Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 12046–12050, October 1996
Psychology

Faithful representation of similarities among three-dimensional shapes in human vision

(object recognition/multidimensional scaling/computational model)

FLORIN CUTZU* AND SHIMON EDELMAN†

Department of Applied Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel

Communicated by David Mumford, Harvard University, Cambridge, MA, August 5, 1996 (received for review June 1, 1996)

ABSTRACT **Efficient and reliable classification of visual stimuli requires that their representations reside in a low-dimensional and, therefore, computationally manageable feature space. We investigated the ability of the human visual system to derive such representations from the sensory input—a highly nontrivial task, given the million or so dimensions of the visual signal at its entry point to the cortex. In a series of experiments, subjects were presented with sets of parametrically defined shapes; the points in the common high-dimensional parameter space corresponding to the individual shapes formed regular planar (two-dimensional) patterns such as a triangle, a square, etc. We then used multidimensional scaling to arrange the shapes in planar configurations, dictated by their experimentally determined perceived similarities. The resulting configurations closely resembled the original arrangements of the stimuli in the parameter space. This achievement of the human visual system was replicated by a computational model derived from a theory of object representation in the brain, according to which similarities between objects, and not the geometry of each object, need to be faithfully represented [Edelman, S. (1995) *Minds Machines* 5, 45–68; *cf.* Shepard, R. N. (1968) *Am. J. Psychol.* 81, 285–289].**

The human visual system possesses an impressive ability to remember and recognize complex three-dimensional (3D) shapes. The nature of the internal representations that underlie this ability and the degree to which they mirror geometrical reality are controversial (1–3). According to some theories, the computational basis for shape representation is faithful encoding of structural (4) or metric (5) properties of individual objects and object classes. Other theories stress the representation of similarities between objects, rather than the geometry of each object in isolation (6–8). We examined the psychophysical and computational plausibility of this latter approach to representation in a series of experiments in which human subjects were confronted with a parametrically controlled family of animal-like 3D shapes.

METHODS

The shape of each stimulus was defined by a point in a common 70-dimensional parameter space (Fig. 1). Assuming that the response times and confusion rates are related systematically to the perceptual (representation space) distances between the stimuli (9), we recovered the structure of the perceptual space by using nonmetric multidimensional scaling (MDS) (10) and compared it with that of the objective parameter space to assess the degree to which the former is a faithful representation of the latter. The planar and regular shape–space configurations formed by the stimuli (Fig. 1) were chosen to facilitate this comparison.

The psychophysical data were gathered using three different methods for estimating perceived similarity. In each pairs of pairs comparison (CPP) experiment, six or seven subjects differentially rated pairwise similarity when confronted with two pairs of objects, each revolving in a separate window on a computer screen. Subject data were pooled using individually weighted MDS (ref. 11; in all the experiments, the solutions were consistent among subjects). In each trial, the subject had to select among two pairs of shapes the one consisting of the most similar shapes. The subjects were allowed to respond at will; most responded within 10 sec. Proximity (that is, perceived similarity) tables derived from the judgments were processed to verify their degree of transitivity (4% of all triplets were found intransitive) and then submitted to MDS.

In the long-term memory (LTM) variant of this experiment, the subjects were first trained to associate a label (a three-letter nonsensical string, such as "BON" or "POM") with each object and then carried out the pairs of pairs comparison task from memory, prompted by the object labels rather than by the objects themselves. Six subjects participated in each of the two LTM experiments (Star and Triangle). The subjects were taught each shape in a separate session and had to discriminate between that shape and six similar nontargets from various viewpoints. Training continued until the recognition rate reached 90%, over a period of several days. The subjects were never exposed to more than one target in one session and were not told the ultimate purpose of the experiment. After 2 to 3 days of rest, they were tested with questions such as: "is the BON more similar to POM than TOC to ROX?", for all pairs of pairs of stimuli. In the LTM experiments, 8% of the comparisons were intransitive.

In the delayed match to sample (DMTS) experiments, pairs of static views of the same or different objects were consecutively and briefly flashed on the screen (the exposure time was 300 msec), in binocular stereo, using liquid-crystal shutter glasses synchronized with the display. The subject had to decide whether or not the two views were of the same object under different orientations or of different objects. A mask consisting of superimposed parts of animal-like shapes was displayed for 0.5 sec upon key press between the two frames in each trial, as well as between trials. The subjects received no training (that is, they were never shown the objects rotating on the screen) and no feedback during the experiment. Three or four distinct viewpoints were used for each object, depending on the type of experiment, to keep the length of the experimental sequence within reasonable limits. The response time

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: 3D, three-dimensional; 2D, two-dimensional; MDS, multidimensional scaling; CPP, compare pairs of pairs; LTM, long-term memory; DMTS, delayed matching to sample; RBF, radial basis functions.

*Present address: Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912.

†To whom reprint requests should be addressed.

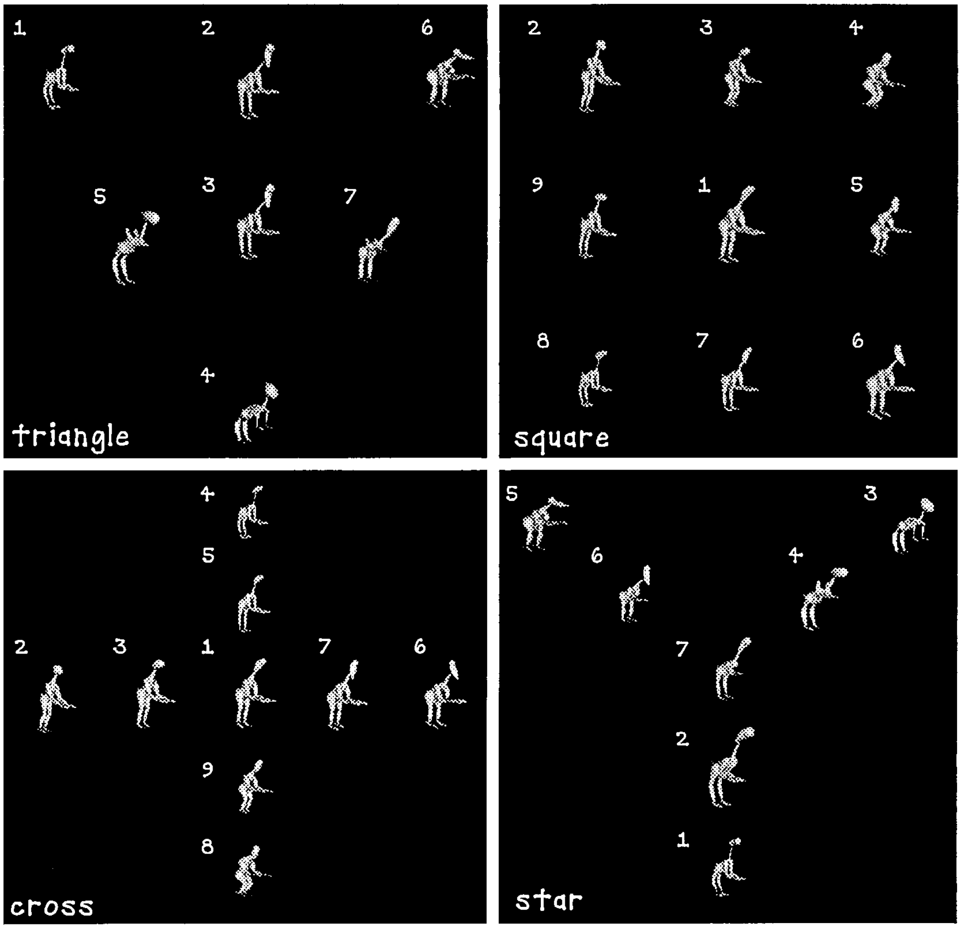

FIG. 1. The four planar parameter-space configurations illustrating the similarity patterns built into the experimental stimuli. All these animal-like shapes had the same parts, modeled by generalized cylinders, and were controlled by 70 parameters that determined the geometry of the parts, as well as their mutual arrangement, via nonlinear functions (thus, the image of a shape situated at the midpoint between two other shapes in the parameter space could not be obtained by image-space interpolation). The orientation of the plane defined by each configuration in the parameter space was arbitrary with respect to the axes; that is, all the 70 parameters were varied in generating the stimuli. Furthermore, the parameterization itself was generic, as verified by a control experiment, in which we used MDS to recover the parameter-space configurations (Triangle, Star, etc.) from interobject distances, computed in the space of the vertices of the high-resolution 3D triangular mesh encoding the detailed geometry of each object. During the experiments, the shapes were rotated in 3D space and rendered on the screen of a computer workstation (SGI Indigo 2).

and error rate data were entered into a proximity table (as described in an unpublished report) and were submitted to MDS.

PSYCHOPHYSICAL RESULTS

In the CPP experiments, the parameter-space configurations built into the stimuli (Cross, Star, etc.) were easily recognizable in the MDS plots (a typical example appears in Fig. 2). Procrustes analysis (a technique for measuring nonrigid distortion of one pattern relative to another; see ref. 14) indicated that the similarity between the MDS-derived and the objective configurations was significantly above chance, as estimated by bootstrap analysis (that is, by comparing the experimentally obtained similarity with a Monte Carlo estimate of its value as expected by chance; see ref. 15). Similar results were obtained in the speeded-discrimination DMTS experiments, where the subjects were able to group together different views of the same object and to tell apart views of different objects, responding correctly in about 75% of the trials, despite the absence of prior exposure to the objects. In each DMTS experiment, MDS was applied to obtain separate view-wise and object-wise solutions; in the former, each point corresponded to an individual view of some object and, in the latter, each point corresponded to an entire object. The configurations recovered by MDS (of points in the object-wise solutions and of point clusters, each corresponding to the different views of the same object, in the view-wise solutions) closely resembled the shape-space configurations built into the stimuli (Fig. 3). Moreover, the clustering of views in the view-wise MDS solutions was consistent with the subjects' performance, as shown by nonparametric (nearest-neighbor) discriminant analysis (16). The parameter-space configurations of the stimuli were also recovered in the LTM experiments (Fig. 2), in which the subjects could not rely on immediate percepts or short-term memory representations of the stimuli (*cf.* ref. 7).

Note that none of the complex shape-space configurations we have tested was ever revealed to the subjects in its entirety. The two dimensions of variation built into the stimuli were well hidden, first in the 70 dimensions of the parametric shape space and then in the quarter of a million or so of the pixel-wise dimensions of the images rendered on the screen. Moreover, the relationship between the parametric representation of a stimulus and its physical appearance (i.e., the values of the pixels in an image of the stimulus) was highly nonlinear: a point halfway between two shapes in the parameter space did not correspond to an image-space interpolation between the views of the shapes. The computational feat of the recovery of the two relevant dimensions is much more difficult than finding the proverbial needle in a haystack [this difficulty is a manifestation of a general property of high-dimensional spaces that has

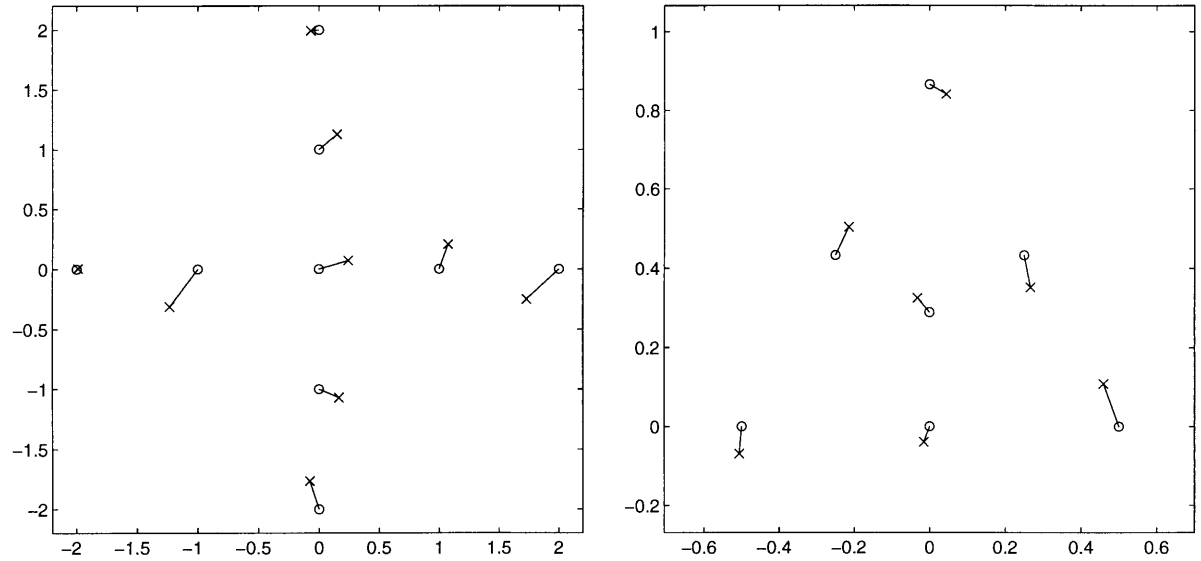

FIG. 2. (*Left*) The CPP experiment, configuration Cross; the two-dimensional (2D) MDS solution for all subjects; stress (12) 0.14 (after ref. 13; reproduced by permission). Symbols: open circle, true configuration; ×, configuration derived by MDS from the subject data, then Procrustes-transformed (14) to fit the true one; lines connect corresponding points. To quantify the visual impression of similarity between objective and data-derived configurations, we computed the optimal Procrustes transformation (combination of scaling, rotation, reflection, and translation) between the MDS-derived and the true configurations. The residual distance that remained after Procrustes transforming the MDS-derived configuration to fit the true one was 0.66 [expected random value, estimated by bootstrap (15): 3.14 ± 0.15, mean and SD; 100 permutations of the point order were used in the bootstrap computation]. We also computed the coefficient of congruence (a correlation-like measure applied to interpoint distances; ref. 14) between the two configurations: 0.99 (expected random value: 0.86 ± 0.03). (*Right*) The LTM experiment, configuration Triangle, all subjects (stress 0.12). Coefficient of congruence, 0.99 (expected random value: 0.87 ± 0.04); Procrustes distance, 0.18 (expected random value: 0.78 ± 0.05).

been termed "the curse of dimensionality" (17); for an illustration of problems associated with finding structure in multidimensional spaces, see ref. 18]. This feat was performed by the subjects' visual system; the role of MDS was merely to help visualize the relevant information present in the subjects' response patterns.

The ability of the subjects to represent the low-dimensional pattern of similarities among stimuli does not extend to nonsense objects, as indicated by the results of CPP and DMTS control experiments involving "scrambled" shapes (ref. 19 and unpublished results). The stimuli in these experiments were obtained by translating the parts of the animal-like shapes to

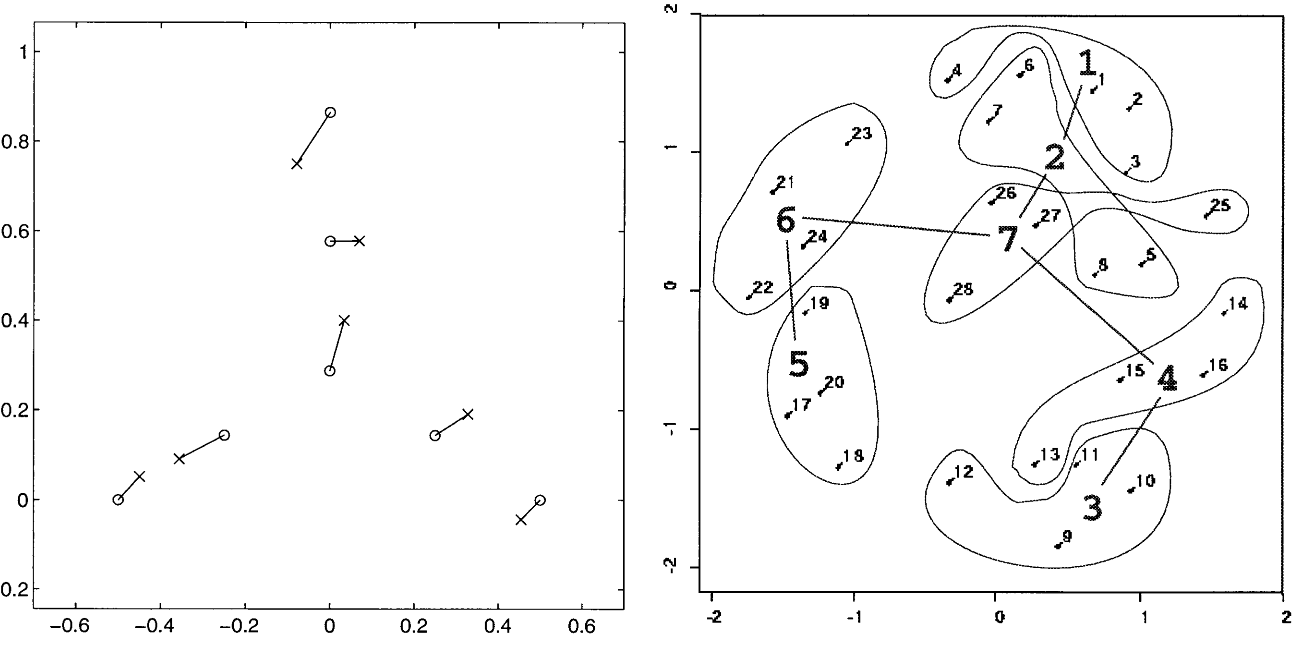

FIG. 3. The DMTS experiment, configuration Star. (*Left*) The 2D MDS object solution (stress 0.14). Coefficient of congruence, 0.98 (expected random value, 0.86 ± 0.04); Procrustes distance, 0.25 (expected random value, 0.77 ± 0.09). (*Right*) The 2D MDS views solution (stress 0.33); views of each object are enclosed in a common contour. The 28 points correspond to 7 objects × 4 views per object. The views are clustered by object identity: the error rate estimated by nearest-neighbor discriminant analysis was 28.6%; the actual pooled-subjects mean error rate was 25.2% [note that the two are not directly comparable; the former is for 1-out-of-7 classification, while the latter is for (easier) same/different decision].

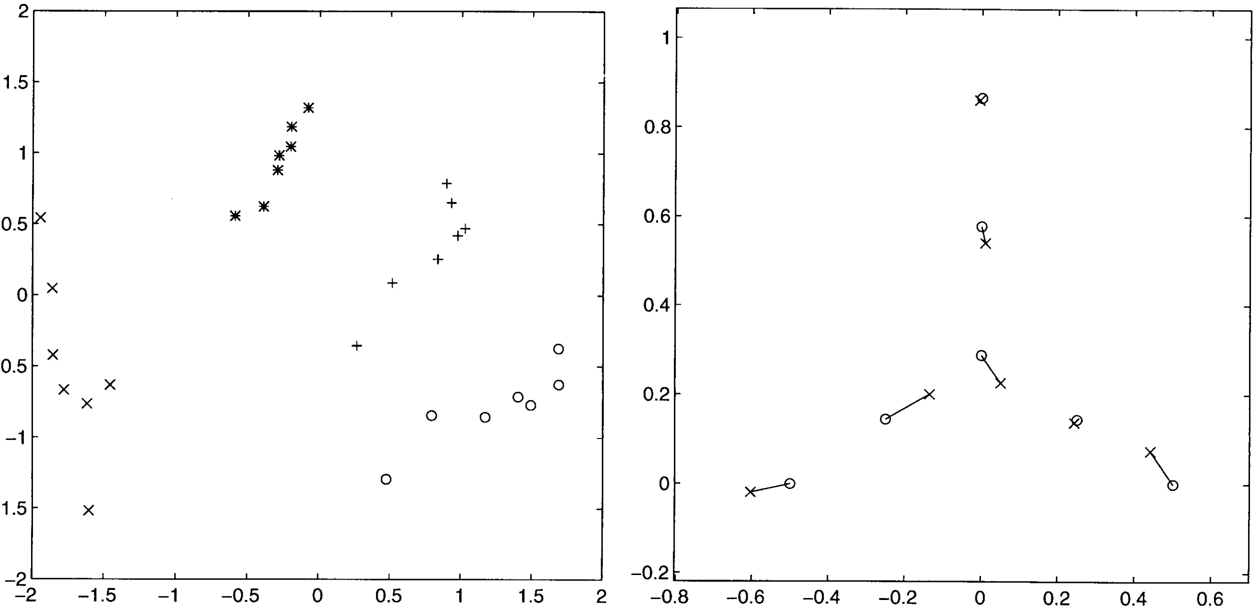

FIG. 4. Simulated DMTS experiments, Star configuration. (*Left*) The image-based model: views solution derived from similarities among receptive field activities evoked by the stimuli. Note the four clusters, one for each view (not object), marked by the different symbols. Each cluster contains seven points, one for each object, in the same view. The object solution (not shown) was essentially random: coefficient of congruence, 0.89 (expected random value, 0.86 ± 0.04); Procrustes distance, 0.64 (expected random value, 0.78 ± 0.07). (*Right*) The object-based model: object solution derived from similarities among RBF network activities evoked by the stimuli. The model contained three RBF networks, one for each extremal vertex. Herein, the clustering in the views solution (not shown) was by object identity, not by orientation. The object solution was highly significant: coefficient of congruence, 0.99 (expected random value, 0.86 ± 0.04); Procrustes distance, 0.17 (expected random value, 0.76 ± 0.09).

a common center, resulting in star-like nonsense objects. For these objects, Procrustes similarity between true and MDS-recovered configurations was consistently lower than for animal-like shapes.

We note that the relevance of the MDS solutions derived from the experimental data is supported by the bootstrap-validated Procrustes analysis and not by mere low stress. To assess the power of this approach, we conducted a 50-trial Monte Carlo study involving seven points in 70 dimensions (*i*) in a randomly oriented planar (Triangle) configuration and (*ii*) in a random configuration. The repeated application of MDS to these statistically controlled data allowed us to ascertain the reliability of the figures of merit we had used to judge the quality of the solutions. Specifically, we found that in example *i*, the Procrustes distance between true and MDS-recovered configurations was, on the average, 2.83 standard deviations below the mean expected random-configuration distance (as estimated by bootstrap); in example *ii*, the average difference was 0.08 standard deviations. Thus, if the pattern to be recovered is known in advance (as it is in our experiments), MDS can be applied in confirmatory mode (14) and can produce reliable results with as few as seven points.

COMPUTATIONAL MODELS

To elucidate the possible computational basis for the performance of our subjects, we replicated the DMTS experiments with two computer models of similarity perception. In the first model, designed to illustrate the behavior of a raw image-based measure of similarity, object views were convolved with an array of overlapping Gaussian receptive fields. The proximity table for each parameter-space configuration was constructed by computing the Euclidean distances between the views, encoded by the activities of the receptive fields. In the MDS-derived view-wise configurations, views of different objects were grouped together by object orientation, not by object identity (Fig. 4 *Left*). Thus, a simple image-based representation (which may be considered roughly analogous to an initial stage of processing in the primate visual system, such as the primary visual area V1) could not reproduce the results observed with human subjects.

Our second model corresponded to a higher stage of object processing, in which nearly viewpoint-invariant representations of familiar object classes (but, presumably, not of nonsense shapes such as those in our control experiments; see refs. 20 and 21) are available; a rough analogy is to the inferotemporal visual area IT (22, 23). Such a representation of a 3D object can be relatively easily formed, given several views of the object (24), e.g., by training a radial basis function (RBF) network to interpolate a characteristic function for the object in the space of all views of all objects (25). Responses of several such object-specific modules, each coarsely tuned to a different reference shape, may be able to support veridical representation of a range of shapes similar to the reference ones (19, 26). We chose a number of reference objects (e.g., in the Star configuration, the three corners were used) and trained an RBF network to recognize each such object (i.e., to output 1.0 for any of its views, encoded by the activities of the underlying receptive field layer). At the RBF level, the (dis)similarity between two stimuli was defined as the Euclidean distance between the vectors of outputs they evoked in the RBF modules trained on the reference objects. Unlike in the case of the simple image-based similarity measure realized by the first model, the MDS-derived configurations obtained with this model showed significant resemblance to the true parameter-space configurations (Fig. 4 *Right*). Computational considerations, described in detail in ref. 27, suggest that this was due to a combination of (*i*) the monotonic dependence of each module's output on the parameter-space dissimilarity between its preferred object and the stimulus, and (*ii*) the relative independence of the module's output on the orientation of the stimulus. Thus, veridical representation is a generic property of a relatively wide class of systems that includes the RBF ensemble model as a special case.

DISCUSSION

Although the recovery of the metric dimensions of stimulus variation has been demonstrated in the past in a wide range of perceptual tasks (10), the full power of MDS as a tool for mapping the internal representation space of subjects can be realized only if the experimental approach constrains the interpretation of the potential outcome to a sufficient degree (cf. ref. 28). An application of MDS always produces a solution that, furthermore, will have a low stress (i.e., low residual discrepancy with the data) if the number of points is small. It is important to realize that our conclusions are based not on low stress *per se* but rather on the recovery of a specific pattern built into the stimuli (see the note in Fig. 2). This effectively allowed us to invoke MDS in a confirmatory mode (14) and to demonstrate the statistical significance of the solution.

We note that the remarkably faithful reconstruction of the parameter-space arrangements of the stimuli from the subject data would have been impossible if the subjects stored merely the distinctive features of each shape, although successful recognition would still be possible in this case. All our objects shared the same structural description; thus, our results cannot be explained by Biederman's Recognition By Components theory (4), according to which objects are represented in terms of the structural layout of generic parts. On a more positive note, our findings suggest that the biological substrate for object representation may be not unlike a "chorus of prototypes"—an ensemble of recognition mechanisms, each coarsely tuned to a reference shape (8). This interpretation is consistent with the results of recent single-unit studies of the inferotemporal cortex in the monkey (23, 29, 30).

We thank N. Intrator, D. Mumford, R. N. Shepard, T. Poggio, M. Tarr, K. Tanaka, and S. Ullman for discussions. This research was supported in part by a grant from the Basic Research Foundation, administered by the Israel Academy of Arts and Sciences. S.E. is an incumbent of the Sir Charles Clore Career Development Chair at the Weizmann Institute of Science.

1. Palmer, S. E. (1978) in *Cognition and Categorization*, eds. Rosch, E. & Lloyd, B. B. (Erlbaum, Hillsdale, NJ), pp. 259–303.
2. Cummins, R. (1989) *Meaning and Mental Representation* (MIT Press, Cambridge, MA).
3. Tanaka, K. (1993) *Science* **262,** 685–688.
4. Biederman, I. (1987) *Psychol. Rev.* **94,** 115–147.
5. Ullman, S. (1989) *Cognition* **32,** 193–254.
6. Shepard, R. N. (1968) *Am. J. Psychol.* **81,** 285–289.
7. Shepard, R. N. & Chipman, S. (1970) *Cognit. Psychol.* **1,** 1–17.
8. Edelman, S. (1995) *Minds Machines* **5,** 45–68.
9. Tomonaga, M. & Matsuzawa, T. (1992) *J. Comp. Psychol.* **106,** 43–52.
10. Shepard, R. N. (1980) *Science* **210,** 390–397.
11. Carroll, J. D. & Chang, J. J. (1970) *Psychometrika* **35,** 283–319.
12. Kruskal, J. B. & Wish, M. (1978) *Multidimensional Scaling* (Sage, Beverly Hills, CA).
13. Edelman, S., Cutzu, F. & Duvdevani-Bar, S. (1996) *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (Erlbaum, Hillsdale, NJ), pp. 260–265.
14. Borg, I. & Lingoes, J. (1987) *Multidimensional Similarity Structure Analysis* (Springer, Berlin).
15. Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap* (Chapman & Hall, London).
16. SAS Institute (1989) *SAS/STAT User's Guide* (SAS Institute, Cary, NC), Version 6.
17. Bellman, R. E. (1961) *Adaptive Control Processes* (Princeton Univ. Press, Princeton).
18. Huber, P. J. (1985) *Ann. Stat.* **13,** 435–475.
19. Edelman, S. (1995) *Neural Comput.* **7,** 407–422.
20. Edelman, S. & Bülthoff, H. H. (1992) *Vision Res.* **32,** 2385–2400, 1992.
21. Bülthoff, H. H. & Edelman, S. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 60–64.
22. Young, M. P. & Yamane, S. (1992) *Science* **256,** 1327–1331.
23. Logothetis, N. K., Pauls, J. & Poggio, T. (1995) *Curr. Biol.* **5,** 552–563.
24. Ullman, S. & Basri, R. (1991) *IEEE Trans. Pattern Anal. Machine Intell.* **13,** 992–1005.
25. Poggio, T. & Edelman, S. (1990) *Nature (London)* **343,** 263–266.
26. Edelman, S. (1996) in *Vision*, ed. Watt, R. (MIT Press, Cambridge, MA), in press.
27. Edelman, S. & Duvdevani-Bar, S. (1996) *Neural Comput.* **8,** in press.
28. Shepard, R. N. & Cermak, G. W. (1973) *Cognit. Psychol.* **4,** 351–377.
29. Tanaka, K. (1992) *Curr. Opin. Neurobiol.* **2,** 502–505.
30. Sakai, K., Naya, Y. & Miyashita, Y. (1994) *Learning Memory* **1,** 83–105.