A Perceptual Grouping Hierarchy for Appearance-Based 3D Object Recognition *

Andrea Selinger and Randal C. Nelson
Department of Computer Science
University of Rochester
Rochester, NY 14627
(selinger, nelson)@cs.rochester.edu

Abstract

In this report we consider the problem of 3D object recognition, and the role that perceptual grouping processes must play. In particular, we argue that a single level of perceptual grouping is inadequate, and that reliance on a single level of grouping is responsible for the specific weaknesses of several well-known recognition techniques. Instead, we argue that recognition must utilize a hierarchy of perceptual grouping processes, and describe an appearance-based system that uses four distinct levels of perceptual grouping, the upper two novel, to represent 3-D objects in a form that not only allows recognition, but reasoning about 3D manipulation of a sort that has been supported in the past only by 3D geometric models.

Key Words: Perceptual grouping, Object recognition, Appearance-based representations.

*Support for this work was provided by ONR grant N00014-93-I-0221, and NSF IIP Grant CDA-94-01142

1 Introduction

1.1 Overview

When designing a computer vision system, one may try to demonstrate a similarity to or take inspiration from biological vision. Although some researchers in the computer vision community argue that such parallels are unneeded we shouldn't overlook the fact that "biological vision is currently the only indication we have that the general vision problem is even open to solution" [11]. A particularly fruitful topic in this regard has been the the demonstration and categorization of psychological grouping phenomena. The initial contributions to this area came from the field of Gestalt psychology, developed by Max Wertheimer [21] which divided grouping on the basis of proximity, similarity, continuation, closure, symmetry and familiarity. This taxonomy, the induced categorization and implied functional separation has been enormously influential, not only in later psychological an psychophysical studies of vision, but in the the study of machine vision as well, particularly in the area known as perceptual grouping.

In this paper we consider the problem of 3D object recognition, and the role that perceptual grouping processes must play. In particular, we argue that a single level of perceptual grouping is inadequate, and that reliance on a single level of grouping is responsible for the specific weaknesses of several well-known recognition techniques. Instead, we argue that recognition must utilize a hierarchy of perceptual grouping processes, and describe an appearance-based system that uses four distinct levels of perceptual grouping to represent 3-D objects in a form that not only allows recognition, but reasoning about 3D manipulation of a sort that has been supported in the past only by 3D geometric models.

The use of a hierarchy of grouping processes is not new, of course. A very nice recent example is the system developed by Havaldar, Medioni and Stein [7]. What distinguishes our system is the grouping processes we use at the higher levels, which give certain exceptional capabilities. Our first two levels use good continuation and image proximity, in common with many approaches. Our third level of grouping however, is closest to what the Wertheimer termed "familiarity", in that it uses experience memory to drive the grouping process. This category has not been much explored in the perceptual grouping literature, but it turns out to be the essence of our recognition process.

Our fourth level of grouping uses proximity, not in image space, but rather in manipulation space (which amounts to out-of-plane rotation space for rigid objects). This induces a topology on view-based object appearances, that when coupled with the lower-level representations can permit complex 3D manipulations to be planned, specified, and carried out on a purely visual basis, without the need for 3D geometric models. Again, this is a means of grouping that has received little attention in the grouping literature, but which produces representations that are practical to obtain, and of considerable potential value.

1.2 Object Recognition

Object recognition is an important and much-researched problem in the study of both machine and human vision. Until recently, the most successful computational work on object recognition has used model-based approaches in which the image is matched against ex-

plicitly represented 3-D geometric models. Such is the work of Lowe (1987), Lamdan and Wolfson (1988), Huttenlocher and Ullman (1990), and Grimson (1990) [12, 10, 9, 5]. While explicit models provide a framework that allows powerful geometric constraints to be used to good effect, they are severely limited in the sort of objects they can represent, and obtaining models is typically a difficult and time-consuming process.

Appearance-based object recognition methods have been proposed as an alternative, in order to make recognition systems more general, and more easily trainable from visual data. Most of these operate by comparing a two-dimensional, image-like representation of object appearance against many prototype representations stored in a memory, and finding the closest match. The system developed by Poggio and Edelman (1990) [16] recognizes wire objects, Rao and Ballard (1995) [17] use the responses of a set of steerable filters to images of objects, Murase and Nayar (1993) [13] and later Huang and Camps (1997) [8] use principal component analysis. Schmid and Mohr (1996) [18] have recently reported good results for an appearance based system with a local-feature approach similar in spirit to what we use, though with different features and without using feature likelihood measures in the evidence combination scheme.

In general, the appearance-based approach has proven to be a useful technique. However, matches are generally made to representations of complete objects, and require that the image be first segmented into regions that represent entire objects. In other words, all the pixels belonging to an object are grouped together and these groups are later used for matching. Unfortunately this kind of high-level perceptual grouping seems to be infeasible, at least as a bottom-up process in any but rather constrained imaging conditions (e.g. objects against uniform background). In particular, whole-object segmentation systems fall apart quickly in the presence of even minor amounts of occlusion or adjacent background clutter. Thus the reliance of the method on a single (strong but non-feasible) level of perceptual grouping is a significant weakness.

In order to overcome the dependence on good whole-object segmentation, evidence combination schemes such as Hough transform methods (and other voting techniques), have been employed to allow evidence from disconnected low-level features to be effectively combined. These features tend to be individual curves or line segments, thus making use of perceptual grouping at a low level. But this method also has its drawbacks [6]. Most obviously, the size of the voting space increases exponentially with the number of degrees of visual freedom. The size of this space makes it difficult to apply such techniques directly when more than about 3 DOF are involved, thus limiting the use of the technique for 3-D object recognition, which generally involves at least 6 DOF. More serious however, is the problem of false positives in the presence of clutter due to the low information content of individual features. From another viewpoint, this weakness can be attributed to relying on the results of a single (feasible but weak) level of perceptual grouping.

In this paper we describe a method that, by using a hierarchy of grouping processes, overcomes the problems of single-level grouping. It addresses the difficulties in the case of clutter and occlusion that arise in traditional memory-based methods that use a single high-level grouping process. It also resolves the problems of space and false-positives seen in voting methods for high DOF problems that arise from single low-level grouping processes. This system demonstrates robust recognition of a variety of 3-D shapes, ranging from sports cars and fighter planes to snakes and lizards, over full spherical or hemispherical ranges (and

planar scale, translation and rotation) and is robust against clutter. This is in contrast to some recent results, e.g. Murase and Nayar [13] where essentially only one of the two out-of-plane rotational degrees of freedom is spanned, and clutter is a significant problem. To our knowledge these represent the best reported results for full-sphere recognition of general shapes with occlusion and clutter resistance.

Our system operates on four levels of perceptual organization. First, perceptual organization is used to group the pixels into contours. Then the contours are grouped into context patches. The context patches are grouped into 2-D views, and last, but not least, the 2-D views are grouped into the characterization of the 3-D object.

The paper is organized as follows: after an overview of the system (section 2) we describe the levels of perceptual organization in more detail, give the biological relevance and speculate on further applications. Finally we give some results of our experiments.

2 Overview of the Method

Our system represents a 3D object as a fourth-level perceptual group, consisting of a topologically structured set of flexible, 2-D views each derived from a training image. In these views, which represent third-level perceptual groups, the visual appearance of an object is represented as a loosely structured combination of a number of local context regions. These local context regions represent second-level perceptual groups, and can be thought of as image patches that surround key first-level features and contain a representation of other first-level features that intersect the patch. The first level features are the result of first level grouping processes run on the image, typically representing connected contour fragments, or locally homogeneous regions. In the current implementation we use only local contour fragments.

The object recognition system is based on the idea that under different conditions (e.g. lighting, background, or small changes in orientation) the first level feature extraction will find some of the first level key features occurring in the visually derived model, but in general not all of them. The subsequent grouping processes allow us to deal effectively with this fact. In particular, the information content of the first-level features is too low for evidence combination techniques to work well for object recognition in the presence of clutter. The power of the features is thus augmented by forming the second-level groups, referred to as *keyed context patches* which embed the key features in local context. Even these local context regions are frequently consistent with several object/pose hypotheses; hence we use the third-level grouping process to organize the context patches into globally consistent hypotheses about object identity and pose. At this point, we have effectively achieved object recognition. The fourth level of grouping is intended primarily to allow interaction with a manipulation of the recognized objects, though as we discuss later it can also be used to increase the reliability of the recognition process.

In more detail, we make use of distinctive local features we have called *keys* (these are the first level groups in our hierarchy), embedded in and seeding a local context (second level groups). A key is any robustly extractable part or feature that has sufficient information content to specify a configuration of an associated object together with enough additional, pose-insensitive (sometimes called semi-invariant) parameters to allow efficient indexing into

the database. The second level of grouping into keyed context patches amplifies the power of the key features by providing a means of verifying whether the key is likely to be part of a particular object. This local verification step is critical, because the invariant parameters of the key features are relatively weak evidence. If only this weak evidence is used in an evidence combination scheme, a proliferation of high-scoring false object hypotheses results. This is a well known problem with voting schemes, but can be alleviated if the voting features are sufficiently powerful.

The basic recognition strategy is to use a database (here implemented as an associative memory) of keyed context patches which is organized so that access via an unknown keyed context patch evokes associated hypotheses for the identity and configuration of all known views of objects that could have produced such context patch. These hypotheses are fed into a second associative memory, indexed by the view parameters, which lumps the hypotheses into clusters that are mutually consistent within a loose geometric framework (these clusters are the third level groups). In the current implementation, the requisite looseness is obtained by tolerating a specified deviation in the position, size, and orientation of keyed context patches relative to a nominal position.

The secondary database maintains a probabilistic estimate of the likelihood of each third level group (cluster) based on statistics about the occurrence of the keys in the primary database. The idea is similar to a multi-dimensional Hough transform without the space problems that arise in an explicit decomposition of the parameter space. In our case, since 3-D objects are represented by a set of views, the clusters represent two dimensional rigid transforms of specific views. As mentioned above, the use of keyed contexts rather than first-level groups gives the voting features sufficient power to substantially ameliorate well known problems with false positives in Hough-like voting schemes.

The output of the system is a set of third level groupings that represent hypotheses as to the identity and pose of objects in the scene, ranked by the total evidence for each hypothesis. Each hypothesis also retains pointers to the supporting context patches. At this point, it would be possible to undertake a top-down verification of the top hypotheses, making a broader search for features that should be present, but did not contribute evidence to the hypothesis (e.g. due to differing bottom-up boundary segmentation). We do not currently perform this step; however, unlike appearance-based systems based on whole-object appearance, the structure of our representation is such that this could be performed to advantage, and such a step has the potential to significantly improve the performance of the system as a whole. The results given should thus be interpreted as representing the power of an initial hypothesis generator or indexing system.

The approach has several advantages. First, because it is based on a merged percept of local contexts rather than global properties, the method works well in the presence of occlusion and background clutter, and does not require prior segmentation of the image into whole objects. This is an advantage over systems based on principal components template analysis, which are sensitive to occlusion and clutter. Second, entry of objects into the memory can be an active, automatic procedure. Essentially, the system can explore the object visually from different viewpoints, accumulating 2-D views, until it has seen enough not to confuse it with any other object in the database. This is an advantage over conventional alignment techniques, which typically require a prior 3-D model of the object. Third, the method lends itself naturally to multi-modal recognition. Because there is no single, global

structure for the model, evidence from different kinds of keys can be combined as easily as evidence from multiple keys of the same type.

3 Levels of Perceptual Grouping

As mentioned before, our system works on several levels of perceptual grouping (see Figure 1). First, pixels are grouped into contours using a stick growing algorithm. The contours are then grouped in context patches that are further grouped into the 2-D views of the object, and finally the 2-D views form the description of the 3-D object. Here we describe the levels and the grouping algorithms in more detail.

3.1 From Pixels to Contours

The contours are extracted from an image using a stick-growing method developed by Nelson [14]. In order to provide robustness and sensitivity, perceptual grouping based on proximity is used, so that extended local information is obtained. The method utilizes both gradient magnitude and direction information, and incorporates explicit lineal and end-stop terms. These terms are combined non-linearly to produce an energy landscape in which local minima correspond to lineal features that can be represented as line segments. A gradient descent process is used to find these minima. The effective result is a set of boundary fragments terminated at corners (regions of high curvature). No specific attempt is made to group these into closed contours (though closed contours may sometimes result) because the corners are perceptually significant features, and our second level of grouping does not utilize closure. This sort of boundary segmentation procedure, if not the exact method we used to carry it out is fairly common, so we will not describe it further here.

3.2 From Contours to Keyed Context Patches

The recognition technique must be based on features that are robustly extractable and pose invariant. These features should be complex enough to specify the configuration of the object and have additional parameters that can be used for indexing and matching. They must have a substantial probability of detection and be as insensitive to pose as possible (i.e. change relatively slowly as the object configuration changes). Many classical features do not satisfy these criteria. Line segments are not sufficiently complex, full object contours are not robustly extractable, and simple templates are not pose-insensitive.

Our next level of perceptual grouping resolves the conflict between feature complexity and robust detectability that 3-D object recognition systems have to face in general, by obtaining keyed context patches from contours on the basis of spatial proximity. These patches are complex enough to reduce multiple matches to a manageable level.

Since pose invariant features are hard to design, especially for 2-D projections of curved 3-D objects, we settle for pose insensitive context patches and compensate by a combination of two strategies. First, we take advantage of the statistical unlikelihood of close matches for complex patterns (another advantage of relatively complex features). Second,

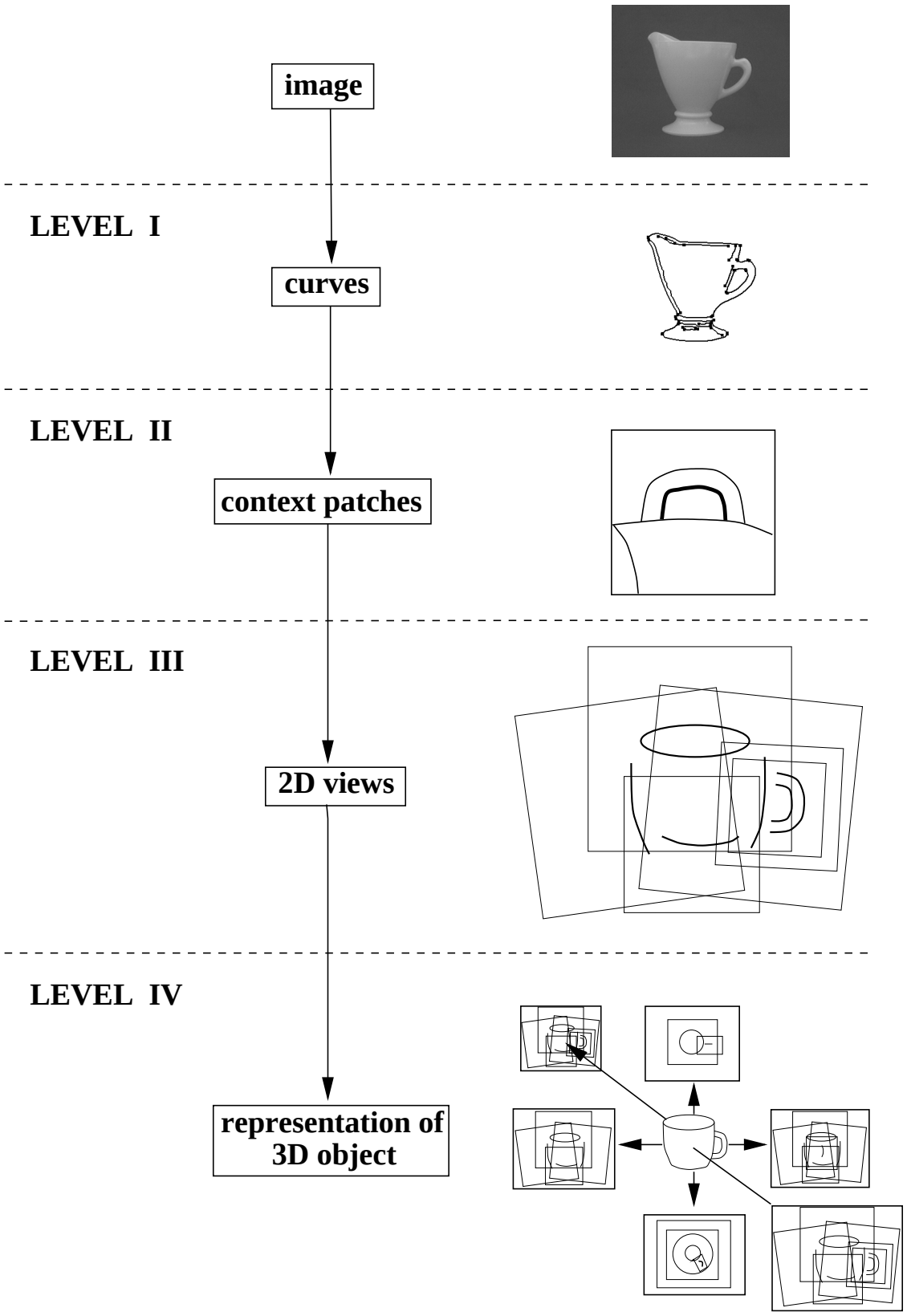

Figure 1: The perceptual grouping hierarchy

the appearance-based recognition strategy provides what amounts to multiple representations of an object in that the same physical attribute of the object may evoke several different associations as the object appears in different views. The pose insensitive nature of the features prevents this number from being too large.

The keyed context patches that form our second level groups are constructed by taking contour fragments (key curves) from the first level groups, which are probabilistically segmentable in similar views of an object, and embedding them in a local context consisting of a square image region, oriented and normalized for size by the key curve, which is placed at the center. Each keyed context patch contains a representation of all other segmented curves, key or not, that intersect it. This canonical placement of the key curve in the context patch, and consequent normalization of the context patch itself permits the planar components (translation, scaling, and plane rotation) of an object pose hypothesis to be computed directly whenever a match between a model patch and an image patch is observed.

In more detail, after the curve-finding algorithm is run on an image, we get a set of segmented contour fragments broken at points of high curvature. The longest curves are selected as key curves, and every one of these provides a seed for a context patch. The number of keys selected depends on whether we are constructing a model, in which case we use the 20-30 longest fragments, or attempting to determine what is in an image, in which case we use all contours exceeding some minimum length threshold. A fixed-size template (21 x 21) is constructed in which a base segment determined by the endpoints (or the diameter in the case of closed or nearly closed curves) of the key curve occupies a canonical position in the template. All image curves that intersect the normalized template are mapped into it with a code specifying their orientation relative to the base segment. Since the templates are of fixed size, regardless of the size of the keying curve, this is, to a certain extent, a multiple resolution representation.

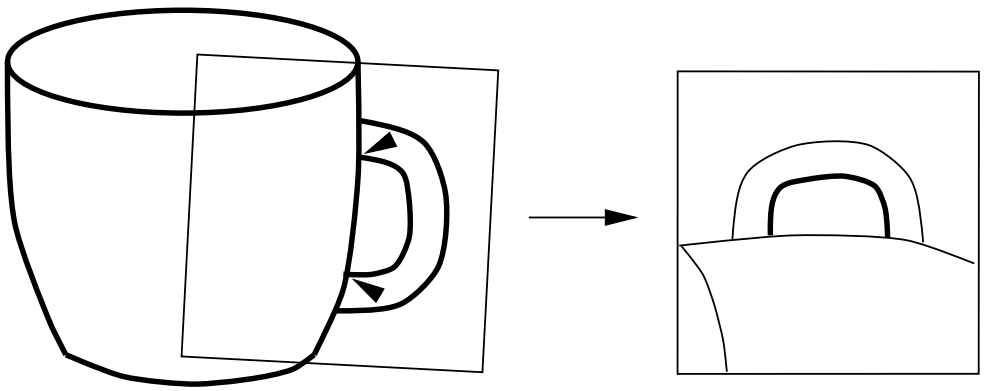

Figure 2: Example of a patch generated by a boundary fragment in a simple cup sketch. In this case the keying fragment is the inner loop of the handle, shown in canonical position in the center of the template square. The template represents not just the keying fragment, but all portions of other curves that intersect the square.

Figure 2 shows how a single context patch is generated by a boundary fragment in a simple sketch of a cup.

3.3 From context patches to 2-D views

The third level grouping provides the object recognition ability of our system. During recognition of objects in an unknown scene, context patches described are constructed for every sufficiently long curve in the image. Groups of these context patches that are mutually consistent (in a loose geometric fashion) with model groups previously constructed from views of the various objects the system knows about are assembled. This can be viewed as grouping by familiarity. The groups so constructed represent integrated hypotheses about what known objects occur in the scene, and incorporate a numerical evidence score that can be used to impose an interpretation on the scene (e.g. if we think there is just one object of interest in the scene and want to answer the question "what is it?" we grab the integrated hypothesis with the highest score).

Figure 3 shows the patches that would be generated by the indicated set of boundary fragments in the sketch. The left-hand side of the figure shows the key curves displaced, while preserving loose geometric relationships. This illustrates both the flexibility that allows the representation to deal with the distortion due to modest (20 degree) changes in viewpoint, and the implicit fragmentation that allows the representation to deal with occlusion, clutter, and missing components. Note also that the representation is redundant, and that local contexts arising from large curves may contain all or most of the curves in an object. This redundancy is important, since the output of the segmentation process may vary over the range of views that need to be covered by a particular 2-D training view, and a substantial fraction of the key fragments may not be matchable in a new view.

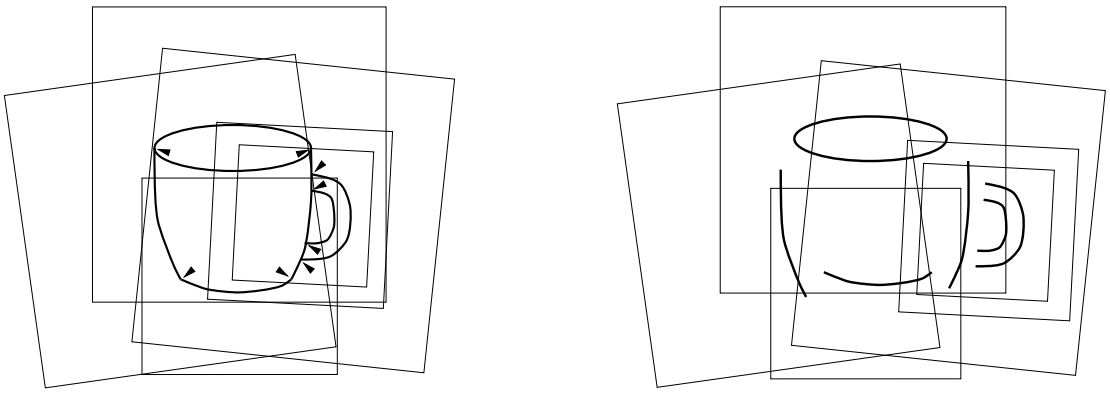

Figure 3: Right, example of patches generated by a set of boundary fragments for the cup sketch; arrows indicate the location of the fragment endpoints or diameters. Left, key fragments displaced, while preserving loose global relationships. Our representation implicitly contains this kind of distortion.

Verifying a local context match between a candidate patch keyed by a curve fragment and a stored model patch involves taking the model patch curve points and verifying that a curve point with similar orientation lies nearby in the candidate template. Essentially this amounts to loose directional correlation. The matching process is modified in that curves that lie parallel to the base segment and within half a diameter of it do not contribute to the match. The reason for this is that close parallel structure is so common in the world,

(narrow objects, shadows, highlights, steep gradient effects) that such structures contribute little evidence while adding enormously to the "accidental" match population.

The model groups (representing familiar 2D views) that are used to drive the grouping process during recognition are constructed from keyed context patches extracted from clean views of the object. These are stored in a database (indexed by 2D invariants of the key contours for efficient access) which represents the long-term memory of the system. During recognition, the input image undergoes a similar feature extraction process. The context patches obtained are matched against those stored in the long term memory, and the results used to assemble clusters that are mutually consistent with a particular model view under a particular 2D transformation (translation, scale, and rotation). The process is described in more detail below.

In order to prepare the database used for matching, we take a number of images of each object, covering the region of the viewing sphere over which the object may be encountered. The exact number of images per object may vary depending on the features that are used. In our case obtaining images at about 20 degree intervals on the viewing sphere is sufficient. Covering the sphere at this sampling requires about 100 images. Since translation, scaling, and in-plane rotatation are handled by the context patch matching process (recall that context patches are normalized by the keying curve), these 100 views give us recognition over the full 6 orthographic degrees of freedom.

For every image so obtained, the boundary extraction procedure is run, and the best 25 or so boundaries are selected as keys, from which patches are generated and stored in the database. With each patch is associated the identity of the object that produced it, the viewpoint it was taken from, and three geometric parameters specifying the 2-D size, location, and orientation of the image of the object relative to the key curve. As mentioned above, this information permits a hypothesis about the identity, viewpoint, size, location and orientation of an object to be made from any image match to the model context patch.

The basic recognition procedure consists of four steps. First, keyed context patches are extracted from the image using the first and second level grouping processes. In the second step, these keyed context patches are used to access the database memory and retrieve information about what objects could have produced them, and in what relative configuration. The third step groups the information got from matching process to produce integrated hypotheses about the identity and configuration of potential objects. This is the third level process in our perceptual grouping hierarchy, where grouping is achieved through familiarity.

In more detail, each context patch from the image may match zero, one or more context patches from view models in the database (long term memory). Each such match generates a hypothesis about the identity and pose of an object that could have produced it (actually a 2D configuration of a 2D view). Patches that are consistent with the same view model in the same configuration form a group that accumulates evidence for that configuration and view model. The object identity and pose specification parameters are used as indices to achieve efficient access into a second associative memory, where the evidence accumulated. After all features have been so processed, the hypothesis corresponding to the group with the highest evidence score is selected. Secondary hypotheses can also be reported.

In the final step described above, an important issue is the method of combining evidence. The simplest technique is to use an elementary voting scheme - each piece of evidence contributes equally to the total. This is clearly not well founded, as a feature that occurs in

many different situations is not as good an indicator of the presence of an object as one that is unique to it. For example, with 24 3-D objects stored in the database, comprising over 30,000 patches, we find that some image features match 1000 or more database features, while others match only one or two. An evidence scheme that takes this into account would probably display improved performance. An obvious approach in our case is to use statistics computed over the information contained in the associative memory to evaluate the quality of a piece of information. It is clear that the optimal quality measure, which would rely on the full joint probability distribution over keys, objects and configurations, is infeasible to compute, and thus we must use some approximation. What we do is to use the first order feature frequency distribution over the entire database and do a Bayesian maximum likelihood evidence combination based on it.

One last step that we do not take in the current system is whole-object verification of the top hypotheses. Unlike appearance-based systems based on whole-object appearance, the structure of our representation is such that this could be performed to advantage, and such a step has the potential to significantly improve the performance of the system as a whole. The results given should thus be interpreted as representing the power of an initial hypothesis generator or indexing system.

3.4 A Connection to Artistic Perception

This representation of views, based on local context patches, reminds us of cubist drawings [15] where fragmentary, but distinctive parts appear in a local context. Similarly to our system, in cubism not every piece is present, and generally, not all the pieces present are correctly parsed. Also, there is frequent duplication of contextual features. The vision scientist is struck by the fact that the appearance of fragmentary but distinctive parts that serve to key the percept (such as the sound holes, partial profile, and scroll of a violin). These are often accompanied by other features that, though not particularly distinctive alone, in the local context established by distinctive keys become meaningful and tend to verify an overall impression (see Figure 4).

Cubism includes the most important problems of machine vision: clutter, mis-labeling, missing parts and geometric distortion.

3.5 From 2-D Views to the 3-D Object

As mentioned before, the database used for recognition is constructed from views of the object taken around the whole viewing sphere. The group of these views, together with the low-level groupings composing them form a characterization of the 3D object which displays a remarkable consistency with some recent psychophysical results, and promises to be useful for 3D reasoning about manipulation of objects.

Grouping at this fourth level of our hierarchy, like the grouping at the second level, is based on proximity. However, rather than proximity in image space, the proximity is in view or manipulation space. This produces a topological organization of the views, which can described through a neighborhood relationship.

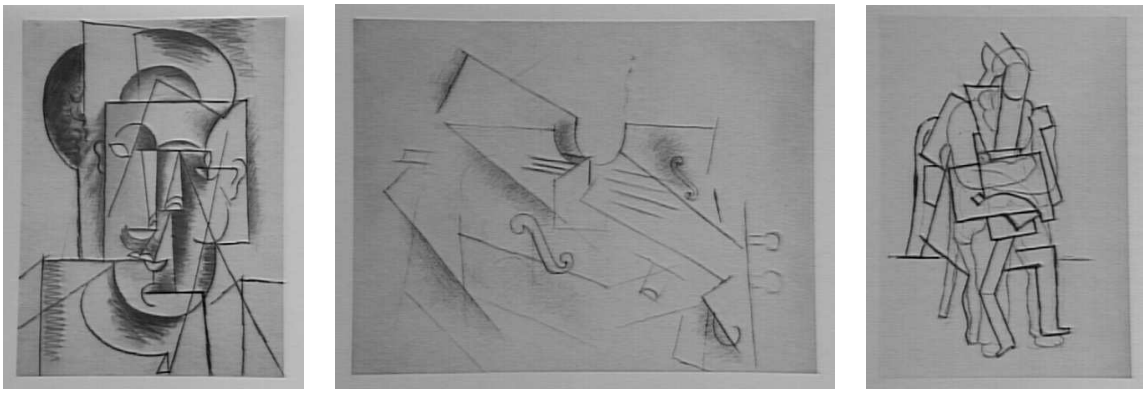

Figure 4: Cubist drawings illustrating use of fragmented linear features and suggesting loosely organized local context frames organized about distinctive key features. Right, Picasso: *Head of a Man*, 1912; Center, Braque: *Violin*, 1912; Left, Picasso: *Seated Man*, 1914;

3.5.1 Psychophysical Relevance

This top-level representation is consistent with the representation of 3-D objects in the human brain. Some early psychophysical work addressed the problem of mental rotation of images of 3-D objects, and determined that people were, in general able to do this, and in a way that took increasing amounts of time as the required rotation was increased (Shepard and Cooper 1982, Tarr and Pinker 1989) [19, 20]. This was taken as evidence for the existence of internal 3-D object models. More recent work, however, while confirming that people are indeed able to perform mental operations that seem most consistent with the existence of 3-D, object-centered representations, has raised questions about whether these representations are what is used for fast recognition (Bulthoff et al. 1995, Edelman and Bulthoff 1992) [1, 2]. It can be plausibly argued that the 3-D representations are used for example, for planning manipulations, while fast recognition uses a separate representation.

Based on psychophysical experiments, as well as on experiments in which they looked at the expected performance of several representations used in 3-D object recognition, Bulthoff, Edelmann and Tarr [1] argue that the representation most similar to the one used by the human visual system is the viewpoint dependent two-dimensional representation. Methods using this representation try to achieve object constancy by storing multiple 2-D viewpoint-specific representations and using mechanisms for matching input images to stored views or to views derived computationally from stored views.

When presented with a new image, our method looks for the stored view that is most similar to the image. This can be thought of as an interpolation of stored views. A lower error rate is obtained for familiar test views than for novel test views, depending on the distance from the novel view to the nearest familiar stored view. By storing a large number of views that are placed at modest (20 degree) distances from each other we managed to obtain a very low error rate even in the case of novel views.

The human visual model advocated by Bulthoff et al. is similar to our system in that it represents objects by small sets of canonical views and uses a variant of mental rotation to recognize objects at attitudes other than the canonical ones. Each canonical view is

essentially an image-based representation of the object as it is seen from a certain viewpoint and might be augmented by limited depth information. Their experiments showed that even in the case of human observers, generalization to novel views was severely limited, with performance dropping to chance levels at a misorientation of about 40° relative to familiar views (Edelman and Bulthoff 1992) [3].

In this human visual model, as in certain computational models, e.g. Edelman and Weinshall (1991) [4], views that "belong" together are more closely associated with each other. Computationally, this method of recognition is analogous to an attempt to express the input as an interpolation of the stored views, and it can also be viewed as a perceptual organization at a higher level. In this case, recognition normally requires neither 3-D reconstruction of the stimulus, nor the maintenance of a library of 3-D models of objects. Instead, information sufficient for recognition can be found directly in the 2-D image locations of object features.

In psychological experiments there are several levels of category organization in recognition performance. The *basic level* is the most salient according to psychological criteria. The *entry level* is the first categorical label generally assigned to a given object. Objects whose recognition implies finer distinctions than those required for entry-level categorization are said to belong to a *subordinate level*. The patterns of response times and error rates in recognition experiments are influenced by the category level at which the distinction between the different stimuli is to be made. Error rates and response times are viewpoint invariant for classification (determining the entry-level category) and viewpoint-dependent for indentification.

From this point of view, our experiments deal with the entry-level classification of objects. When the objects present some similarities, the categorization needs to be done on the subordinate level and the error level gets higher. For example in the generic experiments we sometimes saw confusion between planes and fighter jets.

3.5.2 Machine Uses of the Fourth Level

At the moment the recognition process uses only groups up to the third level in the hierarchy. We could use the fourth level to enhance recognition such that it is able to retrieve the exact position and orientation of objects with respect to each other for manipulation purposes.

Grouping the views of the same object gives us the idea of a possible extension that could be implemented in our system. If the recognition of an object yields a low score, the system could analyze other views of the same object, using an active vision system where a low recognition score would trigger the robot to rotate the object with a specific angle and acquire a new view of it. If the result obtained from this view proves to be consistent with the first one, then the recognition is finished. Otherwise further hypotheses have to be considered.

This method is again consistent with the way the human brain operates. If a person doesn't recognize a given object, then it rotates it and looks at different views trying to use this additional information in the recognition process.

More importantly, the fourth level forms the basis for interacting with the object. For example, suppose we need to move an object into a configuration where certain currently hidden features are visible. If we know our current view of the object, and we know a view where the points of interest are visible, then by running a search through the view-space

topology we can efficiently find a path through the view-space from our current location to the desired view. We can then actively servo the camera through the view space to a goal position. Such servoing can be performed even in the absence of a prior model of the effects of manipulation, though it may initially be slow. The servoing operation can be performed much more efficiently if, in addition to the neighborhood information, we also store a local direction to the neighbor, producing a directional topology.

As another example, suppose that for some operation, the hand of a robot needs to be in a particular orientation with respect to the object. In a view-based system, the easiest way to indicate this is to show the system the correct alignment. We don't want the user to do this for all views. Instead, we show the correct alignment for one view, and use the orientation information available to check if any other view represents the same relative alignment, and if not, what reorientation is needed to obtain alignment. If desired, this reorientation can be represented as a path in view space using the view topology. We can also check whether a particular reorientation is feasible - just get hand close to where the action will be taking place, then check whether the change needed to align is within the robot's range.

The perceptual organization hierarchy that is used in our system could be taken even further, and different objects could be grouped together on the basis of their functionality, or using any other criteria to obtain a fifth level (for instance the buckle on a seat-belt on the seat in a car); such higher level grouping might even be useful for some applications. This is a subject for further thought, however, and is more in line with classical AI representations (except it is grounded) than with traditional perceptual grouping.

4 Experiments

The following section is intended to give an idea as to just how well a system based on the ideas described above can work. To our knowledge, the performance with general objects in the presence of clutter and occlusion is the best reported in the literature (as of 1997).

One measure of the performance of an object recognition system is how the performance changes as the number of classes increases. To test this, we obtained test and training images for a number of objects, and built 3-D recognition databases using different numbers of objects "different" from each other in that they were easy for people to distinguish on the basis of shape. Data was acquired for 24 different objects (34 hemispheres because some objects were either unrealistic or painted flat black on the bottom)(see Figure 5).

Clean image data was obtained automatically using a combination of a robot-mounted camera, and a computer controlled turntable covered in black velvet. Training data consisted of 53 images per hemisphere, spread fairly uniformly, with approximately 20 degrees between neighboring views. The test data consisted of 24 images per hemisphere, positioned in between the training views, and taken under the same good conditions. Note that this is essentially a test of invariance under out-of-plane rotations, the most difficult of the 6 orthographic freedoms. The planar invariances are guaranteed by the representation, once above the level of feature extraction, and experiments testing this have shown no degradation due to translation, rotation, and scaling up to 50%. Larger changes in scale have been accommodated using a multi-resolution feature finder, which gives us 4 or 5 octaves at the cost of doubling the size of the database.

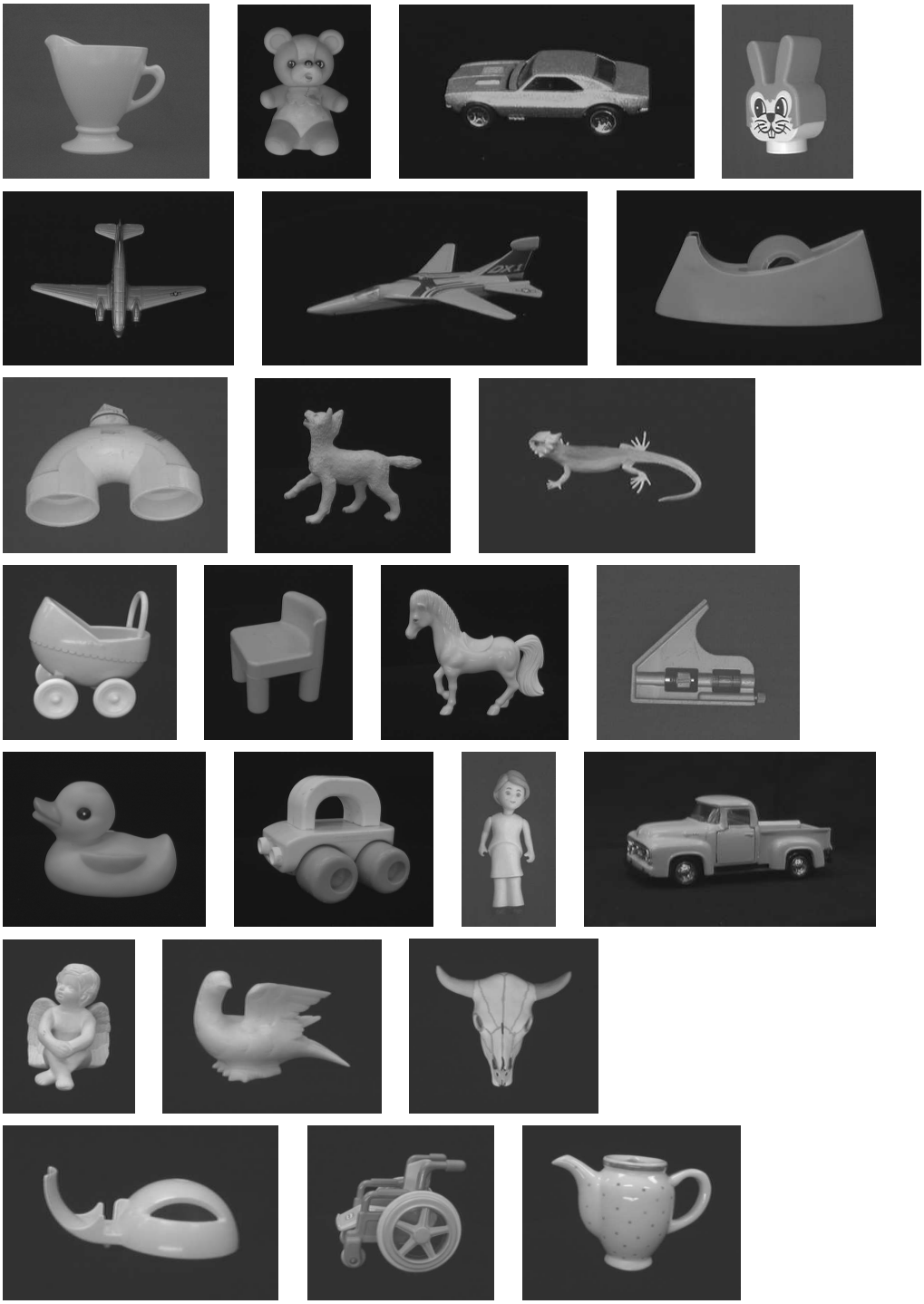

Figure 5: The set of objects used in testing the system

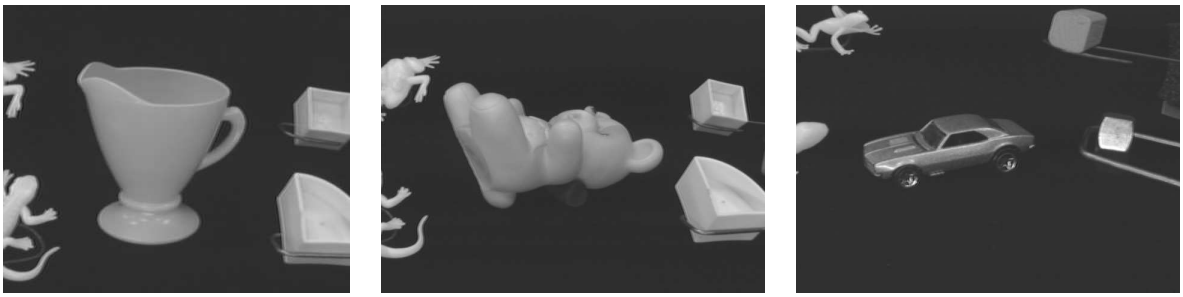

Figure 6: Examples of test images with modest dark-field clutter

We ran tests with databases built for 6, 12, 18 and 24 objects, shown in Figure 5, and obtained overall success rates (correct classification on forced choice) of 99.6%, 98.7% 97.4% and 97.0% respectively. (To find out which objects are in which database, just count the images left to right, top to bottom.) The results are summarized in the following table. The worst cases were the horse and the wolf in the 24 object test, with 19/24 and 20/24 correct respectively. On inspection, some of these pictures were difficult for human subjects. None of the other examples had more than 2 misses out of the 24 (hemisphere) or 48 (full sphere) test cases. Overall, the performance is fairly good. In fact, we believe this represents the best results presented anywhere for this sort of problem.

no. of objects	no. of hemispheres	no. of images	no. correct	% correct
6	11	264	263	99.6
12	18	408	403	98.7
18	26	576	561	97.4
24	34	768	745	97.0

Table 1: Performance of forced-choice recognition for databases of different sizes

4.1 Performance in the Presence of Clutter

The feature-based nature of the algorithm provides some immunity to the presence of clutter in the scene, in contrast to appearance-based schemes that use the structure of the full object, and require good global segmentation. For modest dark-field clutter, the method is quite robust. To test this, we acquired test sets of the six objects used in the previous 6-object case in the presence of non-occluding clutter. Examples of the test images are shown in Figure 6. Out of 264 test cases, 252 were classified correctly which gives a recognition rate of about 96%, compared to 99% for uncluttered test images. A confusion matrix is shown in Table 2. We obtained a recognition rate above 90% even in the case of more difficult clutter caused by textured backgrounds (see Figure 7).

The recognition rate stays above 90% even in the case of simple occlusion. Many of the objects are sufficiently complex that they can be chopped in half, and still recognized by the system (see Figure 8). Our system does not handle extreme occlusion, or occultation by multiple narrow objects (e.g. tree branches) well, because too many of the first-level key

class	index	smpls	0	1	2	3	4	5
cup	0	48	47	0	1	0	0	0
bear	1	48	2	46	0	0	0	0
car	2	24	0	0	24	0	0	0
rabbit	3	48	0	0	1	47	0	0
plane	4	48	0	0	2	1	45	0
fighter	5	48	0	0	1	0	4	43
Totals			49	46	29	48	49	43

Table 2: Error matrix for object classification experiment with clutter.

Figure 7: Examples of manageable images with textured backgrounds

features are broken up. This may be an instance where an additional low-level grouping process (between levels one and two) could help put broken key contours back together.

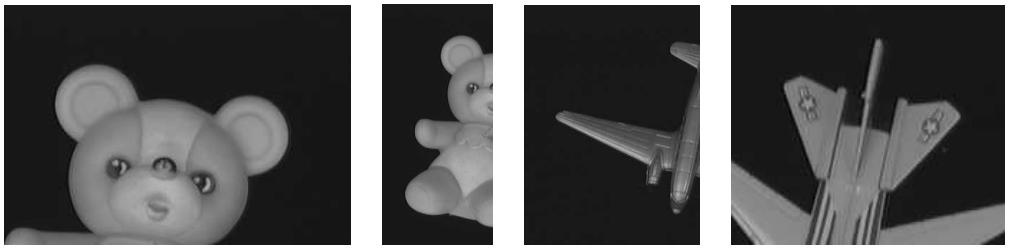

Figure 8: Examples of manageable occluded images

To demonstrate that the clutter resistance is not dependent on whole-object segmentability, we took a number of individual pictures of known objects with adjacent and partially overlapping distractors (moderate clutter, minor occlusion). Figure 9 shows some pictures containing objects from the database that are not trivially segmentable examples where the system correctly answered the question "what is this?". It is hard to quantify the performance in this case, because it is not easy to generate hundreds of test cases of "comparable" difficulty. From our experiments, accuracy on images with 50%-75% clutter and 25% occlusion seems to be around 90% and this number is supported by the statistical framework that we present below. Up to 50% occlusion seems manageable if there are few distractors.

Figure 9: Examples of manageable images with adjacent and slightly occluding clutter

5 Conclusions and Future Work

We have described a framework for keyed appearance-based 3-D recognition, based on a perceptual grouping hierarchy. By doing perceptual grouping at four levels, we avoid some of the problems of previous appearance-based schemes, which did grouping at either only high level, or only low-level. We ran various large-scale tests and found good performance for full-sphere recognition of up to 24 complex, curved objects, robustness against clutter.

Future plans include adding enough additional objects to push the performance below 75%, both to better observe the functional form of the error dependence on scale, and to provide a basis for substantial improvement. We also want to see how the performance can be improved by adding a final verification stage, since we have observed that even when the system provides the wrong answer, the "right" one is generally in the top few hypotheses. In another direction, we have some preliminary results indicating that the system, when coupled with a simple memory-constraint protocol, functions very well for finding particular objects in large, highly cluttered scenes. We plan to gather enough data for this problem to generate statistically significant performance data. Finally, we want to experiment with adapting the system to allow fine discrimination of similar objects (same generic class) using directed processing driven by the generic classification.

References

[1] H. H. Bulthoff, S. Y. Edelman, and M. J. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5(3):247–260, 1995.

[2] S. Edelman and H. H Bulthoff. Modeling human visual object recognition. In *Int. Joint Conference on Neural Networks (IJCNN92)*, Baltimore, MD, June 7-11 1992.

[3] S. Edelman and H. H Bulthoff. Modeling human visual object recognition. In *International Joint Conference on Neural Networks (IJCNN92)*, pages 37–42, Baltimore, MD, June 7-11 1992.

[4] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3d objects. *Biological Cybernetics*, 64:209–219, 1991.

[5] W. E. L Grimson. *Object Recognition by Computer: The role of geometric constraints.* The MIT Press, Cambridge, 1990.

[6] W. E. L. Grimson and Danial P. Huttenlocher. On the sensitivity of the hough transform for object recognition. *IEEE PAMI*, 12(3):255–274, 1990.

[7] P. Havalder, G. Medioni, and F. Stein. Percetual grouping for generic recognition. *Internation Journal of Computer Vision*, 20(1-2):59–80, October 1996.

[8] Chien-Yuan Huang and Octavia I. Camps. Object recognition using appearance-based parts and relations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR97)*, pages 877–883, San Juan, Puerto Rico, June 1997.

[9] Daniel P. Huttenlocher and Shimon Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[10] Y. Lamdan and H. J. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *Proc. International Conference on Computer Vision*, pages 238–249, Tampa FL, December 1988.

[11] David G. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Boston, MA, 1986.

[12] David G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[13] H. Murase and S. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995.

[14] Randal. C. Nelson. Finding line segments by stick growing. *IEEE Trans PAMI*, 16(5):519–523, May 1994.

[15] Randal C. Nelson and Andrea Selinger. A cubist approach to object recognition. In *Proc. Int. Conference on Computer Vision (ICCV98)*, Bombay, India, January 1998.

[16] Thomaso Poggio and Shimon Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[17] R. Rao and D. Ballard. An active vision architecture based on iconi representations. *Artificial Intelligence*, 78:461–505, March 1995.

[18] C. Schmid and R. Mohr. Combining greyvalue invariants with local constraints for object recognition. In *Proc. CVPR96*, pages 872–877, San Francisco CA, June 1996.

[19] R. N. Shepard and L. A. Cooper. *Mental Images and Their Transformations.* MIT Press, Cambridge, MA, 1982.

[20] M. J. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[21] M. Wertheimer. Principles of perceptual organization. In D. Beardslee and M. Wertheimer, editors, *Readings in Perception*. 1958.