

A Cubist approach to Object Recognition *

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

Abstract

We describe an appearance-based object recognition system using a keyed, multi-level context representation reminiscent of certain aspects of cubist art. Specifically, we utilize distinctive intermediate-level features in this case automatically extracted 2-D boundary fragments, as keys, which are then verified within a local context, and assembled within a loose global context to evoke an overall percept. This system demonstrates good recognition of a variety of 3-D shapes, ranging from sports cars and fighter planes to snakes and lizards with full orthographic invariance. We report the results of large-scale tests, involving over 2000 separate test images, that evaluate performance with increasing number of items in the database, in the presence of clutter, background change, and occlusion, and also the results of some generic classification experiments where the system is tested on objects never previously seen or modeled. To our knowledge, the results we report are the best in the literature for full-sphere tests of general shapes with occlusion and clutter resistance.

Key Words: Object recognition, Appearance-based representations, Visual learning.

1 Introduction

In the late 19th and early 20th centuries certain European schools of art made a deliberate and dramatic move away from the notions of photographic realism that had been popular in previous centuries. There were a variety of reasons for this - one may simply have been that the invention of photography made realism trivial. In any case, a number of increasingly abstract movements emerged, which challenged classical notions of spatial representation. Although these artists were not looking for a scientific model of human perception per-se, they pushed a lot of boundaries in terms of discovering what sort of informa-

tion was necessary to evoke visual perception. This work is often interesting from machine vision standpoint because low-level veridical cues are deliberately corrupted. The ways in which the remaining chunks function, can be quite suggestive about the higher level organization operating in recognition

Particularly interesting from this standpoint is the cubist movement, pioneered by Picasso and Braque, mostly over a 6-year period between 1908 and 1914. As with most artistic movements, cubism was complex in its genesis, involving evolution and revolution around previous artistic traditions, inside jokes, personal rivalries etc. However, a central theme involved pushing the limits of the human ability to synthesize perceptions of individual objects from a general sense of the overall relations between parts. The operative term above is "general" - the focus on loose relationships and suggested shape rather than photographic exactness. Within this framework, cubism explored a number of issues, including the materialization of form from ambiguous cues, the fragmentation of primary percepts into suggestive pieces, and the use of space, to provide local and global context. Coincidentally, or perhaps not so coincidentally, these same basic issues lie at the heart of a lot of work on machine vision.

Arising initially from abstracted landscape, portraiture, and still life, cubism evolved to the representation of single isolated percepts. Some of these later works made almost exclusive use of fragmentary linear features, stripped of texture, color, and shading, as well as coherent shape (see Figure 1). It is these works, stripped of the baggage of detail, as it were, that are most suggestive from a machine vision perspective.

Looking at these drawings as a vision scientist one is struck by several aspects. The first is the appearance of fragmentary but distinctive parts that serve to key the percept (e.g. the sound holes, partial profile, and scroll of a violin) These are often accompanied by other features that, though not particularly

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

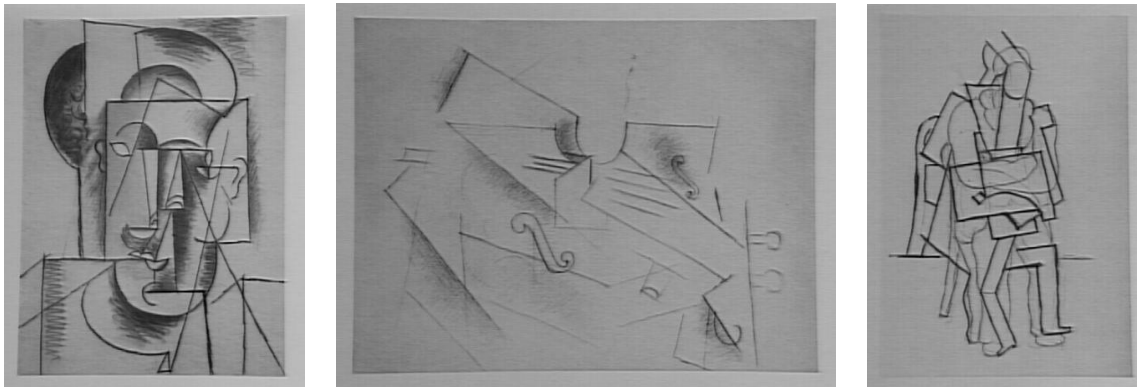


Figure 1: Later Cubist drawings illustrating use of fragmented linear features, and suggesting loosely organized local context frames organized about distinctive key features. Right, Picasso: *Head of a Man*, 1912; Center, Braque: *Violin*, 1912; Left, Picasso: *Seated Man*, 1914;

distinctive alone, in the local context established by distinctive keys become meaningful and tend to verify an overall impression. Such local spatial frames are sometimes indicated explicitly - frequently with the rectangular regions that popularly stereotype cubist art. The spatial organization both within and between local contexts is loose, violating geometry, and sometimes topology as well; but the whole is still organized globally by the human observer. In addition, not every piece is present, and generally, not all the pieces present are correctly parsed; there is frequent duplication of contextual features (and even key features, though this is rare). Basically, whatever the cubist representation is, it not only tolerates, but includes as an essential aspect a huge amount of clutter, mis-labelling, missing parts, geometric distortion - basically, all the problems that plague machine vision systems.

The point is, we think that the way cubism invites and then handles the above multitude of perceptual problems is highly suggestive from the standpoint of machine perception. To elaborate, the fragmented nature and loosely specified spatial contexts of cubist art suggest a representation that turns out to be extremely useful for dealing with the problems of distortion, clutter, and missing information that typically plague machine recognition systems. In particular, the idea of using distinctive key features, enhanced by local context, and assembled in a loose global context to form an overall percept turns out to be extremely powerful.

In this paper we describe an object recognition system based on such an Cubist organization, and present experiments on large databases of 3D objects. To our knowledge these represent the best reported results

for full-sphere recognition of general shapes with occlusion and clutter resistance.

2 Background

Object recognition is probably the most researched area of computer vision, and we hit only highlights in this summary. Much work to date has used model-based systems, e.g [6, 5, 3]. The 3D geometric models on which these systems are based are both their strength and their weakness. [4]. On the one hand, explicit models provide a framework that allows powerful geometric constraints to be utilized to good effect. On the other, model schemas are generally severely limited in the sort of objects that they can represent, and obtaining the models is typically a difficult and time-consuming process.

Appearance-based object recognition methods have been proposed in order to make recognition systems more robust, and more easily trainable from visual data. In recent work, Poggio has recognized wire objects and faces [1]. Murase and Nayar [8] find the major principal components of an image dataset, and use the projections of unknown images onto these as indices into a recognition memory. Rao and Ballard [10] describe a similar approach using steerable filters. Mel [7] uses multiple low-level cues (e.g. color). Schmid and Mohr [12] 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 a much simpler evidence combination scheme.

3 The Method

3.1 Overview

Our basic (cubist) idea is to represent the visual appearance of an object as a loosely structured com-

bination of a number of local context regions keyed by distinctive features. A local context region can be thought of as an image patch surrounding the key feature and containing a representation of other features that intersect the patch. The key features provide parameters for indexing, while verification of local context amplifies the statistical power of the features. This local verification step is critical, because the invariant parameters of the key features are relatively weak evidence, leading to a proliferation of high-scoring false hypotheses if used alone. A Hough-like evidence combination scheme provides the loose global structure that allows evidence from individual context regions to be flexibly combined. The hope is that, although changing environmental conditions (e.g. lighting, background, changes in orientation, occlusion etc.) may disrupt the detection of many of the key features, the combined evidence provided by the ones that are found will be sufficient to identify objects in the scene.

The approach has two main advantages. First, because it is based on a merged percept of local contexts rather than global properties, the method is robust to occlusion and background clutter, and does not require prior global segmentation. 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 mix it up with any other object it knows about. This is an advantage over conventional alignment techniques, which typically require a prior 3-D model of the object.

One step that we do not take in the current system is whole-object verification of the highest-scoring hypotheses. Unlike appearance-based systems using whole-object appearance, the structure of our representation is such that such verification 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.

3.2 Representation: Key Features and Local Context

The recognition technique is based on the assumption that robustly extractable, semi-invariant key features can be efficiently recovered from image data. We currently make use of a single key feature type consisting of robust boundary fragments (curves). These

fragments are placed in a local *context patch* consisting of a fixed size template (21x21), oriented and normalized for size by the key curve, which is placed at the center. All image curves that intersect the normalized template are mapped into it with a code specifying their orientation relative to the base segment. Two, 2-D curve invariants, compactness and total curvature, provide indexing into the main database. Verification in local context is performed using a distance transform coupled with directional correlation, which provides local flexibility. Close parallel structure is suppressed.

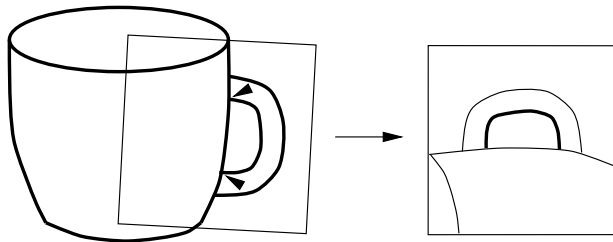


Figure 2: Example of a patch generated by a boundary fragment in a simple cup sketch. In this case the keying fragment is the inner loop of the handle, shown in canonical position in the center of the template square. The template represents not just the keying fragment, but all portions of other curves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

Figure 2 shows show a single patch context is generated by a boundary fragment in a simple sketch of a cup. 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, cubist style, while preserving loose global relationships. This illustrates the sort of fragmentation that is implicit in our representation. Note 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.

Referring back to Figure 1 we see examples of cubist art exhibiting similarly displaced fragmentary features, and additional lines that suggest (rectangular) local contextual frames. Not all of these frames are *centered* on distinctive features, but they are clearly organized around them. The leftmost image is interesting in that it represents an intermediate exper-

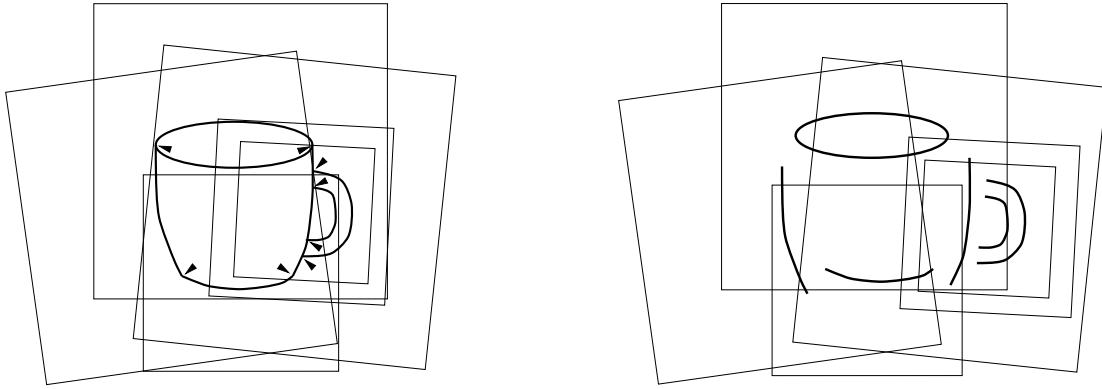


Figure 3: Right, example of patches generated by a set of boundary fragments for the cup sketch; arrows indicate the location of the fragment endpoints or diameters. Left, key fragments displaced, cubist style, while preserving loose global relationships. Our representation implicitly contains this kind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

iment, with a veridical sketch visible underlying abstracted features and local frames. We will leave it to the experts to argue what the artists actually intended, and note only that to this vision scientist, these pictures strongly suggest a representation that has proven to be effective for dealing with the problems of object recognition.

3.3 Overall Recognition Procedure

In order to recognize objects, we first must build a database. We take a number of images of each object, covering the region of interest on the viewing sphere. For our patch features, sampling every 20 degrees is sufficient. To cover the entire sphere at this sampling requires about 100 images. For every image so obtained, the boundary extraction procedure is run, and the best 20 or so boundaries are selected as keys, from which patches are generated and stored in the database. With each context 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. This information permits a hypothesis about the identity, viewpoint, size, location and orientation of an object to be made from any match to the patch feature.

The recognition procedure consists of three steps. First, potential key features and local contexts are extracted from the target image. In the second step, these keys are used to access the database memory, retrieve match information, and generate hypotheses about the identity and configuration of the objects that could have produced them. In the third step, loosely consistent groupings of these “pose” hypotheses are identified. This integration is performed by us-

ing the pose hypotheses themselves as keys into a second associative memory, where evidence for the various global hypotheses is accumulated. Specifically, any global hypotheses in the secondary memory that are consistent (in our loose sense) with a new hypothesis have the associated evidence updated. If no pre-existing global hypothesis matches, a new one is generated. After all features have been so processed, the global hypothesis with the highest evidence score is selected. Secondary hypotheses can also be reported.

3.4 Global Context and Evidence Combination

In the final step described above, an important issue is the method of combining evidence within a loose global context. An elementary voting scheme is clearly not optimal, 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. On the other hand, 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.

We take an intermediate approach and use the first order feature frequency distribution over the entire database in a Bayesian framework, where “feature” means the entire key curve plus local context, since this is what is being matched. The actual algorithm is to accumulate evidence, for each match supporting a pose, proportional to $F \log(k/m)$ where m is the number of matches to the image feature in the whole database, and k is a proportionality constant that attempts to make m/k represent the actual geometric probability that some image feature matches a particular patch in the pose model by accident. F represents an additional empirical factor proportional

to the square root of the size of the feature in the image, and the 4th root of the number of key features in the model. These modifications capture certain aspects, namely the importance of feature size, and the importance of a simple description, that seem relevant to the recognition process, but are difficult to model using Bayesian probability.

It can be shown formally that maximizing the summed log terms in the above formula is equivalent to Bayesian MAP reasoning using the match frequency as an estimate of the prior probability of the feature (including local context), and assuming independence of observations. Because the independence assumption is not entirely valid in the real world, the log evidence values actually obtained are serious underestimates if interpreted as actual probabilities. However, the rank ordering of the values, which is all that is important for classification, is fairly robust to distortion due to this independence assumption.

3.5 Implementation

Using the principles described above, we implemented a recognition system for rigid 3-D objects. The system needs a characteristic shape or pattern to index on, and does not work well for objects whose character is statistical, such as generic trees or pine cones. Component boundaries were extracted by modifying a stick-growing method for finding segments developed recently at Rochester [9] so that it could follow curved boundaries. The system is trained using images taken approximately every 20 degrees around the sphere, amounting to about 100 views for a full sphere, and 50 for a hemisphere.

Figure 4 illustrates the operation of the recognition system on an image of a cup from the test set. The boundary extraction system finds 15 curves in the image; of these, 5 key patches contribute to the best hypothesis (which happens to be the “correct” answer in all the experiments where this image was used). This image illustrates several of the problems that make matching key curves a probabilistic process: boundaries that wash out, ambiguous “corners”, boundaries due to highlights, and boundaries produced by shading effects. However, there is enough repeatability so that the process works.

4 Experiments

4.1 Variation in Performance with Size of Database

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



Figure 4: Operation of the recognition system: 1) Image of a cup, 2) boundaries extracted, and 3) curves which keyed matching patches.

built 3-D recognition databases using different numbers of objects. The objects used were chosen to be “different” in that they were easy for people to distinguish on the basis of shape. Data was acquired for 24 different objects and 34 hemispheres. Examples of the objects are shown in Figure 5. The number of hemispheres is not equal to twice the number of objects because a number of the objects were either unrealistic or painted flat black on the bottom which made getting training data against a black background difficult.

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.

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. The results are summarized in the following table. The worst cases were a horse and a wolf in the 24 object test, with 19/24 and 20/24 correct respectively. 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. A naive estimate



Figure 5: Some of the objects used in testing the system.

num. of objects	num. of hemi-spheres	num. of test images	num. correct	percent 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

of the theoretical error trends in this sort of matching system would lead us to expect a linear increase in the error rates as the size of the database increased (best-case). Our results are consistent with this, though we don't have enough data points to provide convincing support for a linear trend.

The resource requirements are high, but scale more or less linearly with the size of the database. Memory use is about 3Mbytes per database hemisphere, and indexing time is about 3 seconds per hemisphere on a 160MHz Ultrasparc. Both the memory use and the indexing could probably be substantially improved with various standard techniques, and the whole process is efficiently parallelizable.

4.2 Performance with clutter and occlusion

The feature-based nature of the algorithm provides some immunity to the presence of clutter and occlusion in the scene; this, in fact, was one of the design goals. This is in contrast to appearance-based schemes that use the structure of the full object, and require good prior segmentation. Our algorithm, in fact, seems reasonably robust against modest dark-field clutter in high quality images, that is, extra objects or parts thereof in the same image as the object of interest. We ran a series of tests where 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.

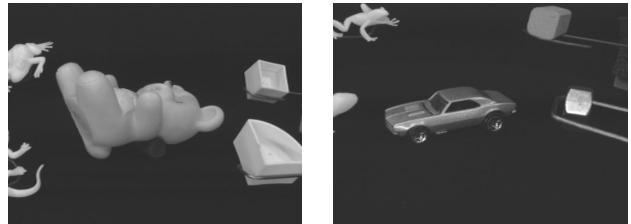


Figure 6: Examples of test images with modest dark-field clutter

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 classification with dark-field clutter.

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 7 shows some examples from the 6 object database where the system correctly answered the question "what is this?". These pictures are not trivially segmentable, but on the other hand it is not easy, as in the previous cases, to automatically generate hundreds of test cases of "comparable" difficulty over the full test sphere. It thus is hard to quantify performance but accuracy with images of this "difficulty" (50%-75% clutter, 25% occlusion) seems to be somewhere around 90%. Some recent results from a new statistical model of performance support this number.



Figure 7: Examples of manageable images with adjacent and slightly occluding clutter

4.3 Experiments on “Generic” Recognition

This set of experiments was suggested when, on a whim, we tried showing our coffee mugs to an early version of the system that had been trained on the creamer cup in the previous database (among other objects), and noticed that even though the creamer is not a very typical mug, the system was making the “correct” generic call a significant percentage of the time. Moreover, the features that were keying the classification were the “right” ones, i.e., boundaries derived from the handle, and the circular sections, even though there was no explicit part model of a cup in the system. Though the notion of generic visual classes is ill defined scientifically, the generalization to different objects in the same “human class” was suggestive.

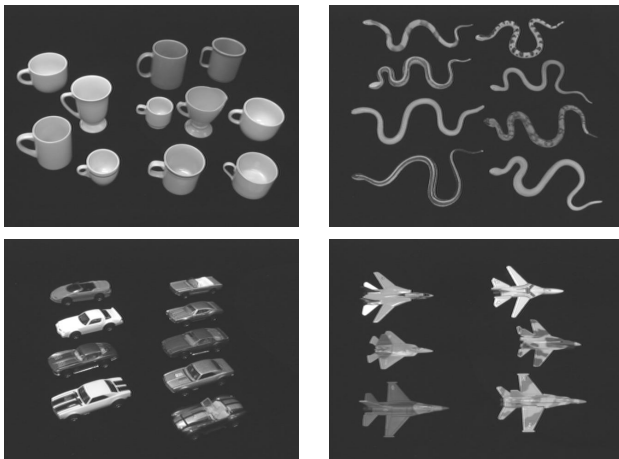


Figure 8: Test sets used in generic recognition experiment. The training objects are on the left side of each image (4 cups, 3 planes, 3 fighters, 4 cars, 4 snakes) and the test objects are on the right.

For the experiment, gathered multiple examples of objects from five generic classes, (11 cups, 6 “normal” airplanes, 6 fighter jets, 9 sports cars, and 8 snakes).

The recognition system was trained on a subset of each class, and tested on the remaining elements. The training sets consisted of 4 cups, 3 airplanes, 3 jet fighters, 4 sports cars, and 4 snakes. Four of these classes are shown in Figure 8, with the training objects on the left of each picture, and the test objects on the right. The cups, planes, and fighter jets were sampled over the full sphere; the cars and snakes over the top hemisphere (the bottom sides were not realistically sculpted). Overall performance on forced choice classification for 792 test images was 737 correct, or 93.0%, with the best performance on the cups (98%) and the worst for the fighter planes (83%), some of which were camouflaged.

class	index	smples	0	1	2	3	4
cup	0	288	282	0	6	0	0
fighter	1	144	0	120	7	16	1
snake	2	96	5	0	88	1	2
plane	3	144	0	2	7	135	0
car	4	120	1	0	6	1	112
Totals			288	122	114	153	115

Table 3: Error matrix for generic classification.

These results are very preliminary, and do not say anything conclusive about the nature of “generic” recognition, but they do suggest a route by which generic capability could arise in an appearance based system that was initially targeted at recognizing specific objects, but needed enough flexibility to be able to deal with inter-pose variability and environmental lighting effects. We want to look more at this in the future.

5 Comparisons to Other Methods

As far as we have been able to ascertain, the above results represent the most accurate reported in the literature for fully (orthographically) invariant recognition of general 3-D shapes tested on large sets of real images. There is some model-based work that seems accurate for shapes describable with planar regions or line segments; however none of these techniques are applicable to the sort of complex, curved shapes that form the majority of our examples. Furthermore, almost all of the papers illustrate the results on just a few examples, without the sort of full-sphere verification we present here.

Of the appearance-based techniques not using color, the best results on large, real image databases have been reported by groups directed by Nayar and Mohr (e.g. [8] [12]). Both groups present large scale tests on databases of real images. Nayar presents

results using eigenspace techniques for 3-D recognition in databases containing several tens of objects with accuracy comparable to what we report. Since system is trained only over a circle on the viewing sphere rather than the full sphere as we do, the results should be scaled accordingly. The eigenspace techniques also require accurate global segmentation to work, and would fail with several of the problem classes where we demonstrate success. On the other hand, the eigenspace techniques are much faster than ours, operating in a fraction of a second, whereas we take several seconds.

Mohr's methods are based on differential invariants, and exhibit good tolerance for clutter and occlusion. The group shows results for 3-D recognition and gets good results for a few tens of objects, again training over a circle rather than the full sphere (Nayar's database in fact). The drawbacks of this method are that (as of the most recent report) it does not handle geometric scaling gracefully, and since the features are differential invariants of the gray-scale image, it is somewhat sensitive to dramatic lighting and contrast changes. Our method is less sensitive in this respect, and handles geometric scaling implicitly. On the other hand, Mohr's techniques probably perform better in the presence of clutter and occlusion since they work with many more, and much smaller features than we do.

6 Conclusions and Future Work

In this paper we have described a framework for 3-D recognition based on loose assemblage of local context fragments keyed by distinctive features. The representation is similar in some striking ways to certain dramatic aspects of cubist art. We ran various large-scale performance tests and found good performance for full-sphere/hemisphere recognition of up to 24 complex, curved objects, robustness against clutter and occlusion, and some intriguing generic recognition behavior.

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. 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] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Trans. PAMI*, 15(10):1042–1062, 1993.
- [2] F. Stein and G. Medioni. Efficient 2-dimensional object recognition. In *Proc. ICPR*, pages 13–17, Atlantic City NJ, June 1990.
- [3] W. E. L. Grimson. *Object Recognition by Computer: The role of geometric constraints*. The MIT Press, Cambridge, 1990.
- [4] W. E. L. Grimson and D. P. Huttenlocher. On the sensitivity of geometric hashing. In *3rd International Conference on Computer Vision*, pages 334–338, 1990.
- [5] 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.
- [6] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.
- [7] B. Mel. Object classification with high-dimensional vectors. In *Proc. Telluride Workshop on Neuromorphic Engineering*, Telluride CO, July 1994.
- [8] H. Murase and S. K. Nayar. Learning and recognition of 3d objects from appearance. In *Proc. IEEE Workshop on Qualitative Vision*, pages 39–50, 1993.
- [9] R. C. Nelson. Finding line segments by stick growing. *IEEE Trans PAMI*, 16(5):519–523, May 1994.
- [10] R. P. Rao. Top-down gaze targeting for space-variant active vision. In *Proc. ARPA Image Understanding Workshop*, pages 1049–1058, Monterey CA, November 1994.
- [11] R. K. Ruud M. Bolle and D. Sabbah. Primitive shape extraction from range data. In *Proc. IEEE Workshop on Computer Vision*, pages 324–326, Miami FL, Nov-Dec 1989.
- [12] 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.
- [13] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Trans. PAMI*, 13(10), 1991.