MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING

A. I. Memo No. 1518
C. B. C. L. Paper No. 106

November, 1994

View-Based Strategies For 3D Object Recognition

Pawan Sinha and Tomaso Poggio

Abstract

A persistent issue of debate in the area of 3D object recognition concerns the nature of the experientially acquired object models in the primate visual system. One prominent proposal in this regard has expounded the use of object centered models, such as representations of the objects' 3D structures in a coordinate frame independent of the viewing parameters [Marr and Nishihara, 1978]. In contrast to this is another proposal which suggests that the viewing parameters encountered during the learning phase might be inextricably linked to subsequent performance on a recognition task [Tarr and Pinker, 1989; Poggio and Edelman, 1990]. The 'object model', according to this idea, is simply a collection of the sample views encountered during training. Given that object centered recognition strategies have the attractive feature of leading to viewpoint independence, they have garnered much of the research effort in the field of computational vision. Furthermore, since human recognition performance seems remarkably robust in the face of imaging variations [Ellis et al., 1989], it has often been implicitly assumed that the visual system employs an object centered strategy. In the present study we examine this assumption more closely. Our experimental results with a class of novel 3D structures strongly suggest the use of a view-based strategy by the human visual system even when it has the opportunity of constructing and using object-centered models. In fact, for our chosen class of objects, the results seem to support a stronger claim: 3D object recognition is *2D* view-based.

This report describes research done within the Artificial Intelligence Laboratory and the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences. Support for the laboratory's artificial intelligence research is provided by ARPA contract N00014-91-J-4038. Support for the Center is provided in part by a grant from the National Science Foundation under contract ASC-9217041 (funds provided by this award include funds from DARPA provided under the HPCC program) and by a grant from the National Institutes of Health under contract NIH 2-S07-RR07047.

1. Introduction

Viewer-centered recognition strategies have often been dismissed as being inelegant or, worse, infeasible on account of the large memory resources their naive implementations require. Some recent work on interpolation networks [Poggio and Girosi, 1990; Poggio and Edelman, 1990] has, however, mitigated this problem significantly. Having a few exemplars and the ability to interpolate between them essentially does away with the need for storing all the infinitely many views of any given object. While viewer-centered recognition has thus been rendered feasible from an engineering point of view, there remains open the important issue of whether this strategy has any biological significance. Does the primate visual system, for instance, use such a scheme for recognizing three-dimensional objects? This is the question we attempt to address psychophysically in the present paper.

2. Methods

The 3D objects (target as well as distractor) we chose for our experiments resembled thin bent paper clips with no cues to three-dimensionality other than the binocular disparities in their stereo images. They had 10 segments and were closed-loop. The target and distractor objects had a special relationship: the distractor objects were designed so as to have the same 2D projection as the target when viewed from one specific direction (which was designated to be the training direction for our experiments) but otherwise had unconstrained 3D structures (see figure 1). Corresponding to each target, either a single or several distractor objects were constructed. This manipulation did not affect the experimental outcome and the results reported here are from the single distractor condition.

Our six subjects were naive as to the purpose of the experiment. All of them had normal or corrected vision. A preliminary test with ten different random-dot stereograms was run to ensure that they were not stereo-blind. All displays were presented stereoscopically on a Silicon Graphics workstation. Subjects were required to wear stereo glasses to view the stimuli in depth. A chin-rest placed at a distance of 70 cm from the display screen served to minimize head movements. No feedback was provided until the conclusion of all experimental sessions.

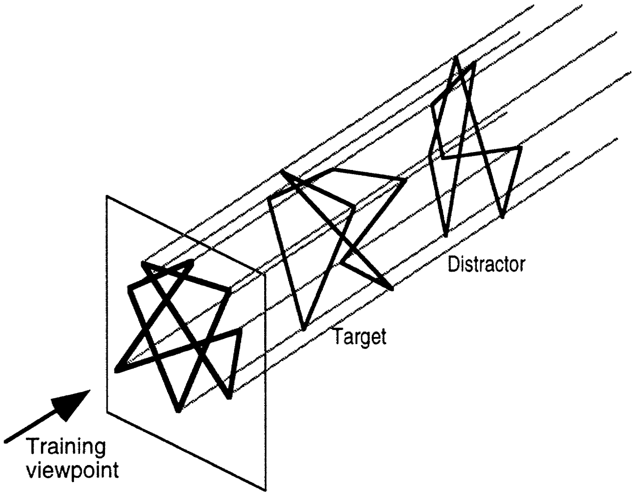

Figure 1. The nature of target and distractor objects used in our study.

Our experimental sessions were divided into two phases: a training phase lasting for 25 seconds during which the subject was stereoscopically shown a static 3D object, and a test phase that examined the subject's recognition performance against a set of distractor objects in a 2-AFC setup. Each trial of the test session stereoscopically presented a pair of objects for 1.8 seconds. One of these was the target and the other a distractor. Subjects were asked to identify the former.

As shown in figure 2, the test pairs were generated by systematically varying the viewing directions for the target and distractor objects. We began with viewing the distractor from the training direction and the target from 90 degrees away (with reference to a vertical axis). The viewing directions were then altered (in opposite directions for the target and distractor) in steps of 10 degrees to ultimately have the target viewed from the training direction and the distractor from the 'side'. This process produced object pairs where the 2D appearances of the distractor and target objects exhibited varying degrees of similarity with the 2D appearance of the target during training. The 3D structures of the target and distractor, of course, remained unchanged throughout the test session. The systematic variation of viewing directions was not evident to the subjects since the pairs were presented in a random sequence. Each pair was presented several times during a test session.

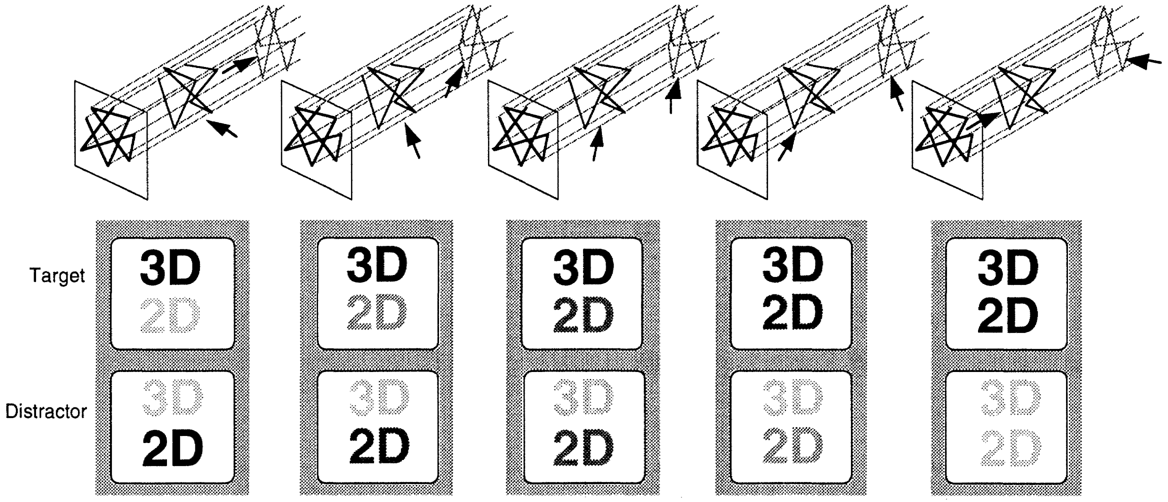

Figure 2. Object pairs presented during the test session. As indicated in the top row, the viewing directions for the distractor and target objects were varied systematically. The distractor viewing direction steadily got less aligned with the training direction while the target viewing direction got increasingly more aligned with the training direction. The degree of similarity (in 2D and 3D) of the resulting pairs of views with the target object (and its projection during training) is shown schematically in the bottom row (darker grays represent higher similarity to the training object/view).

3. Results and Discussion

The object-centered and the view-based recognition schemes make very different predictions about a subject's recognition performance (the percentage of correct responses) for the different pairs generated by a given target-distractor combination. These are illustrated in figure 3. An object centered scheme (say, one that uses object centered 3D models) would predict that it should always be possible to pick out the target from the distractor irrespective of the viewing direction of either by matching their 3D structures against the target 3D model acquired during training (remember that all objects in the experiment are presented stereoscopically with plainly evident 3D structures). Therefore, the psychometric function relating the percentage of correct responses to the systematically varied angular deviation in viewing direction would be expected to be flat. A view-based scheme, however, would predict that subjects would pick the alternative that presented a 2D appearance more like the 2D appearance of the training object. This would lead to selecting the distractor in pairs where the distractor viewing direction is similar to the training direction and the actual target in others. The psychometric function relating the percentage of correct responses to the systematically varied angular deviation in viewing direction would, therefore, be expected to have a sigmoidal form.

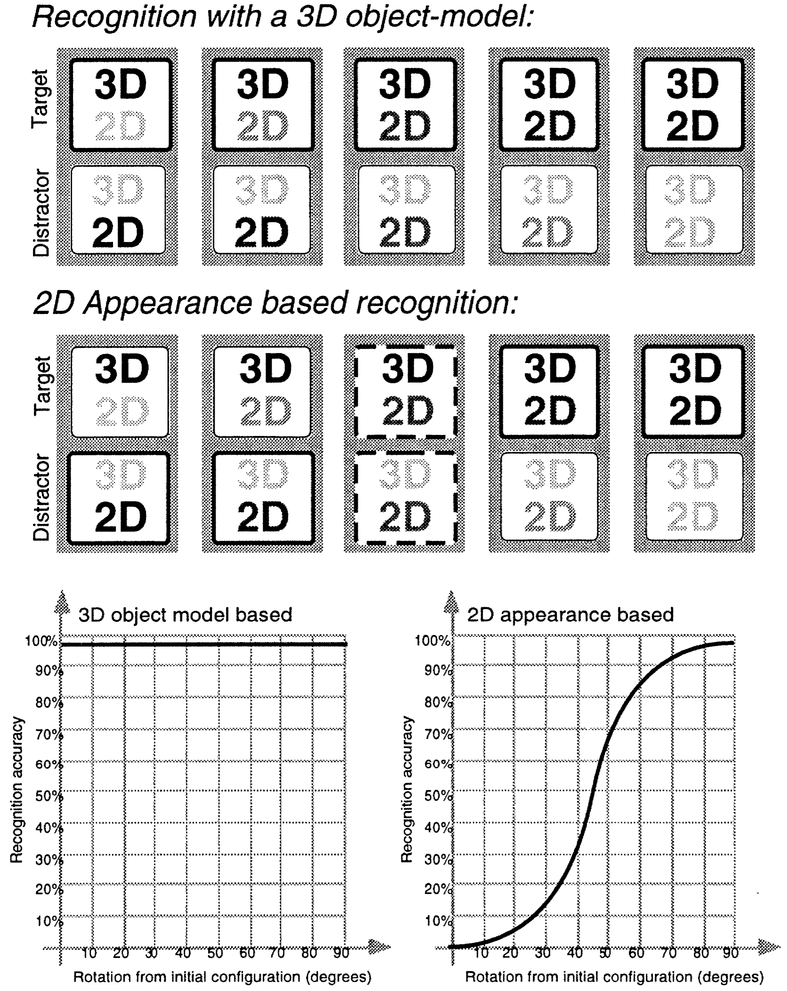

Figure 3. The differing predictions about recognition performance (and the psychometric functions) made by the object-centered and view-based schemes. The highlighted boxes in the top two panels indicate which of the two objects in a pair a subject is likely to identify as the target.

Figure 4 shows the experimental results from six subjects averaged over two sessions each. The sigmoidal tendency in the psychometric functions from all subjects except for BMB is quite evident. This data supports the 2D view-based scheme for recognition over one that calls for use of object-centered 3D models. BMB's results present an interesting exception to the general trend. Her near perfect performance might be justifiably construed to lend support to the object-centered scheme. But, such a conclusion would have to be qualified by the fact that BMB's visual experience is somewhat atypical. She is a computational molecular biologist by training and has extensive experience viewing stereo-images of complex protein molecules. It is very likely that this experience has lent her a measure of facility in memorizing and manipulating 3D structures besides changing her criterion for inter-object similarity. BMB's results are also significant from another point of view. They demonstrate that there is enough information in the displays to make correct judgements all the time. Subjects' use of view-based strategies, therefore, is not a matter of coercion, but a matter of choice.

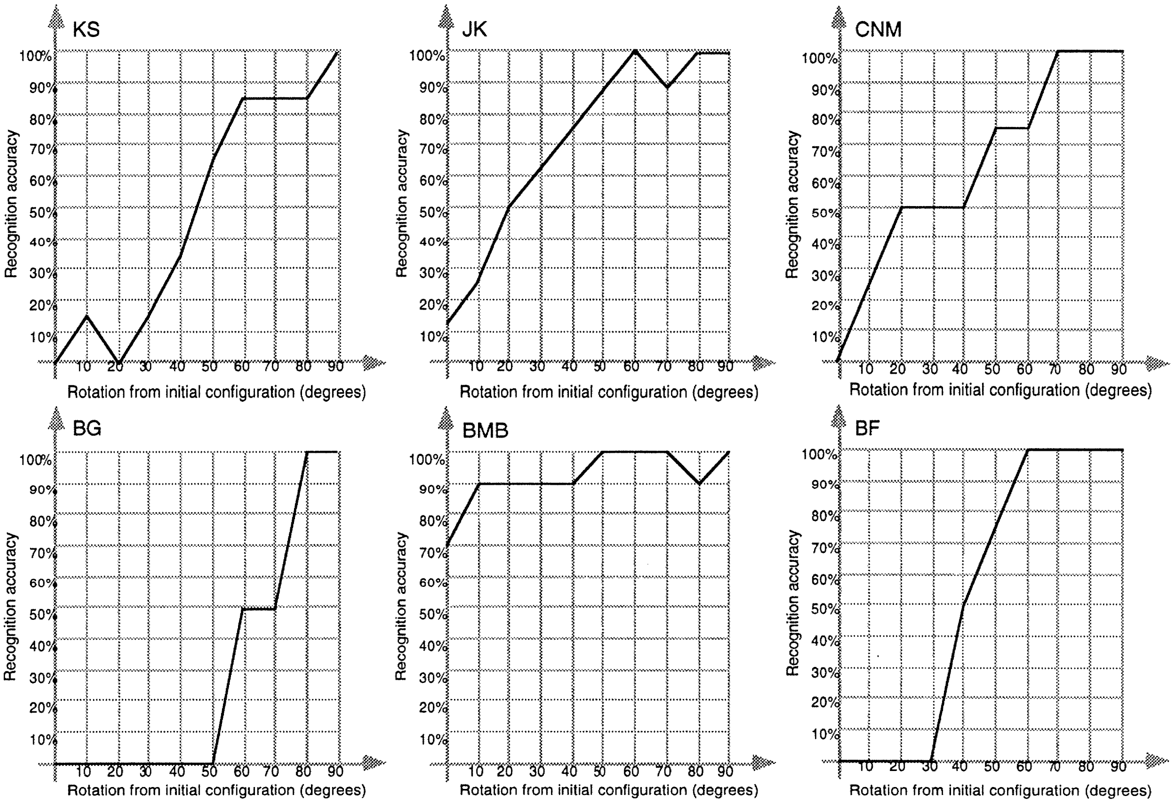

Figure 4. The experimental results showing recognition performance for six subjects averaged over two sessions each. The pronounced sigmoidal tendency of the psychometric functions indicates use of a view-based recognition strategy by the subjects. BMB's results deviate from the average probably because of her previous experience with stereo-imagery. See text for details.

In the context of this data, it is appropriate to consider the question of why the brain might opt for a view-based strategy over the one that involves transforming and projecting 3D object-centered models. We discuss three candidate answers.

First, the bias towards view-based strategies might simply be an evolutionary vestige. Binocular vision is a relative new-comer insofar as the evolutionary history of the visual system is concerned. The view-based strategies that the brain might have been forced to use before the development of binocularity might simply have carried over to this day. A loose parallel may be drawn from the field of color-vision. Possibly because of its rather recent arrival, color vision does not play a big part in several perceptual processes, most notably in those having to do with motion. The brain probably just has not had enough time to develop strategies that incorporate such additional sources of information. It is important to emphasize that this is not a case where information is not available, but rather one where it is not used during the performance of certain tasks. In other words, not all the attributes perceived are necessarily used for recognition.

Second, purely from an information theoretic point of view, 2D information is very often enough enough to uniquely index into a library of stored models in a 'non-malicious' visual world like ours. The conditional probability of correctly identifying a 3D object given its 2D image is, therefore, very high. The recognition strategy used by our visual systems might be designed to implicitly exploit this fact.

Third, a view-based strategy makes sense in terms of how the brain is 'implemented'. Its computational powers are somewhat limited but its memory capacity is truly phenomenal. Accordingly, a memory-intensive view-based strategy would seem more appropriate for the brain than a computation intensive transformationist strategy for object recognition.

In summary, we have presented experimental results that suggest that for certain classes of 3D objects, recognition might be mediated by 2D view-based strategies. We feel that our results are significant because unlike some previous studies [Bulthoff and Edelman, 1992], we made available information about the 3D structures of objects during both the training and test sessions. The subjects, therefore, had the opportunity to memorize and use an object-centered 3D model for their discrimination task, but opted for the view-based strategy instead. It seems that the hypothesis of 3D object recognition being viewer-centered can be refined further to claim that for certain classes of objects, recognition might be 2D view-based.

References:

Buelthoff, H. H. and Edelman, S. (1992) *Proc. Natl. Acad. Sci.*, USA, **89**: 60-64.

Ellis, R., Allport, D. A., Humphreys, G. W., and Collis, J. (1989) *Q. Journal of Exp. Psychology*, **41A**:775-796.

Marr, D. and Nishihara, H. K. (1978) *Proc. of the Royal Society of London: B*, **200**:269-294.

Poggio, T. and Girosi, F. (1990) *Science*, **247**: 978-982.

Poggio, T. and Edelman, S. (1990) *Nature*, **343**: 263-266.

Tarr, M. J. and Pinker, S. (1989) *Cognitive Psychology*, **21**:233-282.