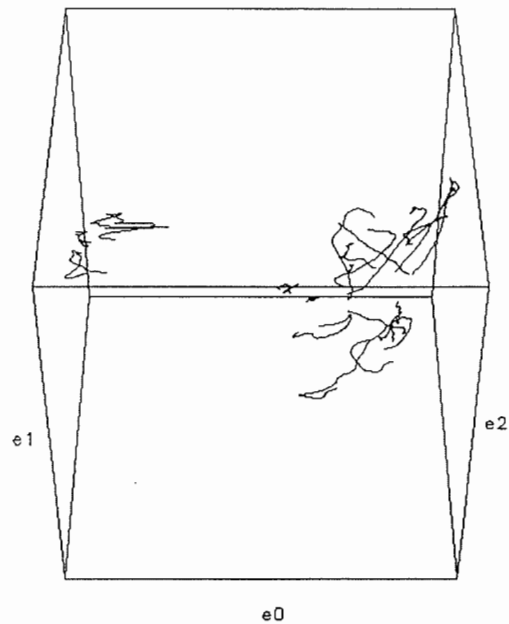


Fig. 16. Each object generates a one-dimensional manifold parametrized by the rotation applied during image acquisition. Here the manifolds (curves) are plotted in a sub-space defined by the first three eigenvectors. The one dimensional manifolds for the three SORs (left) are well separated from the one dimensional manifolds of the other objects (to the right). This may explain why an SOR is correctly distinguished from other objects but can be confused with other SORs

texture of otherwise identical objects must be treated as completely separate object instances.

A significant advantage of the geometric approach is that geometric class models also provide constraints which support figure-ground segmentation. For example, in the case of rotational symmetry, the relation between corresponding points on the imaged boundary provides a grouping mechanism which can extract SOR features in the presence of complex textured backgrounds and significant occlusion.

On the other hand if grouping is not successful, then the geometric approach cannot proceed. This is the problem with the current implementation state of Morse: a very powerful grouping constraint is available for SORs, but at present it is only applied to curves if they contain a bitangent. Consequently, if the bitangent points are occluded no grouping occurs. However, first, it is clear why the system has failed; and second, there is considerable potential for improvement by employing the constraint more fully (i.e. considering other curves for grouping). The appearance approach is inferior in both these respects.

For appearance models, some alleviation of the segmentation and occlusion problems can be obtained by partial appearance matching — finding subparts

of objects and checking the consistencies of the geometrical arrangements of the subparts for recognition, see [13] and also Schmid & Mohr (these proceedings). Nevertheless, SLAM, as a specific implementation of appearance modelling, currently has the deficiency that it can only recognize one object in an image (even though the image may contain several modelled objects) and this can be attributed to the difficulty of figure-ground segmentation in appearance systems.

A related problem to segmentation is the normalization required by appearance systems. In the presence of mutual illumination, and mutual shadows, this normalization is difficult to achieve. For example, it can be shown that effects of mutual illumination and shadowing lead to complex and unpredictable patterns of intensity in real scenes [4]. For surfaces with a Lambertian reflectance map, the dimension of the illumination manifold (for each view) is reduced to three, even in the presence of mutual illumination (see papers in this collection). However, varying shadow structures are still a significant obstacle. 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.

Finally there is the issue of statistical variation. Geometric models impose hard constraints on image geometry. This is an advantage — in that it facilitates powerful grouping mechanisms and allows strict verification, but also a disadvantage in that variations from these constraints can result in recognition failing. This was exemplified by the relative motion of the disk and its case. This non-rigidity caused the Lewis system to fail on two examples, whilst SLAM tolerated the variation because the disk was still the nearest manifold. However, the lack of a verification stage in SLAM does result in false positives.

## 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. On the other hand, currently there are suitable geometric representations for only a small number of classes, and also a 'model' should be more than geometry alone.

Clearly, a resolution 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 use eigenfunctions after geometric and intensity normalization. In some applications, starting with local features and following with geometry, may be a useful alternative.

tation  
space  
SORs  
to the  
it can

arate

class  
. For  
ond-  
can  
sig-

each  
te of  
sent  
bit-  
the  
it by  
ng).

sion  
arts

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 model-base? It is expected that the effective separation of objects in eigenspace (for SLAM) and in invariant space (for Lewis/Morse) will be reduced as the number of library objects is increased. A preliminary answer to this question has been given in this paper.
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 or object textures. Will the geometric grouping constraints be sufficiently powerful to isolate an object with surface markings and texture from a textured and cluttered background?
4. How will the computational complexity of recognition scale with complexity of the scene? Can geometric grouping be made efficient in a textured and cluttered scene?
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 [5].

#### Acknowledgements

Lewis was originally developed by Charlie Rothwell, and was subsequently implemented in C++ primarily by Charlie Rothwell, with contributions from Bill Hoffman and Chien-ming Huang. Morse, has been implemented using the same basic C++ libraries as Lewis primarily by Nic Pillow, with contributions from Jane Liu and Sven Utcke. The SLAM software was developed by Sameer Nene, Shree Nayar, and Hiroshi Murase at Columbia University. We are grateful to Sameer A. Nene for his technical assistance in using the SLAM software. Financial support was provided by several agencies: ESPRIT BRA Project 'IMPACT'; the UK EPSRC; and General Electric.

Re

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

13.

14.

15.

16.

17.

18.

## References

1. Beymer D., 'Face Recognition Under Varying Pose', *Proc CVPR*, 756-761, 1994.
2. Craw I. and Cameron P., 'Parametrizing 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. Belhumeur N., Hespanha J. and Kriegman D., 'Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection', *Proc. ECCV*, 45-58, 1996.
6. Lanitis A., Taylor C.J. and 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. Murakami H. and Kumar V., 'Efficient calculation of primary images from a set of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4, 511-515, 1982.
11. Murase H. and Nayar S.K., 'Visual Learning and Recognition of 3-D Objects from Appearance', *IJCV*, 14, 1, 1995.
12. Murase H. and Nayar S., 'Illumination Planning for Object Recognition Using Parametric Eigenspaces' *IEEE Trans. PAMI*, 16, 12, 1219-1227, 1995.
13. Murase H. and Nayar S., 'Image Spotting of 3D Objects Using the Parametric Eigenspace Representation', *Proc. of 9th Scandinavian Conference on Image Analysis*, 325-332, June 1995.
14. Nene S.A., Nayar S. and Murase H., 'SLAM: Software Library for Appearance Matching,' *Proc. of ARPA Image Understanding Workshop, Monterey*, November 1994.
15. Pentland A., Moghaddam B. and Starner T., 'View-based and modular eigenspaces for face recognition', *Proc CVPR*, 84-91, 1994.
16. Pillow N., Utcke S. and Zisserman A., 'Viewpoint-Invariant Representation of Generalized Cylinders Using the Symmetry Set', *Image and Vision Computing*, 13, 5, 1995.
17. Rothwell C.A. *Object Recognition through Invariant Indexing.*, OUP, 1995.
18. Zisserman A., Forsyth D., Mundy J., Rothwell C., Liu J. and Pillow N., '3D Object Recognition using Invariance', *AI Journal*, 78, 239-288, 1995.