IETE Journal of Research
Vol 49, No 2, March-April, 2003, pp

Use of 2D Similarity Metrics for 3D Object Recognition

PAWAN SINHA
E25-229, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
45 Carleton Street, Cambridge, MA 02142, USA.

How the brain recognizes three-dimensional objects is one of the fundamental open questions in visual neuroscience. The challenge is to determine the information the visual system uses for making recognition decisions. It is generally accepted that shape attributes play an important role in the recognition of many object classes. However, the term 'shape attributes' is loosely defined and encompasses not only the projected two-dimensional shapes of the objects, but also their three-dimensional forms. It is unclear whether, for the purpose of recognition, the visual system favors the use of projected shape over depth structure or vice-versa. We have developed an experimental paradigm that allows us to directly compare the relative efficacies of the two kinds of information in a simplified object domain, and thereby provides a rigorous method for addressing this important question. Our results indicate that while it is possible for humans to use depth information when explicitly instructed to do so, their default recognition strategy is overwhelmingly biased towards the use of two-dimensional projected shape. We discuss possible reasons for this bias and also consider the implications of these results with regard to the issue of the nature of internal representations for three-dimensional objects.

Indexing terms: 3D object recognition, representations, similarity metrics

While material properties such as color and texture are undoubtedly important for recognition in certain situations [31], object shape seems to have a more general significance [6, 8, 9]. Recognition of several objects is unaffected by changes in their color and texture but is profoundly disrupted by shape changes. The fact that we can recognize objects in line-drawings, monochrome images and Gauguin's paintings, that often have non-veridical object colors, attests to the importance of shape.

In the context of three-dimensional object recognition, the term 'shape' is somewhat vague. It can refer to an object's 3D structure or to its 2D projected form. In considering the visual system's reliance on shape attributes, it is important to determine whether it is one or the other (or both) of these two kinds of information that is actually used. This is the issue we address in this paper. An answer to this question would shed light on the more general problem of how the visual system is able to achieve its seemingly remarkable recognition performance. It would also provide clues to the nature of the internal representations of three-dimensional objects.

PRIOR WORK

The issue of the nature of shape attributes used by the visual system for 3D object recognition has a rich history of research and debate.

Suggestions favoring the use of objects' 3D structures have been motivated, in large part, by the seemingly robust recognition performance of the human visual system under significant view-point changes. In an influential paper, Marr and Nishihara [14] proposed that the visual system reconstructs the 3D structure of an object using cues such as binocular disparity as a precursor to matching it against stored representations. Work by Biederman and his colleagues [1, 2] has also emphasized the role of 3D shape attributes in recognition. Phinney and Siegel have recently presented results that suggest that the visual system can recognize 3D objects even in the almost complete absence of 2D shape cues [15]. Such results conform well with our introspective sense of the vivid three-dimensionality of our visual world - it seems reasonable to expect that the perceived three-dimensionality of objects would also influence their recognition.

Indirect evidence for the use of 2D shape cues has come from studies showing significant limitations in an observer's ability to recognize novel 3D objects from view-points different from the previously experienced ones [17, 19, 28, 29, 3, 13, 23, 24]. These authors have proposed that it is the 2D appearance of an object, rather than its 3D structure, that is matched against the internal representations.

In our work we seek to rigorously address the question of the nature of shape attributes used by the visual system for 3D object recognition. To do so, we have developed an experimental paradigm that allows a direct comparison of the relative significance of an object's 2D projected shape and its 3D structure for recognition.

Paper No 193-D; Copyright © 2003 by the IETE.

The proposals of using objects' 3D structures on the one hand and 2D projected shapes on the other define two ends of a continuum of possibilities. In the present study, we describe three experiments that attempt to determine for a chosen class of objects, which end of this spectrum the strategy used by the human visual system is closer to. We are mindful, however, of the fact that any single answer to this complex question is likely to be an oversimplification. The complete answer will likely involve a collection of different schemes, their applicability determined by the demands of the task at hand. Some of this task-dependent complexity will become evident in our set of results.

EXPERIMENT 1

The goal of experiment 1 was to determine, in a two alternative forced choice paradigm, whether after having been trained on a novel three-dimensional object, the subjects' recognition performance was more consistent with the predictions of a viewpoint independent representation scheme or a viewpoint-dependent one. An important feature of the experiment was that explicit three-dimensional information about the training and the test objects was always available to the subjects.

Stimuli

For use as experimental stimuli, we needed a class of three-dimensional objects that satisfied the following three criteria:

1. Objects should have only shape cues,
2. Viewing the object from different positions should not lead to variable amounts of self-occlusion. In other words, the amount of information in the objects' images should be constant across all viewing positions, and
3. The objects should be novel to prevent any previously acquired familiarity biases in the subjects from influencing the results.

The class of three-dimensional thin-wire sculptures made of sequentially joined straight segments meets all these criteria and is the one we draw our stimuli from. Members of this object-class have no cues to three-dimensionality other than the binocular disparities in their stereo images. This class has a long history of use in perceptual and cognitive studies [30, 16, 17, 18, 3, 7]. The specific objects we used in our experiments had 10 segments and were closed-loop. Care was taken to exclude objects with accidental perceptually distinctive characteristics like aligned vertices, parallel segments or symmetry. Such objects were, however, used in experiment 3.

Subjects

We used eight subjects with differing backgrounds. They were drawn from the staff and students at MIT and the Max-Planck Institute for Biological Cybernetics in Tübingen, Germany. Only two of these knew of the purpose and design of the experiment. The subjects' ages ranged from 20 years to 52 years. All of them had normal or corrected to normal vision. A preliminary test with ten different random-dot stereograms was run to ensure that they were not stereo-blind.

Methods

To allow subjects to view the stimuli in depth, all displays were presented stereoscopically either via free-fusion (for subjects who were able to do so) or via stereo glasses. Viewing distance was 70 cm from the display screen. No feedback was provided until the conclusion of all experimental sessions. Experiments were run for each subject individually. The experimental sessions were divided into two phases: a training phase and a test phase. During the training phase, which lasted uninterrupted for 25 seconds, the subject was stereoscopically shown a static 3D object, which he/she was asked to examine for a subsequent recognition task. The test phase commenced 5 seconds after the conclusion of the training. This phase was designed to examine the subject's recognition performance against a set of distracter objects in a 2-AFC (two alternative forced choice) setup. Each trial of the test session stereoscopically presented a pair of objects for 1800 milliseconds. One of these was the target and the other a distracter. Subjects were asked to identify the former. The target and distracter objects had a special relationship: the distracter 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).

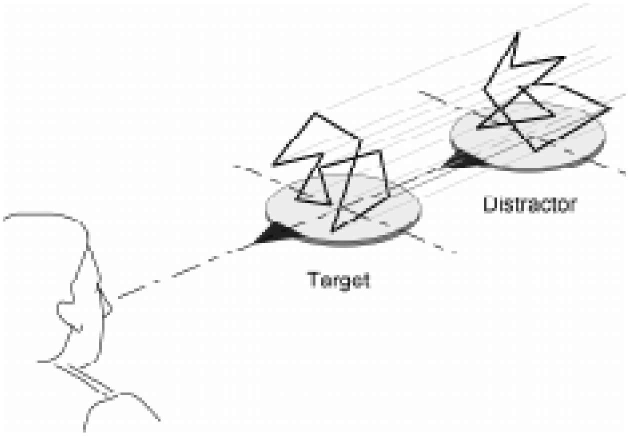

Fig 1 During the training phase, which lasted uninterrupted for 25 seconds, the subject was stereoscopically shown a static 3D object which he/she was asked to examine for a subsequent recognition task. Meanwhile, a distractor object was constructed without the knowledge of the subject. The distractor object was designed to have the same 2D projection as the training object when viewed from the training direction

Corresponding to each target, either a single or several distracter objects were constructed. This manipulation did not significantly affect the experimental outcome and the results reported here are from the single distracter condition.

Figure 1. During the training phase, which lasted uninterrupted for 25 seconds, the subject was stereoscopically shown a static 3D object, which he/she was asked to examine for a subsequent recognition task. A distracter object was constructed without the knowledge of the subject. The distracter object was designed to have the same 2D projection as the training object when viewed from the training direction.

The test pairs were generated by systematically varying the viewing directions for the target and distracter objects. As shown in figure 2, we began with viewing the distracter 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 distracter) in steps of 10 degrees to ultimately have the target viewed from the training direction and the distracter from the 'side' (figure 3). This process produced object pairs where the 2D appearances of the distracter and target objects exhibited varying degrees of similarity with the 2D appearance of the target during training (figure 4). The 3D structures of the target and distracter, 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 three times during a test session. Each subject was tested on three objects (these objects remained the same across all subjects). Figure 5 summarizes the complete experimental procedure.

Figure 2. The test pairs were created by first having the

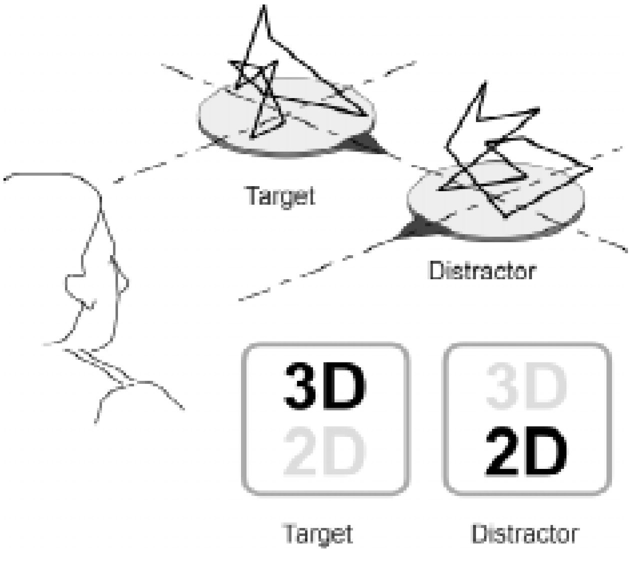

Fig 2 The test pairs were created by first having the distractor be viewed from the training direction and the target from 90 degrees away (with reference to a vertical axis). In this condition, the distractor looks very similar to the training object in 2D. The target, on the other hand, has poor 2D similarity since it is being viewed from the side

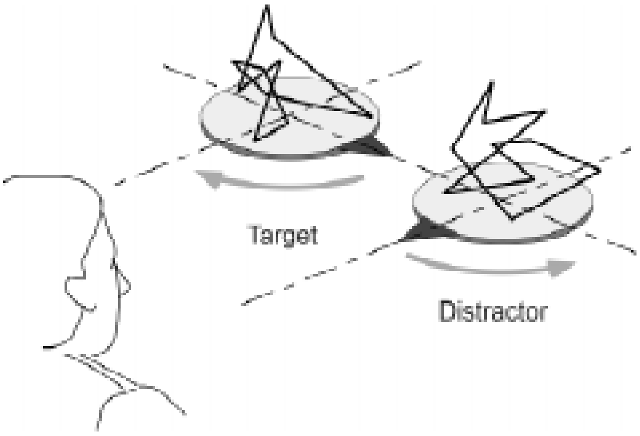

Fig 3 To generate the rest of the test pairs, the viewing directions were altered (in opposite directions for the target and distractor) in steps of 10 degrees to ultimately have the target viewed from the training direction and the distractor from the 'side'

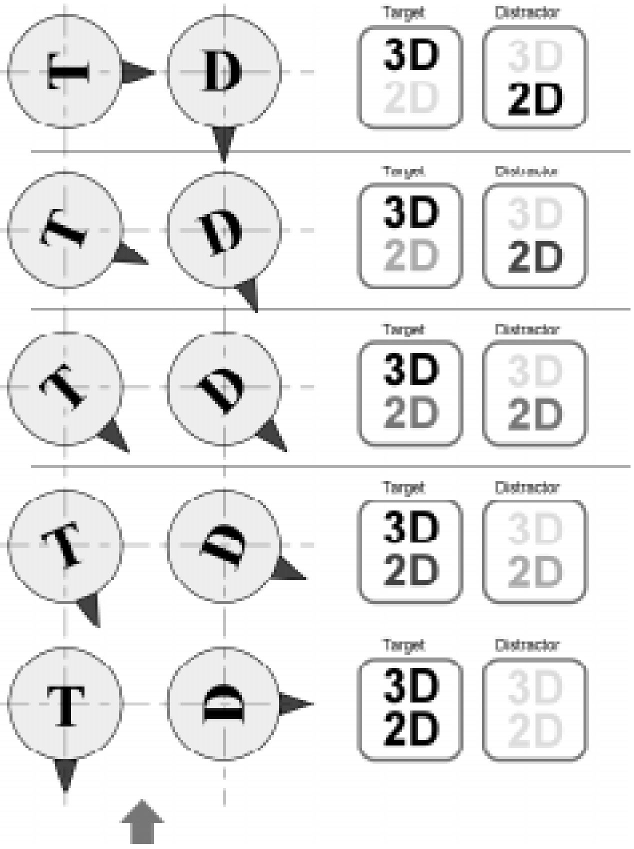

Fig 4 The test pair generation process illustrated in fig 2 and 3 produced object pairs where the 2D appearances of the distractor and target objects exhibited varying degrees of similarity with the 2D appearance of the target during training. The similarity is indicated by shades of gray - the darker the shade, the higher the similarity

distracter viewed from the training direction and the target from 90 degrees away (with reference to a vertical axis). In this condition, the distracter looks very similar to the training object in 2D. The target, on the other hand, has poor 2D similarity since it is being viewed from the side.

Figure 3. To generate the rest of the test pairs, the viewing directions were altered (in opposite directions for the target and distracter) in steps of 10 degrees to ultimately have the target viewed from the training direction and the distracter from the 'side'.

Figure 4. The test pair generation process illustrated in figures 2 and 3 produced object pairs where the 2D appearances of the distracter and target objects exhibited varying degrees of similarity with the 2D appearance of the target during training. The similarity is indicated by shades of gray - the darker the shade, the higher the similarity.

Figure 5. A summary of the complete experimental procedure.

Predictions of the two hypotheses

The viewpoint-independent and the viewpoint-dependent representation 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-distracter combination. These are illustrated in figure 6. A viewpoint-independent scheme would predict that it should always be possible to pick out the target from the distracter irrespective of the viewing direction of either. This should be especially easy in our experiment because all objects are presented stereoscopically with plainly evident 3D structures, and the input 3D structures can be matched against the target 3D model acquired during training. Therefore, the psychometric function (the dependence of perception on an aspect of the stimulus) relating the percentage of correct responses to the systematically varied angular deviation in viewing direction would be expected to be flat. A viewpoint-dependent 2D 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 distracter in pairs where the distracter 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 6. The differing predictions made by viewpoint-independent and the viewpoint-dependent representation schemes about a subject's recognition performance in our experiment. (a) A viewpoint independent scheme would allow a subject to consistently pick out the target object in all target/distracter pairs. (b) Use of a viewpoint dependent scheme would result in choosing the distracter or target depending on which one was viewed from a direction closer to the training direction. (c) and (d) Performance curves predicted by the two strategies.

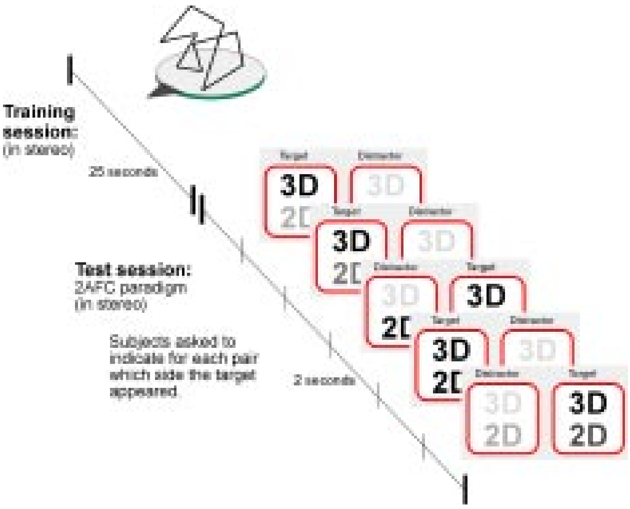

Fig 5 A summary of the complete experimental procedure

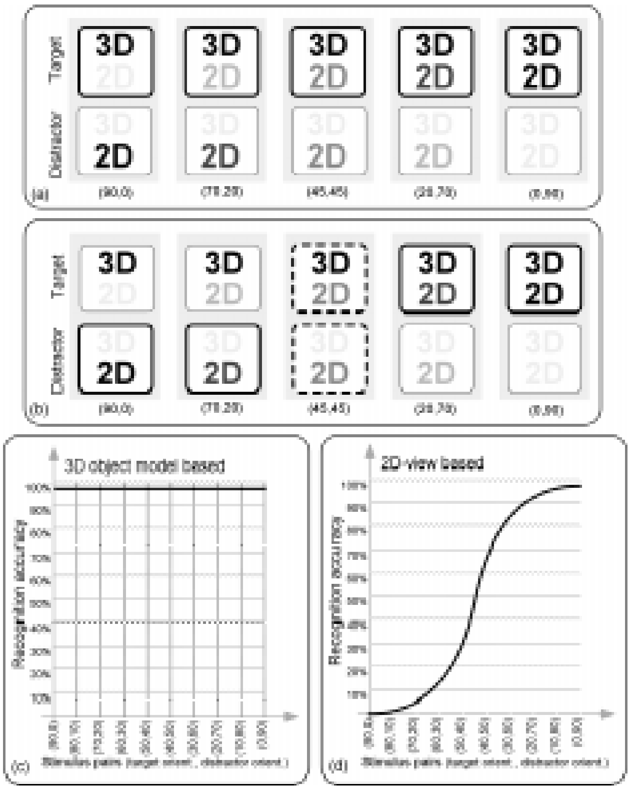

Fig 6 The differing predictions made by viewpoint-independent and the viewpoint-dependent representation schemes about a subject recognition performance in our experiment

Experimental Results

Figure 7 shows data obtained with the two subjects who knew of the experimental purpose and design. Their performance, for nearly all target-distracter pairs is high, consistent with the predictions of the viewpoint-independent representation scheme.

Figure 7. Results obtained with the two subjects who knew of the experimental purpose and design.

Results obtained with the six naive subjects are shown in figure 8. They clearly have a very different form from those in figure 7. All of them except one (S4) exhibit a pronounced sigmoidal tendency, consistent with the predictions of the 2D view-based representation scheme.

Figure 8. Results obtained with six naive subjects. Axis labels are the same as in figure 7.

Discussion

Our results with subjects informed of the experimental design demonstrate that they can indeed encode the 3D structure of the training object and subsequently arbitrarily transform it to match against the test objects. The results with naive subjects, on the other hand, strongly suggest the use of view-based representations. How might we account for the large differences between the performances of these two populations of subjects?

One parsimonious explanation of these differences is that the default tendency in the visual system is to use 2D view-based representations. However, knowledge about the specific demands of a task can result in this scheme being augmented with, or even completely replaced by, the use of viewpoint independent representations that encode, either explicitly or implicitly, the 3D structures of objects. In other words, while the visual system is not incapable of encoding and manipulating 3D object information (results from mental rotation experiments [5, 20, 21] too suggest this), by default it tends not to do so. We shall briefly discuss possible reasons for this default tendency in the general discussion section.

Subject S4's results (see figure 8) are in marked contrast to those of the other naive subjects. This subject exhibits a far greater degree of viewpoint independence than the rest. We believe that the reason for this difference might lie in S4's background. This individual is a computational molecular biologist by profession and has had extensive experience looking at and reasoning about stereo images of (schematically depicted) molecules. It is possible that this experience has modified the default notion of similarity to be biased more towards 3D structural congruence and has also heightened the facility to mentally encode and manipulate 3D structures. S4's results are, therefore, interesting in that not only do they serve as a control by demonstrating that there is enough information in the display to perform the task in a viewpoint independent fashion, they also suggest an influence of past visual experience on subsequent recognition performance.

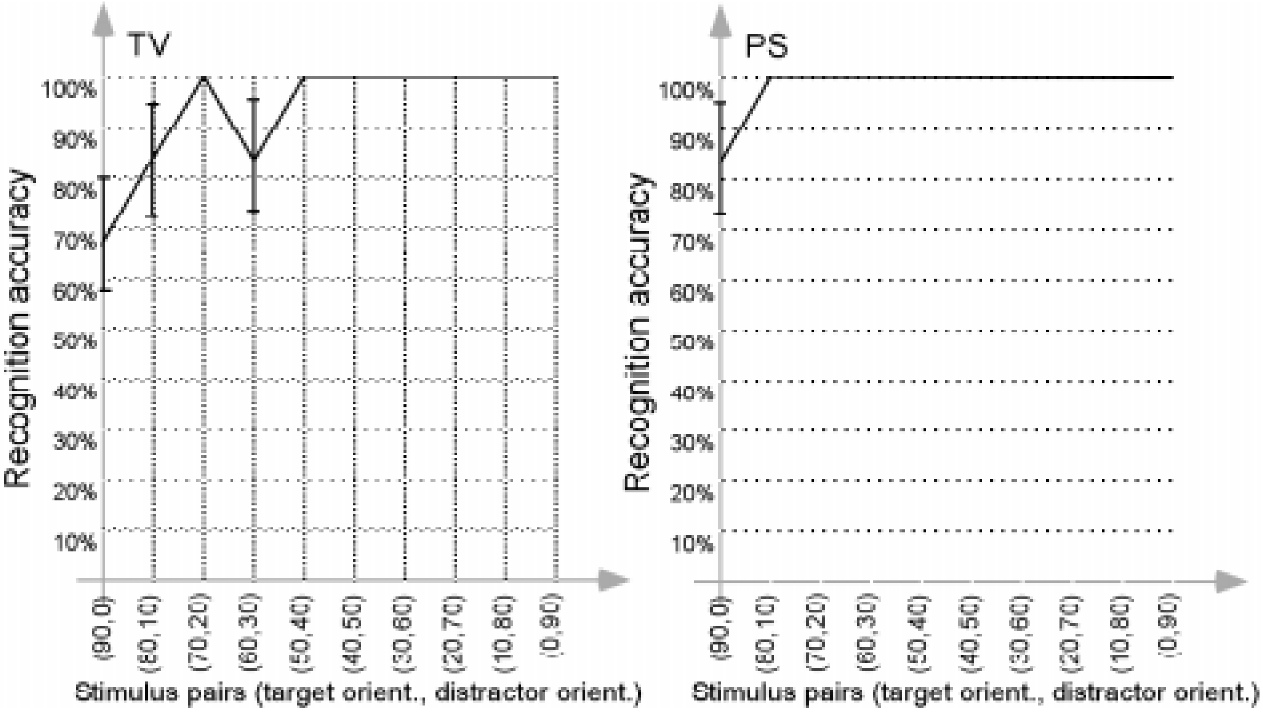

Fig 7 Results obtained with the two subject who knew of the experimental purpose and design.

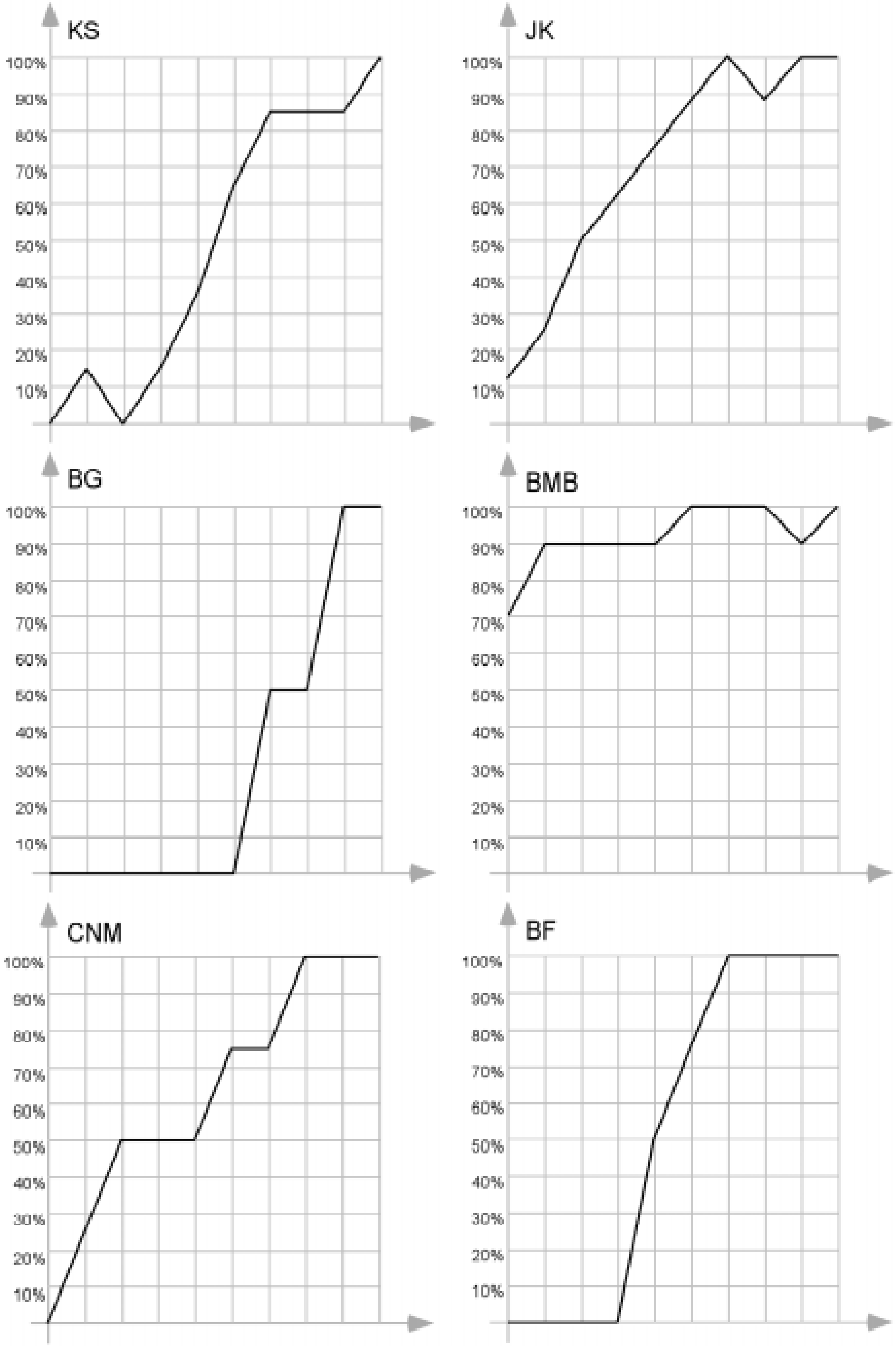

Fig 8 Results obtained with six naive subjects. Axis labels are the same as in Fig 7

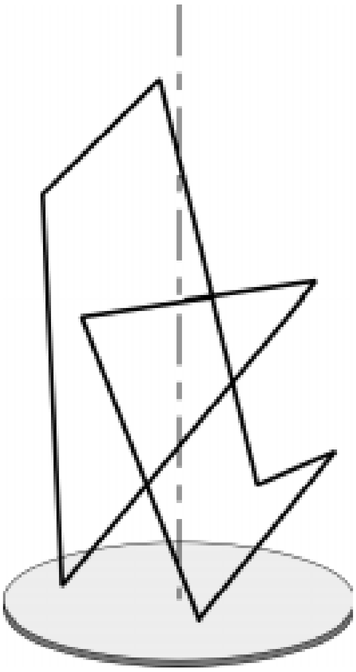

Fig 9 A modification of our class of stimulus objects to indicate a specific medical axis that could provide a frame of reference for encoding the 3D shape of the object

One possible criticism of our experimental design is that the objects we used as stimuli were not 'fair' for some classes of viewpoint-independent representation schemes. Such schemes attempt to construct a 3D structural description wherein the pose of the constituent parts is indicated with respect to a reference frame centered on the object [14]. Given that our experimental stimuli had no discernible axis that could serve to anchor the reference frame, such schemes might have been unfairly handicapped. To address this criticism, we repeated our experiments with elongated objects having a clearly defined medial axis (that also served as the axis of rotation in the experiment) (see figure 9). The results obtained with such objects were virtually the same as those with the original object set.

Figure 9. A modification of our class of stimulus objects to indicate a specific medial axis that could provide a frame of reference for encoding the 3D shape of the object.

The naive subjects' tendency to use 2D similarity as a criterion for recognition is further evidenced by the results of a slightly modified version of experiment 1. Subjects were asked not only to pick the target object in each pair, but also to indicate the confidence of each of their responses on a scale of 1 (unsure) to 3 (very sure). We found that the confidence is high when either the distracter or the target have high 2D similarity to the training object and low otherwise. Thus, even when subjects are making the objectively wrong response, they are confident of being correct so long as the chosen object has high 2D similarity to the training object. The inconsistent 3D information apparently is ignored. Does this result prove that the internal representations themselves are largely two-dimensional? No. It is possible that some depth information is indeed encoded in the representations but is not discernible in our experiments because it is overwhelmed by the high 2D similarity of either the target or the distracter to the training object. Experiment 2 was designed to address this possibility.

Experiment 2

Subjects

Experiment 2 was run on four of the eight subjects who had participated in experiment 1. Only one of the four was informed of the experiment's purpose.

Methods

The experimental procedure was identical to that in experiment 1 except that the target-distracter pairs were generated not by having the two objects move in anti-phase, but rather in-phase in steps of 10 degrees (see figure 10). The starting viewpoint for both objects was aligned with the training direction. As shown in figure 11, this paradigm results in target-distracter pairs where the only distinguishing attribute between the two is their 3D structure. Each subject was tested with three objects and no feedback was provided during the experimental sessions.

Figure 10. The test pairs in experiment 2 were generated by having the training and distracter objects move in-phase in steps of 10 degrees starting with a viewpoint aligned to the training direction.

Figure 11. The target-distracter pairs generated by the procedure illustrated in figure 11.

Experimental Results

Figure 12 shows the data obtained from the four subjects. As in experiment 1, the subject informed of the experimental design performed at ceiling for all trials. The data for the other three subjects individually averaged over all three sessions shows some variations. For two of the subjects, performance is a little above chance for presentations where the viewing angle is close to the training direction and at chance for the other presentations. For the remaining subject, however, performance is significantly above chance for the initial pairs and registers a slight drop for the other pairs.

Figure 12. Results from four subjects in experiment 2. The subject designators here do not correspond to those in figure 8.

Discussion

It seems valid to conclude from the results that at least some depth information is indeed retained in the internal

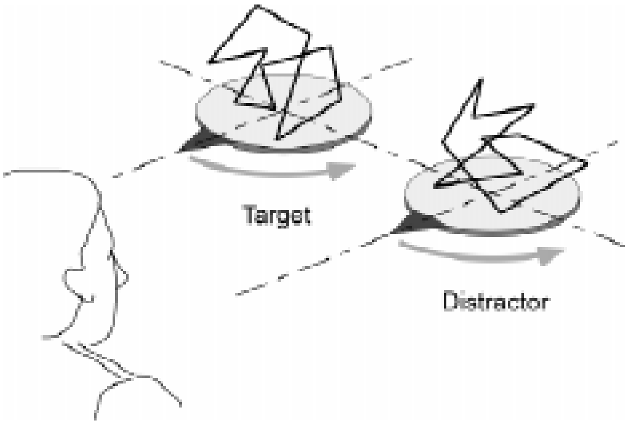

Fig 10 The test pairs in experiment 2 were generated by having the training and distractor objects move in-phase in steps of 10 degrees starting with a viewpoint aligned to the training direction

object representations allowing subjects to exhibit above chance performance while discriminating between the target and distracter objects for viewpoints similar to the training viewpoint. The fall-off of performance as the objects rotate away from the training direction suggests that this depth information cannot readily be manipulated by subjects to match against novel images.

One caveat regarding the design of this experiment needs to be mentioned. The test sessions can potentially allow subjects to guess what stimulus attribute is important to perform the task well. After going through the first session and observing that the two members of some pairs have identical 2D structures, subjects might correctly conclude that it is the 3D shape of the objects that is the discriminatory attribute. This would induce them to pay close attention to the 3D structure of the training object in subsequent experimental sessions. While this would constitute an interesting illustration of a strategy change based on the demands of the recognition task, for our purposes it would have the unfortunate consequence of biasing the results away from their default values. We do not yet have any good ways to address this problem.

Experiment 3

We had mentioned earlier that care had been taken to ensure that the objects used as experimental stimuli in experiments 1 and 2 did not possess any accidental perceptually salient characteristics. However, since such objects constitute an interesting and important subclass by themselves, we decided to study subjects' recognition performance exclusively with them in experiment 3. Our goal was to determine if view-based schemes were still good predictors of perceptual behavior in this situation.

Subjects

Our experiments were run on four subjects drawn from the set of those who had already participated in experiment 1. All of them were naive as to the purpose of the experiment.

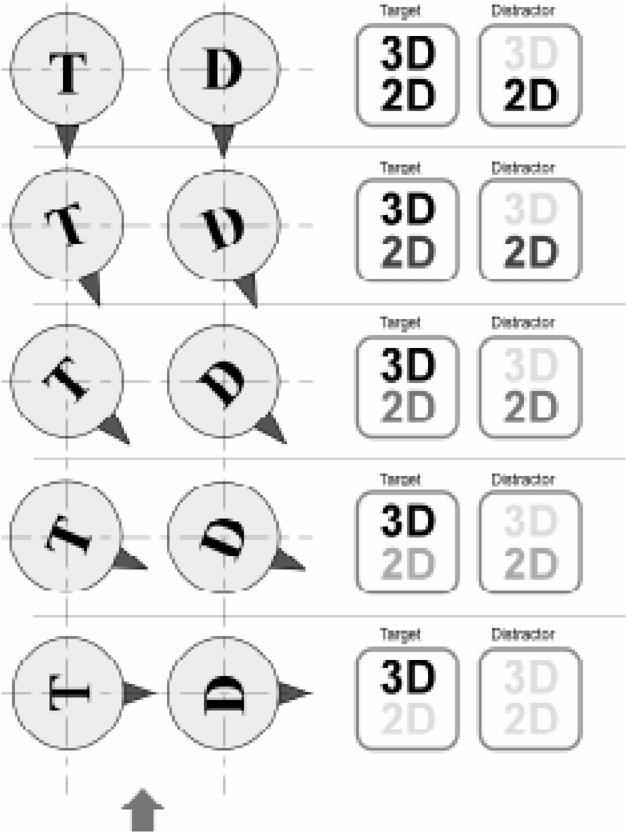

Fig 11 The target-distractor pairs generated by the procedure illustrated in Fig 11

Stimuli

We drew our target stimuli from three specially constructed classes of objects with perceptually salient characteristics (see figure 13). The first class included objects with a pair of vertices aligned in space. The second had a pair of parallel segments and the third had symmetric lower and upper halves. The distracters produced by applying random depth perturbations to the vertices of these objects did not preserve their perceptually distinctive characteristics.

Figure 14. The three classes of objects from which stimuli were drawn for experiment 3.

Methods

The experimental protocol was precisely the same as for experiment 1. Every subject was tested on two objects from each of the three classes. No feedback was provided during the experimental sessions.

Experimental Results

Figure 14 shows the results from our four subjects for each of the three classes of objects. The results suggest that

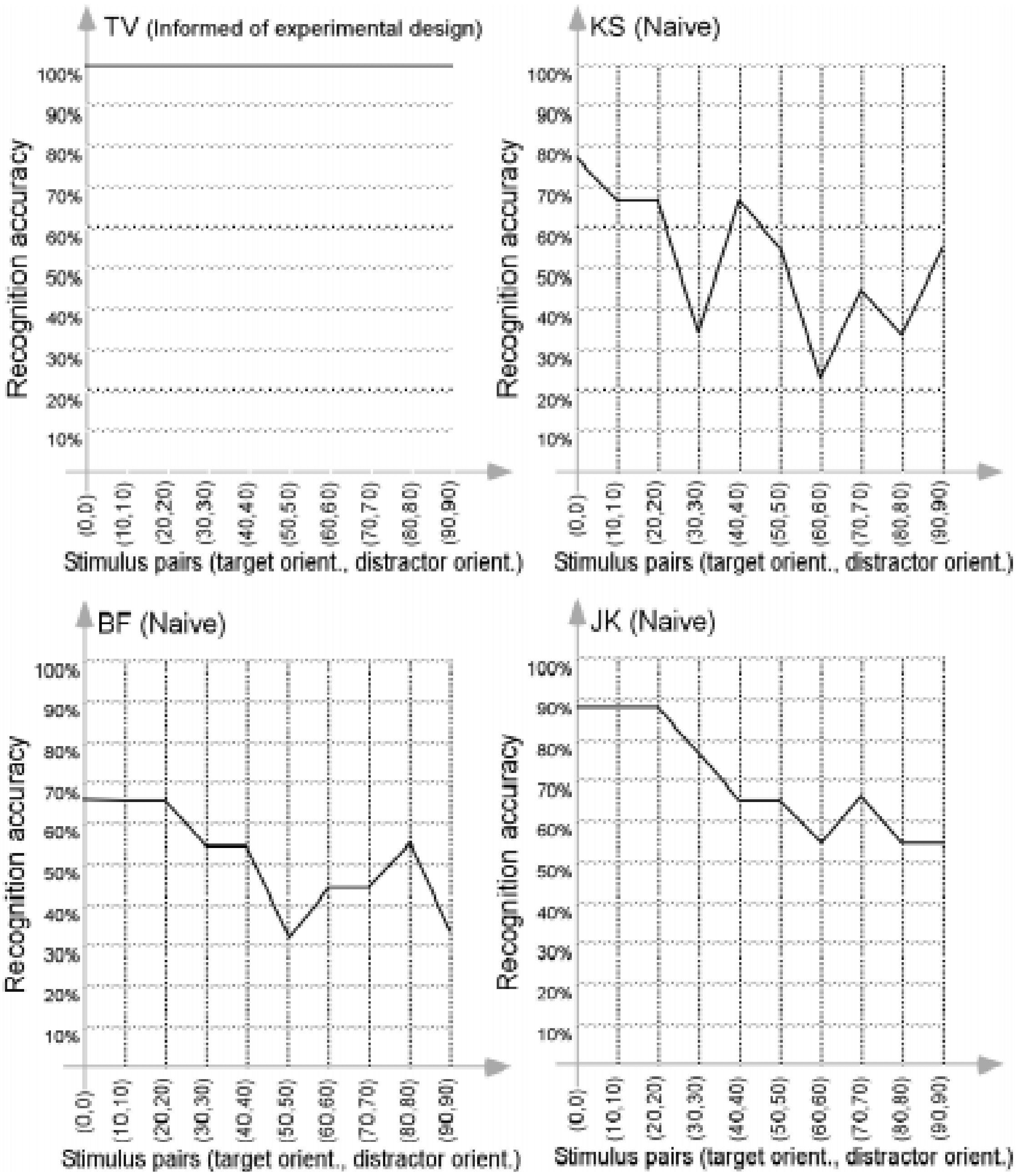

Fig 12 Results from four subjects in experiment 2

although the subjects sometimes tend to confuse the distracter for the target when the former is viewed from near the training direction, for all other viewing angles, their discrimination performance is very high. There are exceptions to this general trend, though, as evidenced in the figure.

Figure 14. Results of experiment 3 from four subjects for each of the three classes of objects.

Across object classes, performance is best for symmetric stimuli, followed by stimuli with a pair of parallel segments and then stimuli with a pair of aligned vertices.

Discussion

Clearly, the results of our four subjects are more consistent with the predictions of viewpoint-independent representation scheme than with those of the viewpoint-dependent one. However, a view-independent representation scheme is not alone in predicting such performance; a scheme based on the use of distinctive features would do so too. Based on the results we have obtained in experiments 1 and 2, we suggest that it is the feature based scheme that is a more accurate model of human recognition processes in this case.

As its name suggests, a feature-based scheme encodes

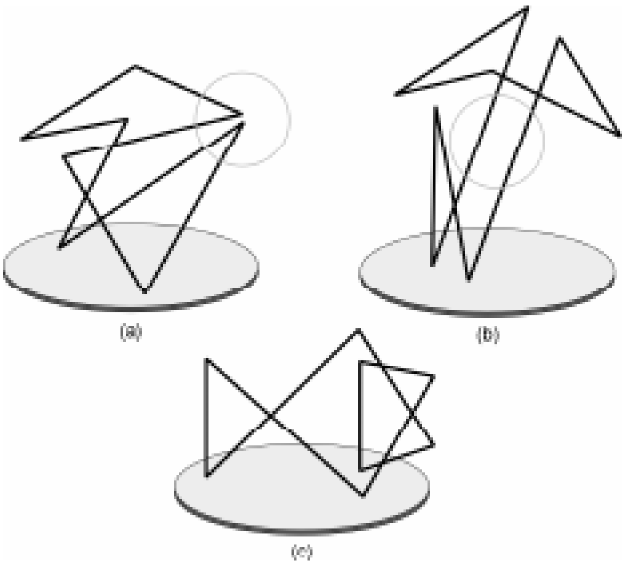

Fig 13 The three classes of objects from which stimuli were drawn for experiment 3.

objects as a set of distinctive characteristics. Recognition involves verifying whether or not the stimulus has those characteristics. In our experiment, the characteristics would be those of vertex-alignment, segment parallelness, and symmetry. The high level of subjects' performance is due, we suggest, to their being able to perceive and encode these distinctive characteristics of the target objects.

How do the subjects know that a particular characteristic is likely to be distinctive and useful for the performance of a recognition task? In this regard, it is important to note that whether or not an object characteristic is distinctive depends on the characteristics of the distracter objects. The white down of a gosling is distinctive with respect to a distracter set of ravens, but not in the context of other geese (where its size may be a more useful discrimination criterion). Similarly, the properties of the objects drawn from the three classes we had chosen were distinctive in the context of randomly generated wire sculptures where the possibility of accidentally creating one or more of these characteristics was very small. Since subjects had experienced several such objects in experiment 1, they could assess the distinctiveness of the object characteristics employed here. It would be interesting to compare the performance of this set of subjects with that of completely naive ones, who have had no experience at all with the general class of wire-sculptures we have here.

The distinctive-feature based scheme makes an interesting experimentally verifiable prediction. If an object's encoding emphasizes its distinctive characteristic at the expense of its other attributes, then a subject's performance on a task that required him/her to discriminate between the original object and another one that had the same distinctive characteristic but was different otherwise, would be expected to be poor (poorer, in fact, than a subject whose encoding gives equal weights to all object attributes). The other-race effect in the context of face-recognition probably has similar roots.

In summary, the results from experiment 3 suggest that the visual system is capable of encoding objects in terms of their distinctive characteristics with respect to the observed (or expected) distractor set. Such encoding can lead to a high level of recognition performance, and if the distinctive features (such as those used in this experiment) are immune to changes in viewpoint, then the resulting recognition performance is too.

General discussion

Let us briefly recapitulate the results from our three experiments. Experiment 1 demonstrated a strong bias in naive subjects towards the use of view-dependent representation schemes. Furthermore, the criterion for model to image matching appeared to be based primarily on 2D appearance similarity. To determine whether this match criterion reflects an absence of any depth encoding, we performed experiment 2 wherein subjects were required to discriminate between two objects that differed only in their depth structures. The results showed that a limited amount of depth information is indeed encoded. However, this information cannot readily be transformed to allow the subject to perform in a view-independent manner. Experiment 3 studied the use of distinctive object characteristics for recognition. The results demonstrated that such characteristics can be used by subjects to achieve a high level of recognition performance.

The result that we wish to emphasize the most is that of the first experiment, which suggests that for at least some classes of three-dimensional objects, the default representation scheme used by the human visual system is highly viewpoint-dependent. The results from experiments 2 and 3, while interesting in their own right, are susceptible to the criticism that they came about by encouraging the visual system to act in an 'opportunistic' fashion - the system had clues about what object characteristics were going to be important for the discrimination task and decided to modify its representation scheme to incorporate those attributes. While such opportunistic modifications in representations might themselves constitute a general strategy for recognition, for the present, we wish to highlight the results obtained under conditions that minimize the use of obvious short-cuts by the visual system. Under such conditions, it seems that: 1. the two-dimensional information in an image is accorded more significance than its depth structure, and 2. the internal representations are viewpoint-dependent.

In the context of these results, 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.

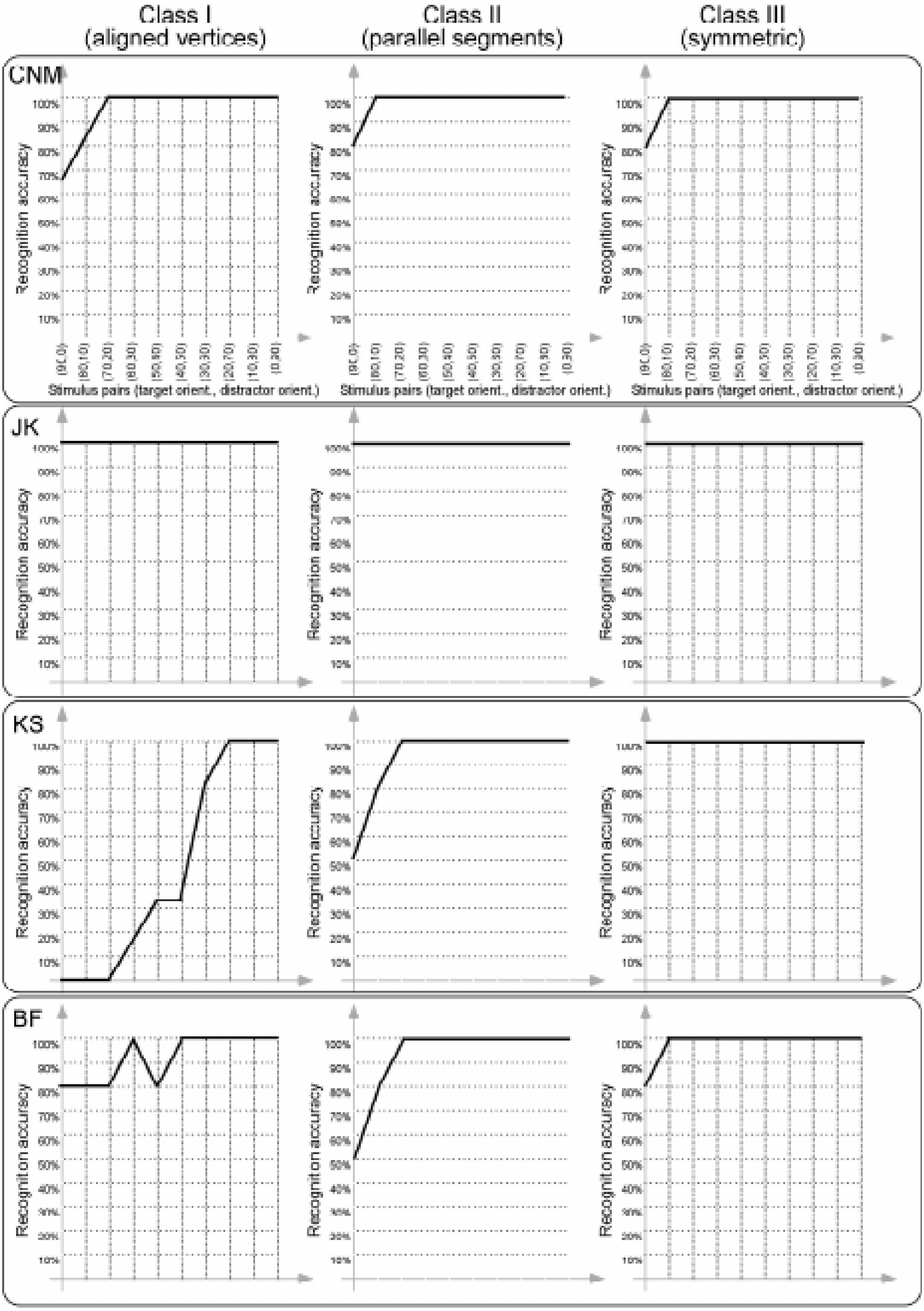

Fig 14 Results of experiment 3 from four subjects for each of the three classes of objects.

First, the bias towards view-based strategies might partly be an evolutionary vestige. Binocular vision is a relative late-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 some perceptual processes, most notably 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 about certain object attributes 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 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'. It has been argued that the brain's computational powers are limited but its memory capacity is impressive. Accordingly, a memory-intensive view-based strategy would seem more appropriate for the brain than a computation intensive transformationist strategy for object recognition.

The issue of minimizing the use of short-cuts by the visual systems calls into question the naturalness of the stimuli we used in our recognition tasks. Clearly, most objects found in real world settings have a multiplicity of cues like shape, color, texture and motion. Any one or more of these might be sufficient to discriminate between objects. An unusual hair color, for instance, might be enough to recognize a person in a crowd. The wire-sculptures we have used in our experiments are unnatural in that they possess only shape cues. The recognition task is, therefore, somewhat unusual and possibly a little difficult. However, this is intentional. As we mentioned at the outset, we wish to examine shape based recognition processes in isolation, without having extraneous cues confound the results. This design decision needs to be kept in mind while interpreting our experimental data. It is unlikely that our conclusions will be valid for arbitrary classes of objects. What we do suggest, however, is that they indicate the nature of the strategies used by the visual system when the latter is constrained to use only shape cues.

Conclusion

In this paper, we described three experiments that attempted to explore the nature of internal representations that the human visual system uses to recognize static three-dimensional objects. Specifically, we investigated whether the internal representations are viewpoint-independent or viewpoint-dependent and how much depth information about the object they encode. The conclusion we arrived at, based on the results obtained with a specific class of static 3D objects was that while the visual system is indeed capable of creating viewpoint-independent representations, its default strategy involves the use of highly viewpoint-dependent representations. Furthermore, though these representations do encode a limited amount of viewer-centered depth information, the metric for image-to-model similarity is heavily biased towards two-dimensional appearance. Congruence in 2D appearance even seems to perceptually compensate for incongruence in 3D structure.

However, the visual system is not limited to the use of holistic view-dependent representations. It is also capable of encoding objects as sets of distinctive features and can exhibit a remarkable degree of viewpoint-independence if the features themselves are immune to viewpoint variations.

In summary, it seems valid to say that the visual system is rather opportunistic, capable of using as simple or as complex an object attribute as is necessary to accomplish the task. Results from experiment 2 and experiment 3 attest to this claim. However, when the stimuli and experimental paradigm are carefully controlled for unintended clues and distinctive features, then it seems that the default shape recognition strategy adopted by the visual system is strongly biased towards the use of view-dependent representations and the match criterion emphasizes 2D similarity over 3D congruence.

Our results are based on a specific class of 3D objects and we do not yet have any firm grounds to claim that their applicability will extend to other object domains. However, results from a few other laboratories have pointed towards similar inferences [3, 10, 11, 12, 13]. In our own work, we have examined recognition strategies for a few other classes of objects. One of the most notable such classes is that of dynamic articulated objects, such as the human body in motion. Here too, we have found experimental support for the use of viewpoint-dependent representations [4, 22].

The idea that three-dimensional objects may be represented via 2D views opens up a host of interesting challenges. These include identifying the specific subsets of information in a view that are encoded for a specific recognition task [25] and also determining how to encode these views so as to obtain the maximal generalization ability [26]. Progress on understanding these issues in the context of the primate visual system will, we hope, eventually lead to better machine based recognition schemes [27].

REFERENCES

1. I Biederman, Recognition-by-components: a theory of human image understanding, *Psychological Review*, vol 94, pp 115-147, 1987**.**

2. I Biederman & P C Gerhardstein, Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance, *Journal of Experimental Psychology: Human Perception and Performance*, vol 19, pp 1162-1182, 1993.

3. H H Bülthoff & S Edelman, Psychophysical support for a two-dimensional view-interpolation theory of object recognition, *Proceedings of the National Academy of the Sciences, USA*, vol 89, pp 60-64, 1992.

4. I Bülthoff, H H Bülthoff & P Sinha, Top-down influences on stereoscopic depth-perception, *Nature Neuroscience*, 1(3), 1998.

5. L A Cooper & R N Shepard, Chronometric studies of the rotation of mental images, In W G Chase (Ed), Visual Information Processing, New York: Academic Press, 1973.

6. M C Corballis, Recognition of disoriented shapes, *Psych Review*, vol 95, pp 115-123, 1988.

7. S Edelman & H H Bülthoff, Orientation dependence in the recognition of familiar and novel views of three-dimensional objects, *Vision Research*, vol 32, pp 2385-2400, 1992.

8. J E Hummel & I Biederman, Dynamic binding in a neural network for shape recognition, *Psychological Review*, vol 99, pp 480-517, 1992.

9. G K Humphrey & S C Khan, Recognizing novel views of three-dimensional objects, *Canadian Journal of Psychology*, vol 46, pp 170-190, 1992.

10. P Jolicoeur, The time to name disoriented natural objects, *Memory and Cognition*, vol 13, pp 289-303, 1985.

11. P Jolicoeur, Identification of disoriented objects: A dual-systems theory, *Mind and Language*, vol 5, pp 387-410, 1990*a*.

12. P Jolicoeur, Orientation congruency effects on the identification of disoriented shapes, *Journal of Experimental Psychology: Human Percep and Perf*, vol 16, pp 351-364, 1990*b*.

13. N K Logothetis, J Pauls, H H Bülthoff & T Poggio, View-dependent object recognition in monkeys, *Current Biology*, vol 4, pp 401-414, 1994.

14. D Marr & H K Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, *Proceedings of the Royal Society of London Series B*, vol 200, pp 269-294, 1978.

15. R E Phinney & R M Siegel, Stored representations of three-dimensional objects in the absence of two-dimensional cues, *Perception*, vol 28, pp 725-737, 1999.

16. T Poggio & S Edelman, A network that learns to recognize three-dimensional objects, *Nature*, vol 343, pp 263-266, 1990.

17. I Rock & J Di Vita, A case of viewer centered object perception, *Cognitive Psychology*, vol 19, pp 280-293, 1987.

18. I Rock, J Di Vita & R Barbeito, The effect on form perception of change of orientation in the third dimension, *Journal of Experimental Psychology: Human Perception and Performance*, vol 7, pp 719-732, 1981.

19. I Rock, D Wheeler & L Tudor, Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, vol 21, pp 185-210, 1989.

20. R N Shepard & L A Cooper, *Mental Images and Their Transformations*. Cambridge, MA: The MIT Press, 1982.

21. R N Shepard & J Metzler, Mental rotation of three-dimensional objects, *Science*, vol 171, pp 701-703, 1971.

22. P Sinha, H H Bülthoff & I Bülthoff, View-based recognition of biological motion sequences, *Invest Ophth and Vis Science*, 36/4: #1920, 1995.

23. P Sinha & T Poggio, The role of learning in 3-D form perception, *Nature*, vol 384(6608), pp 460-463, 1996.

24. P Sinha & T Poggio, High-level learning of early perceptual tasks, In Perceptual Learning, Ed Manfred Fahle, MIT Press, Cambridge, MA, 2002.

25. P Sinha, Identifying perceptually significant features for recognizing faces, *Proceedings of the SPIE Electronic Imaging Symposium*, January, 2002*a*.

26. P. Sinha, *Qualitative representations for recognition*, In Lecture Notes in Computer Science, Springer-Verlag, LNCS 2525, 2002i.

27. P Sinha, Recognizing complex patterns, *Nature Neuroscience*, vol 5 (suppl), pp 1093-1097, 2002*c*.

28. M J Tarr & S Pinker, Mental rotation and orientation dependence in shape recognition, *Cognitive Psychology*, vol 21, pp 233-282, 1989.

29. M J Tarr & S Pinker, When does human object recognition use a viewer-centered reference frame? *Psychological Science*, vol 2, pp 207-209, 1990.

30. S Ullman, *The interpretation of visual motion*. Cambridge, MA: The MIT Press, 1979.

31. A Yip & P Sinha, Role of color in face recognition, *Perception*, vol 31, pp 995-1003, 2002.

AUTHORS