Psychobiology
1998, 26 (4), 309-321

Toward direct visualization of the internal shape representation space by fMRI

SHIMON EDELMAN
University of Sussex, Falmer, England

KALANIT GRILL-SPECTOR
Weizmann Institute of Science, Rehovot, Israel

TAMMAR KUSHNIR
The Chaim Sheba Medical Center, Tel Hashomer, Israel

and

RAFAEL MALACH
Weizmann Institute of Science, Rehovot, Israel

Reports of columnar organization of the macaque inferotemporal cortex (Tanaka, 1992, 1993a) indicate that ensembles of cells responding to particular objects may be both sufficiently extensive and properly localized to allow their detection and discrimination by means of functional magnetic resonance imaging (fMRI). A recently developed theory of object representation by ensembles of coarsely tuned units (Edelman, 1998; Edelman & Duvdevani-Bar, 1997b) and its implementation as a computer model of recognition and categorization (Cutzu & Edelman, 1998; Edelman & Duvdevani-Bar, 1997a) provide a computational framework in which such findings can be interpreted in a straightforward fashion. Taken together, these developments in the study of object representation and recognition suggest that direct visualization of the internal representations may be easier than was previously thought. In this paper, we show how fMRI techniques can be used to investigate the internal representation of objects in the human visual cortex. Our initial results reveal that the activation of most voxels in object-related areas remains unaffected by a coarse scrambling of the natural images used as stimuli and that a map of the representation space of object categories in individual subjects can be derived from the distributed pattern of voxel activation in those areas.

Recognition of visual objects over successive encounters requires that the visual system maintain representations of objects in long-term memory. The main computational challenge posed by this task is the variability in the appearance of objects. A given object will look different, depending on viewing conditions such as the illumination and the orientation of the object with respect to the observer. Moreover, different exemplars of the same visual category may look different, yet must be treated similarly, even when encountered for the first time. Any visual system, natural or artificial, must strive to compensate for the variability due to viewing conditions before attempting to recognize an object—that is, to compare it to an internal representation stored in memory. Furthermore, the format of the memory trace must be such that the variability among exemplar objects is represented explicitly, to make possible not only coarse categorization but also fine distinction among stimuli. Several theories of object recognition and representation that attempt to address these issues have been proposed in the past (see, e.g., Biederman, 1987; Ullman, 1996; see Edelman, 1997, for a review). Both their compatibility with general principles of biological information processing and the possibility of their mapping onto the experimentally determined mechanisms of primate vision remain debatable (Logothetis & Scheinberg, 1996; Tanaka, 1996).

The advent of new imaging techniques, such as functional magnetic resonance imaging (fMRI) and PET, made it possible to study noninvasively the activation evoked by various stimuli in the human brain, in an attempt to characterize directly the nature of object representation in the human visual system. For such an attempt to succeed, the experimenter must have a notion of the mechanism whereby the response to a stimulus is generated, so that the stimuli can be manipulated in an appropriate manner. Most of the current theories of recognition, however, are not specific enough in their predictions of the response properties of large assemblies of cells (the only quantity that can be measured directly by imaging techniques).[1] Thus, as frequently happens when the availability of an experimental tool precedes the development of a detailed model of the process that is subjected to scrutiny (Barlow, 1990), the results of the experiments seem to be conflicting and difficult to interpret.

Correspondence concerning this article should be addressed to S. Edelman, School of Cognitive and Computing Science, University of Sussex, Falmer, Brighton BN1 9QH, England (e-mail: shimone@cogs.susx.ac.uk).

A solid basis for the interpretation of the fMRI data on object recognition in humans is provided by the identification of a region in the lateral occipital (LO) cortex that is activated by images of objects much more strongly than by random dot patterns, repetitive textures, or degraded images (Malach et al., 1995). The same region appears to respond to line drawings of unfamiliar objects more than to visual noise or to highly scrambled line drawings (Kanwisher, Woods, Iacoboni, & Mazziota, 1997). Furthermore, it also appears (Grill-Spector et al., 1998) that the LO region is essentially nonretinotopic in that it integrates information from both the ipsilateral and the contralateral visual fields, indicating that it may be the human homologue of the monkey inferotemporal (IT) cortex (Tanaka, 1997).

Whereas the existence and the location of the object-related areas in humans seem to be a matter of consensus, their presumed principles of operation are controversial. One suggestion stemming from many of the recent studies is that the object areas are subdivided into regions that respond preferentially to certain categories of objects. Thus, significant efforts are currently directed toward the establishment of a topographic characterization of the object areas, in which processing of specific categories, such as faces (Allison, Ginter, McCarthy, Nobre, & Puce, 1994; Allison, McCarthy, Nobre, Puce, & Belger, 1994; Haxby et al., 1994; Kanwisher, Chun, McDermott, & Ledden, 1996; Kanwisher, McDermott, & Chun, 1997), certain object categories (Ishai, Ungerlieder, Martin, Maisog, & Haxby, 1997; Martin, Wiggs, Ungerleider, & Haxby, 1996), or letter strings (Puce, Allison, Asgari, Gore, & McCarthy, 1996); would be associated with specific regions of the cortex. The experimental evidence for such parcellation remains, however, elusive; even when a differential response across categories is found, areas that are shown to be selectively responsive to a certain object category also exhibit a nonnegligible response to stimuli from other object categories (as compared with a low baseline response to texture patterns). Category-based parcellation is also not favored by the neuropsychological data summarized in Farah (1990): "superselective category-specific deficits ... appear to result from impairments outside the the visual recognition system, as they are confined to lexical/semantic operations" (p. 85). Even in face recognition, a function most frequently postulated to be narrowly localized in the brain, the deficits are not category specific: "in prosopagnosia the impairment encompasses some subset of faces, animals, buildings, clothing, and makes of automobile" (Farah, 1990, p. 123).

A Framework for the Understanding of fMRI Results Concerning Object Representation

An intriguing alternative to the hypothesis of anatomical localization by object category emerges when one examines the same imaging data reported in the above studies from a different perspective. In a typical experiment, dozens of voxels are found to respond to objects but not to simpler control stimuli. Adopting the notion of coarse coding (Hinton, 1984), one may hypothesize that it is the *relative activation levels of many voxels at a specific time* that signal the category to which the stimulus belongs and, possibly, its identity.

It is important to realize that coarse coding per se must remain a mere hypothesis, unless it is substantiated on two levels: computational and implementational. On the computational level, a theory is needed that would show how, precisely, a distributed coarse code could work. Specifically, such a theory must address the two issues having to do with the variability of object appearance mentioned earlier—namely, the effects of viewing conditions and of the within-category shape variation. On the implementational level, evidence is needed to the effect that the cellular mechanisms identified by neurobiological means (e.g., electrophysiological and fMRI studies) can support the functionality required by the theory.

Coarse coding for object representation: A computational framework. The variability of object appearance with viewing conditions can be countered, to a large extent,[2] by storing a number of views of the object and attempting to recognize another (potentially novel) view as an interpolation of the stored ones (Poggio & Edelman, 1990). This approach provides the necessary computational basis for a distributed code for visual objects, whose relevance to biological vision is indicated by psychophysical studies (Bülthoff, Edelman, & Tarr, 1995; Logothetis, Pauls, Bülthoff, & Poggio, 1994) and by electrophysiological evidence (Logothetis, Pauls, & Poggio, 1995; Miller, Li, & Desimone, 1993). It does not, however, address the problem of dealing with novel shapes, for which no stored views are available. An extension of the view interpolation idea that can support the representation and processing (e.g., categorization) of novel objects was proposed in Edelman (1995). Whereas, in the view interpolation approach, a novel view is processed on the basis of its similarity to several stored views, the extended method represents a novel object on the basis of its similarity to several stored object traces (each of which, in turn, consists essentially of several stored views; see Figure 1). Empirical support for this representational scheme, whose theoretical underpinnings are discussed elsewhere (Edelman & Duvdevani-Bar, 1997b), stems from two sources. First, a computer implementation of this scheme was shown to perform well both in the traditional recognition task of generalization to novel views of familiar objects and in the more challenging tasks of categorization (Edelman & Duvdevani-Bar, 1997a) and analogy-like generalization from a single view of novel objects (Duvdevani-Bar, Edelman, Howell, & Buxton, 1998). Second, this scheme proved to be capable of modeling psychophysical results concerning human performance in a variety of object (Cutzu & Edelman, 1996, 1998) and face (O'Toole, Edelman, & Bülthoff, 1998) recognition tasks.

In this scheme, the computation of similarities between the stimulus and the stored objects is carried out

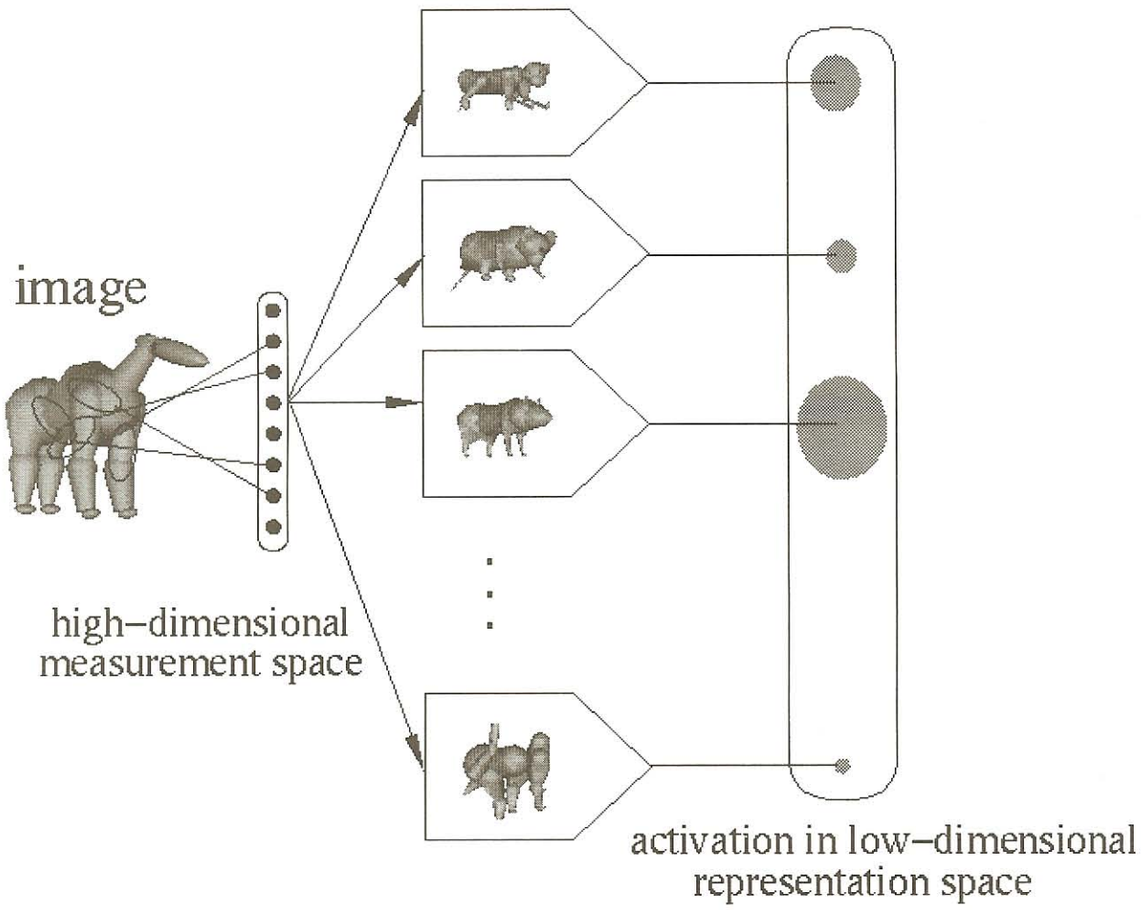

Figure 1. Representation by a chorus of prototypes. An implementation of this scheme consists of two layers. The first of these is composed of a bank of receptive fields, which map the stimulus into a high-dimensional measurement space. The second layer consists of a number of prototype (reference object) modules, each of which encodes a particular class of objects by some of their views (not shown). The stimulus, therefore, is represented by the distributed pattern of activation that it induces across the prototype modules. The level of activation of each module (indicated here symbolically by the size of the circle drawn near it) encodes the degree of similarity between the stimulus and the prototypical object of that module. (See Edelman & Duvdevani-Bar, 1997a, for details.)

in a distributed fashion, by modules tuned to those objects. The resulting representation is also distributed—an object (familiar or novel) is represented by the pattern of activities its views elicit in the existing modules. Note that each module must be trained to respond to shape changes more strongly than to changes in the viewing conditions of objects and that its shape selectivity profile must be wide enough, so that several modules respond to each stimulus (Edelman & Duvdevani-Bar, 1997b).

It is convenient to think of this representation scheme as a *shape space*, in which each point (i.e., vector of activities of the modules) corresponds to some object. If this space is endowed with a distance measure, various tasks, such as categorization (clustering) and recognition (pinpointing a location), become possible. Most importantly, a novel object is projected (by activating the various modules to varying degrees) into some point in the shape space, which can then be attributed to the nearest cluster (category) or used to create a new category, if it falls too far from any of the familiar ones.

The nominal dimensionality of the shape space is equal to the number of modules that span this space. However, the actual dimensionality might be much smaller. To realize this, consider the table of all pairwise distances of a set of cities, as measured off a map. Although the nominal dimensionality of this data set is equal to the number of cities, the actual dimensionality is equal to two, and can be recovered by a proper algorithm, such as multidimensional scaling (Kruskal, 1964; Kruskal & Wish, 1978; Shepard, 1962). It should be noted that the location of an additional city on the map can be represented by its distances to any subset of (three or more) cities. In the context of the above shape representation scheme, the number of modules is equivalent to the number of cities

(nominal dimensionality), whereas the actual dimensionality might be much lower (in the example, the actual dimensionality was two). An object is, therefore, representated by the distributed activation elicited by the modules that span the shape space. (Note that an implicit assumption is that the activation is proportional to the distance between the object and the module). It should not be taken to mean that the categories for which modules exist need to be the same for all representational systems; consistency among individuals only requires that the different choices of reference categories lead to sufficiently close values of *similarities* among stimuli in the resulting shape spaces (Edelman, 1998).

We are now in a position to formulate a computationally sound and empirically testable framework for the interpretation of fMRI results concerning object representation: (1) Objects are represented by a relatively widely distributed activity of functional modules. (2) Pairwise distances among objects, computed in the space of activities of the functional modules, correspond to their pairwise (dis)similarities, defined geometrically or psychophysically. (3) Activity related to the functional modules mentioned above may be observable in the activation level of voxels responding to objects (but not to textures) in fMRI experiments. The first statement seems to be true of the reported data obtained in fMRI studies; it is, however, phrased in terms that are insufficiently quantitative to allow, say, a statistical test. In comparison, the second point is a concrete prediction: Given a set of stimuli, the configuration they form in the voxel activation space (or in its low-dimensional replica) can be matched quantitatively against the configuration derived from geometry or from psychophysics, and the significance of the match can be tested statistically (Cutzu & Edelman, 1996). The validity of the observability assumption in the third point is the subject of the fMRI experiments described in the next section.

Coarse coding for object representation: Evidence for the necessary functional architecture. A reason to believe that the pattern of activities of the hypothesized tuned modules can be observed by means of fMRI is provided by the columnar structure of the IT cortex in the monkey (Fujita, Tanaka, Ito, & Cheng, 1992; Tanaka, 1992, 1993b, 1996). In a series of experiments, Tanaka and his collaborators found that IT cells, which respond selectively to various object features, cluster in columns that run perpendicular to the cortical surface, so that cells in the same column tend to respond to similar (but not identical) features. Although the columns, whose size is about 0.5 mm (Wang, Tanaka, & Tanifuji, 1996), are too small to be individuated by fMRI means, patterns of activation of several columns should be amenable to visualization. Specifically, even if each one of the fMRI voxels overlaps several columns, the ensemble of several voxels will carry some information concerning the relative activities of several adjacent columns.[3] The pattern of activities of such columns, which in our interpretation corresponds to the activities of the tuned chorus-like "modules," may then be observable by state-of-the-art fMRI means.

SCRAMBLING EXPERIMENT

A major issue that remains unclear regarding the LO complex in humans and the IT cortex in monkeys is the level of complexity of the features represented there. One possibility here is that entire objects are represented by tightly coupled neuronal clusters—the hypothesis favored by the chorus model. Alternatively, object fragments can constitute the "alphabet" out of which representations of entire objects are constructed. We take *fragments* to mean parts of images, as in "the nose occupies a middle part in a typical face image." Operationally, the extraction of such fragments is a matter of focusing attention and narrowing it down to a proper level—a notion supported by considerable behavioral (Keele & Neill, 1978; Nissen, 1985) and electrophysiological (Chelazzi, Miller, & Desimone, 1993; Moran & Desimone, 1985; Spitzer, Desimone, & Moran, 1988) evidence. Volumetric parts defined in an object-centered reference frame (as in Biederman, 1987) are not excluded in principle but are unlikely to play a significant role, given their lack of privileged attentional status (Brown, Weisstein, & May, 1992) and the computational difficulty of their reliable extraction from raw images (Edelman & Weinshall, in press).

To investigate the effects of the level of complexity or the *scale* of object features on the activation of object-related areas, we designed an experiment (Grill-Spector et al., 1998) in which images of objects were repeatedly fragmented until the independently characterized object areas in the human visual cortex ceased to respond to them. In this *scrambling* experiment, gray-level images of faces or animals were randomly scrambled into an increasing number of blocks (see Figure 2). All the images were low-pass filtered with a finite impulse response filter (cutoff frequency = 15 cycles per image, window size = 21 × 21 pixels) to reduce changes in the spatial frequency spectrum caused by the scrambling process. Epochs of visual stimulation lasting 40 sec were alternated with blank epochs (20 sec long). The subjects were asked to covertly name the visual stimuli, including the scrambled squares. A scrambling index was defined in order to compare the sensitivity of various visual areas to image scrambling. Thus, the scrambling index = (average fMRI signal during most scrambled epoch − blank) / (average fMRI signal during unscrambled epoch − blank). Note that areas that are essentially unaffected by image scrambling should have an index that is close to 1.0, whereas areas that are affected by image scrambling should have an index smaller than 1.0.

Experimental Results

The scrambling experiment differentiated three main foci of activation arranged mediolaterally in both hemispheres of the occipital cortex, as is depicted in Figure 3. The medial focus,[4] located over the calcarine sulcus and

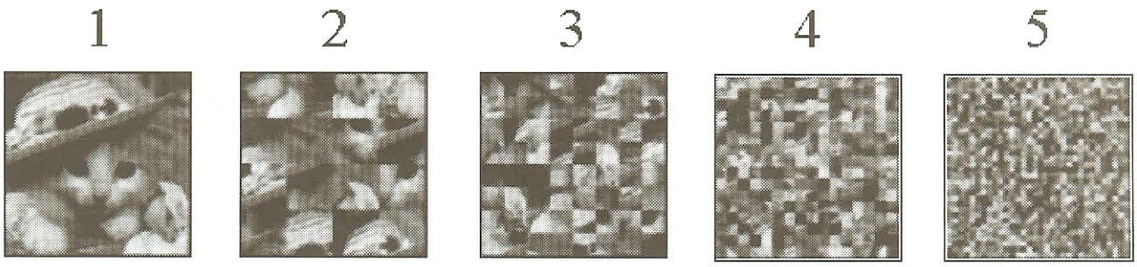

Figure 2. Examples of the images used in the scrambling experiment. Epochs of visual stimulation, which contained 20 different images depicted in 256 gray levels, were presented at a rate of 0.5 Hz and were alternated with blank epochs lasting 20 sec. The visual stimuli were low-pass filtered in all the epochs (this is not visible in the reduced-size images shown here). Visual epochs contained the same images, which were increasingly scrambled into 16, 64, 256, and 1,024 random blocks in epochs 2, 3, 4, and 5, respectively. The first epoch contained entire images of animals or faces.

the medial surface of the occipital lobe corresponding to areas V1–V3 (blue in Figure 3), showed no reduction in activity and even a slight enhancement with mild picture scrambling (see Figure 3b, top, for the average time course). More laterally, V4v (yellow in Figure 3) showed a reduction in the two highly scrambled epochs (256 and 1,024 fragments), as depicted in Figure 3b, middle. Most laterally, LO voxels (red in Figure 3) showed the highest sensitivity to scrambling, in that reduction in activation was achieved with less scrambling, as compared with V4v, and the percent of reduction was greater (see Figure 4).

Examining the behavior of LO voxels in detail, we found that in the majority of LO voxels breaking the pictures into 16 scrambled squares did not cause a severe decrease in activation (see Figure 3, bottom). Thus, the overall activation in LO under this level of scrambling was 82% ± 6% (standard error of the mean, *SEM*) of the maximal activation. However, in a minority (28% ± 9%) of LO voxels, there was a larger degree of reduction (32% ± 5%, time course not shown) for the same scrambled stimuli.

To control further for spatial frequency or edge effects, a second experiment (12 subjects) was conducted in which images were also scrambled, but instead of low-pass filtering the images, a grid was overlaid on the unscrambled images. In another experiment (5 subjects), the order of the epochs was permuted, so that the highly scrambled epochs were presented first. To compare quantitatively the sensitivity to scrambling in the three foci of activation mentioned above, we calculated the scrambling index of areas V1, V4v, and LO for the variations of the scrambling experiment, as is illustrated in the histogram in Figure 4. The results of these control experiments were similar (see Figure 4), indicating that spatial frequency, number of edges, fatigue, and adaptation effects do not account for the scrambling results. LO voxels yielded the lowest scrambling ratio, corresponding to the highest sensitivity to image scrambling; V1 voxels had the highest ratio, as they responded largely to the same extent to scrambled and to entire images.

The use of a gradual scrambling paradigm enabled us to distinguish areas V4v and LO: V4v was highly activated even when the image was fragmented into 64 blocks, whereas the activation LO was significantly reduced by this degree of image scrambling. Thus, the present results suggest that area V4v plays a role in intermediate-scale shape representation. The fact that most LO voxels remained active after the first scrambling indicates the predominance of the representation of something like object fragments rather than entire objects in LO. This result is in line with several physiological studies (Kobatake & Tanaka, 1994; Wachsmuth, Oram, & Perrett, 1994). It should be noted that this result does not necessarily contradict the idea of category-specific representations. For example, it is possible that object fragments common to faces, animals, and so forth are clustered in distinct anatomical subdivisions of the LO complex. Moreover, we did find a smaller subpopulation of LO voxels in which the activation was reduced with the first image scrambling, suggesting the existence of a subdivision of LO in which entire objects are represented.

SHAPE SPACE EXPERIMENT

In this experiment, we explored (1) the viability of distributed coarse coding as a model of object category representation in human visual areas and (2) the relationship between human perception and fMRI activation levels of voxels within object-related visual areas.

The idea of category-specific representations is usually associated with the hypothesis that different object categories occupy distinct anatomical regions (Ishai et al., 1997; Martin et al., 1996). The experimental approach based on this idea has several drawbacks. First, its results depend on the choice of the visual stimuli; different stimuli may yield somewhat different anatomical divisions. Second, given the spatial resolution of fMRI, which is of the order of 1–2 mm, it is difficult to use fMRI to find the optimal stimulus for a given voxel, as was done by electrophysiological means for clusters of IT neurons in the monkey (Fujita et al., 1992). Therefore, a voxel that ap-

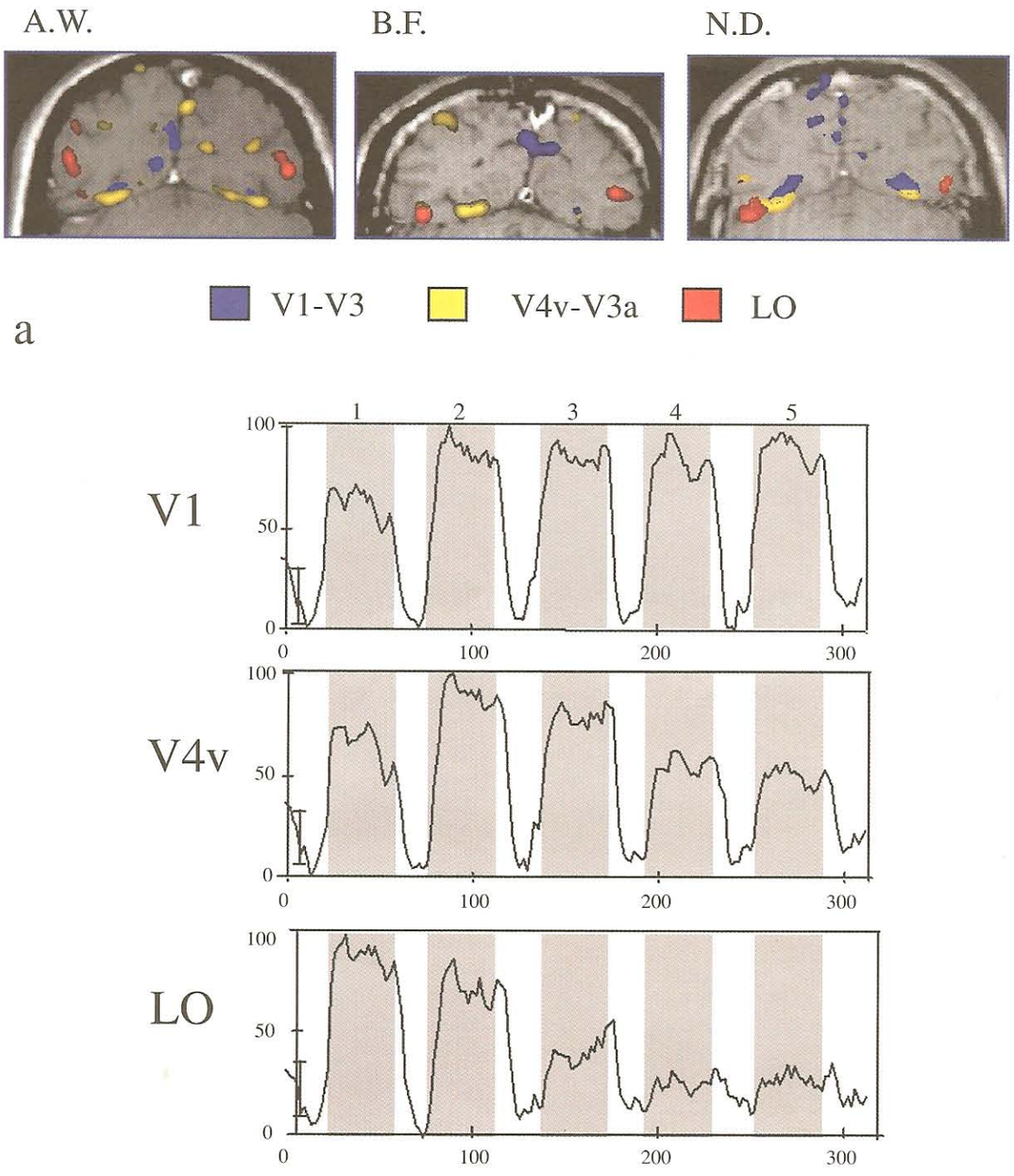

Figure 3. Results of the scrambling experiment. Panel a: Superposition of the activation maps of the most significant voxels of the three functional foci, obtained by regression analysis of the scrambling experiment overlaid on T1 weighted high-resolution anatomical scans of 3 different subjects. The lightness of each color corresponds to the level of the correlation ρ between the time course of the voxel and an ideal paradigm (statistical significance $P < 1\epsilon - 6$, $\rho > 0.4$); voxels below the threshold were not colored: blue, V1–V3; yellow, V4v, V3a; red, lateral occipital (LO) cortex. Note that the anatomical organization is such that low-level visual areas V1–V3 are located in the medial portion of the slice, whereas higher level areas are located more laterally, with area LO being the most lateral. Panel b: Average time courses of 9 subjects derived from each of the three foci of activation depicted in panel a. The abscissa denotes time in seconds and the ordinate normalized fMRI signal strength. Error bars indicate ±1 averaged standard error of the mean. The numbers on top correspond to the epochs shown in Figure 2. Note the increased sensitivity to image scrambling in LO, as compared with V1–V3 and V4v.

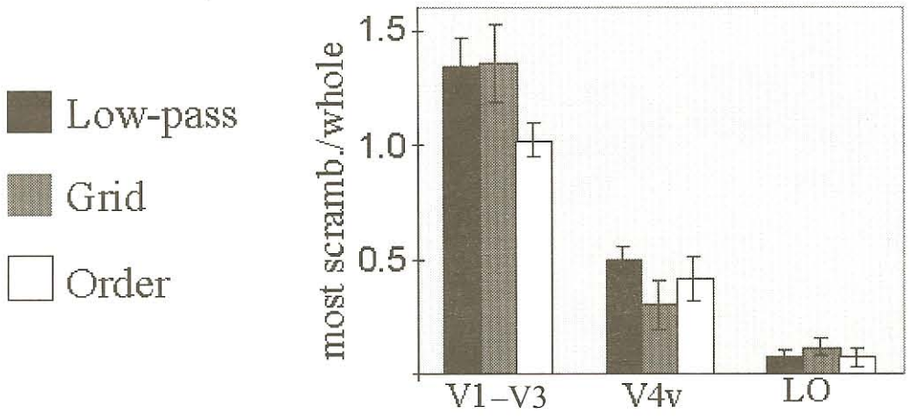

Figure 4. The histogram of the scrambling ratio is plotted for areas V1–V3, V4v, and LO, for different variations of the scrambling experiment. We defined the scrambling ratio as (average fMRI signal during most scrambled images epoch − blank)/(average fMRI signal during unscrambled images epoch − blank). The abscissa denotes the area's label, and the ordinate the scrambling index. Black, low-passed filtered pictures; gray, whole pictures including a grid; white, same as gray but order of images permuted. Note that the variability of the empirical value of the scrambling index in the same area in different scrambling experiments is much smaller, as compared with the differences between the scrambling indices of the functionally different areas.

pears, for example, to belong to a "face" or a "house" area may in fact respond better to images of objects other than faces or houses.

One way to circumvent these problems is to examine the entire distributed pattern of activation of voxels in object-related areas and to attempt to relate it to the category of the stimulus. Because the fMRI signal measured in a voxel is determined by the activity of all the neurons within it throughout the duration of the scan, different stimuli will produce different levels of the signal, depending on the correspondence between the stimulus and the selectivity profile of the neurons within the voxel. This approach has two advantages: It does not depend on finding the optimal stimulus for any of the voxels, and it is based on a computational theory (see above) that can be tested quantitatively.

The Experimental Setup

Seven subjects participated in the *shape space* experiment, in which we examined the distributed patterns of activation of visual object-related brain areas. The stimuli were 32 images of three-dimensional (3-D) computer-generated objects from five categories: planes (4), fish (3), standing human and primate figures (4), four-legged animals (12), and cars (9), as is illustrated in Figure 5. The objects were taken from a commercial graphics library (Viewpoint Datalabs, Inc.) and were rendered with the Silicon Graphics Inventor software library in 256 gray levels, at a rate of 0.5 Hz, which was synchronized with the scanning rate. The objects were rendered in such a manner that the size and the illumination of all the stimuli images were similar. Objects from the same category were presented in the same pose and occupied the central part of the image. Control epochs consisted of 3-D noise patterns created by "exploding" the images into random triangle fragments; blank epochs were used to identify visual areas. The subjects were asked to covertly name the visual stimuli, including the random triangle textures. After the fMRI scan, the subjects participated in a psychophysical experiment, described below, that involved the same images as those presented in the scan. The aim of this psychophysical test was to map the perceived shape space of the subjects (Cutzu & Edelman, 1996; Shepard & Cermak, 1973).

Experimental Results

Psychophysics. The perceptual version of the shape representation space was derived by applying multidimensional scaling (MDS) analysis (Shepard, 1980) to the dissimilarity judgments among object shapes made by the subjects. Because the number of pairwise comparisons required to fill the entire 32×32 dissimilarity matrix is prohibitively large, a tree construction method (Fillenbaum & Rapoport, 1979) was used to obtain dissimilarity data for the 32 stimuli objects. The set of 32 images was randomly arranged and shown to the subjects. The subjects were asked to sort the images according to their similarity and were instructed to base their judgments on the shape of the objects. The subjects began by selecting the 2 most similar objects, then the next 2 most similar objects, and so on. The tree construction method requires that the subject specify directly only $(N - 1)$ out of the $N(N - 1) / 2$ possible proximities. All other dissimilarity scores were derived from the tree of pairwise comparisons given by the subject, using Dijkstra's algorithm for determining the shortest path connecting two vertices in a graph.

Figure 6, left, illustrates the shape space configuration derived from the dissimilarity data pooled from all the subjects. In this figure, points corresponding to the indi-

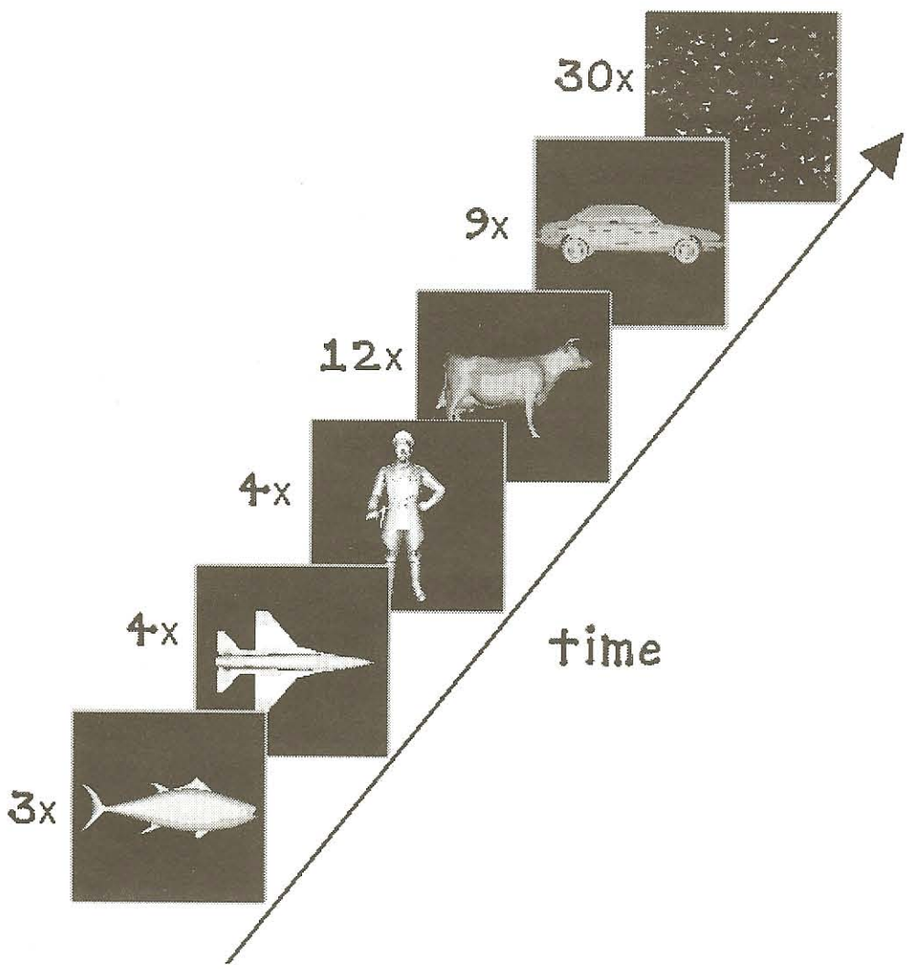

Figure 5. An illustration of the visual stimuli and the sequence of the shape space experiment. The experiment begins and ends with a blank epoch lasting 30 sec. Following the first blank epoch, 32 images of objects appeared from each of the categories displayed, such that they were synchronized with the scanning rate of 0.5 Hz. This epoch was succeeded by an epoch containing noise patterns of random three-dimensional triangle fragments that were displayed at the same rate for 60 sec.

vidual objects are clustered into five categories: cars, four-legged animals, upright figures, fish, and airplanes. Note that the exact location of a point within the cluster is of little importance: It does not really matter whether the pig is mapped to the left or the right of the cow in this shape space. In comparison, it is important that the distances between points (exemplars) belonging to the same category are smaller on the average than the distances between exemplars of different categories; it is this difference that allows the subjects to perceive the categories as distinct.

Voxel-space representation. The fMRI data were first preprocessed, using principal component analysis (Reyment & Jöreskog, 1993), to reduce correlated noise artifacts. This was followed by a Kolmogorov–Smirnov (KS) statistical test (Baker, Hopper, & Stern, 1993; Siegel, 1956) that detected voxels that were activated significantly by images of objects, as compared with noise patterns formed by random triangle fragments. This test highlighted bilaterally voxels located in the LO region. Only the most significant voxels in each slice were submitted to further analysis, as follows.

For each image of an object, the vector of the distributed activation of voxels in object-related areas was created by concatenating the activation of the significant object–voxels at the time point at which this particular image was presented. We call this the *voxel-space* representation of the stimulus. The dimensionality of this representation depends on the number of significant voxels whose activation exceeds the threshold in each subject (mean, 105 ± 44 voxels). To visualize the voxel-space rep-

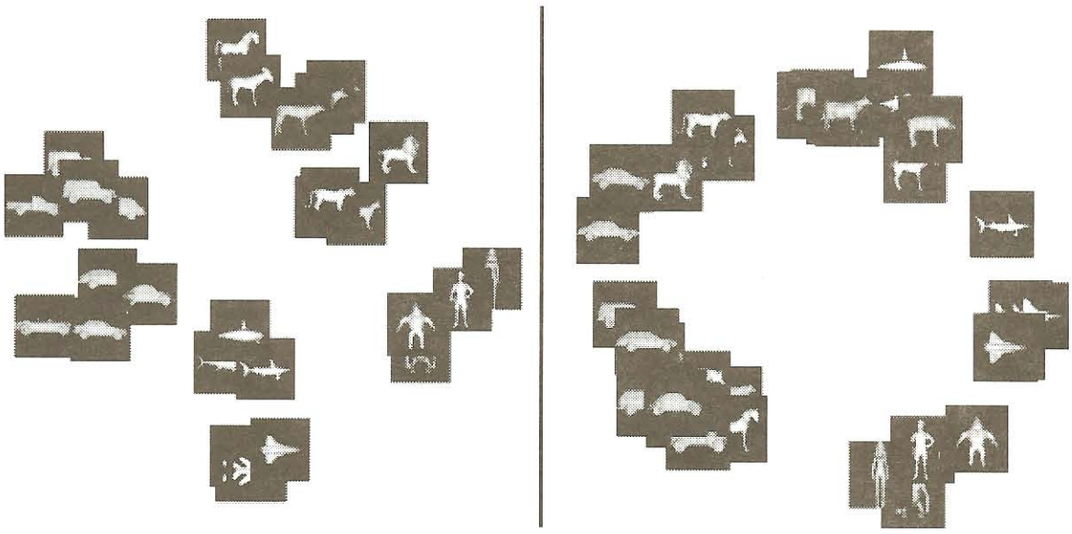

Figure 6. A comparison between the layout of the object representation space derived by multidimensional scaling (MDS; see Shepard, 1980) from perceptual judgment of similarities among objects (left) and that from the similarities among activation patterns measured by fMRI (right). Left: A two-dimensional MDS configuration of the 32 objects, recovered from the *psychophysically determined* dissimilarity matrix combined from all subjects. Right: A two-dimensional MDS configuration of the 32 objects, recovered from the combined *voxel-space representation* derived from the most significant object-related voxels in all 7 subjects. Although the MDS configuration derived from the voxel-space representation is noisier than the configuration retrieved from the psychophysical test, there is some clustering according to object categories. Note that the MDS configuration space does not necessarily correspond in a simple fashion to the physical (anatomical) space in the cortex.

resentation, we used multidimensional scaling to embed it into two dimensions. The multidimensional scaling analysis (SAS procedure MDS; SAS Institute, Inc., 1989) was applied to the Euclidean distance matrix of all pairs of voxel-space representations corresponding to the activations by different images of objects. This procedure was carried out for each subject separately (Figure 7), as well as for all subjects (Figure 6, right).

A comparison between the configuration obtained from the pooled fMRI data and the configuration derived from psychophysics (Figure 6) reveals important similarities. Specifically, some objects, such as airplanes and upright figures, cluster together according to the category. Even in a configuration obtained from the fMRI activation of a single subject (Figure 7), it is possible to distinguish some clustering of object categories—for example, airplanes, four-legged animals, and cars.

To quantify the ability of the distributed voxel-space object representation to support categorization, voxel-space similarity data from each of the 7 subjects were submitted to a standard hierarchical clustering procedure (SAS procedure CLUSTER; SAS Institute, Inc., 1989). We compared the categorization results obtained by this procedure with the object categories that had been characterized psychophysically—namely, four-legged animals, cars, planes, fish, and figures. The mean classification error rate based on voxel-space representation was 0.28 ± 0.05 (*SD*). The significance of this figure was confirmed by a bootstrap procedure (Efron & Tibshirani, 1993) in which clustering was run on randomized data obtained by permuting the 32 slots in each time course record (each voxel was separately permuted). This procedure, which had been used before in evaluating the statistical significance of MDS results (Cutzu & Edelman, 1996; Edelman, 1995), diminished the temporal correlations between the voxels but preserved the basic statistical properties of the data. The classification error rate obtained with the randomized data was 0.55 ± 0.10 (std)—significantly above the error obtained with the actual subject data.

Coarse Coding Analysis of the Scrambling Experiment Data

A direct prediction of the shape space hypothesis is that the same clustering by object category should be revealed if MDS is applied to voxel data from any experiment in which subjects are exposed to a variety of objects. To test this, we returned to the scrambling experiment and carried out the same analysis as above. We used data from the most significant voxels that responded to entire objects and in which the activation was reduced by any scrambling. Data were combined from 7 subjects (see Figure 3). Only the epoch in which entire images of faces or animals had been presented was analyzed. Because this experiment was not originally designed to map the shape representation space of the subjects, stimuli were not precisely ordered by category, and their appearance was not controlled for size, illumination, or viewing position. The results (Figure 8, left) reveal that points corre-

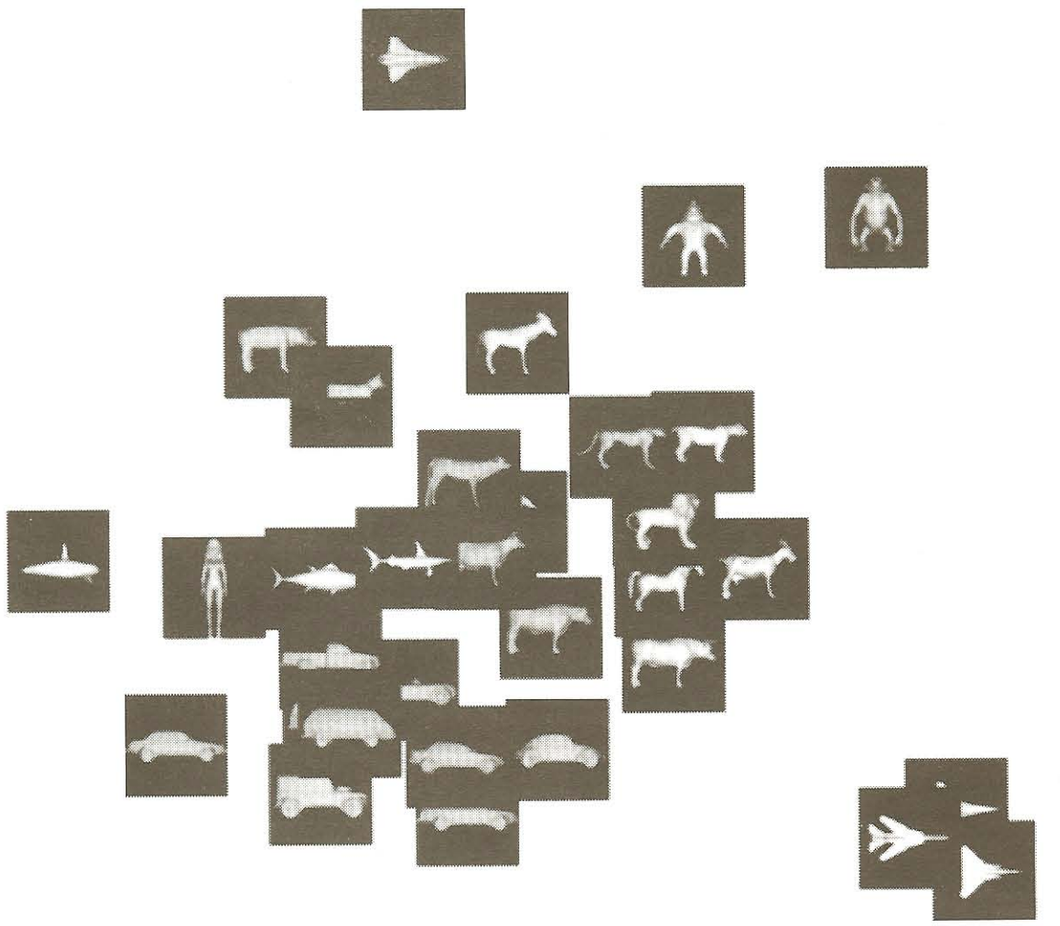

Figure 7. This two-dimensional multidimensional scaling configuration of the 32 objects was recovered from the activation of the 136 most significant object-sensitive voxels in five fMRI slices of a *single subject* (see text for details). Note how airplanes, cars, and four-legged animals are clustered separately.

sponding to animal images and faces form quite separable clusters. The significance of this result was estimated by a bootstrap procedure (Efron & Tibshirani, 1993). To that effect, the MDS analysis was applied to random permutations of the 19 time slots corresponding to the presentations of entire images in that experiment, separately for each voxel. The resulting configuration (Figure 8, right) shows no clustering of either animals or faces, indicating that the clusters derived from the original data are most probably not due to chance or to some bias in the data.

DISCUSSION

The fMRI results, such as those shown in Figures 6 and 7, show that the shape space clustering is correlated with the perceived object categories. This finding can be given a straightforward interpretation in terms of the computational theory outlined above and thereby linked to a much wider corpus of data provided by computer simulation, psychophysics, and electrophysiological studies. According to our theory, objects are represented by their similarity to a number of reference shape classes. Given that each such class is assigned a cortical module, as in the implementation described in Edelman and Duvdevani-Bar (1997a), the similarities between the current stimulus and the reference shapes should be manifest in the relative activity levels of the various modules—precisely what our fMRI experiments measured and what the MDS-based analysis technique helped us to visualize. In this sense, one may say that the cortical activity patterns underlying the configurations revealed by our analysis of the fMRI data *are* the representational substrate of the perceptual shape categories.

For the conclusions suggested above to be reasonably grounded in the empirical findings, additional controls are required. First, the experiments conducted so far do not exclude the possibility that the clustering we found was due to some *low-level* features of the stimulus images, rather than to the shape categories to which the depicted objects belonged. For example, images of upright figures

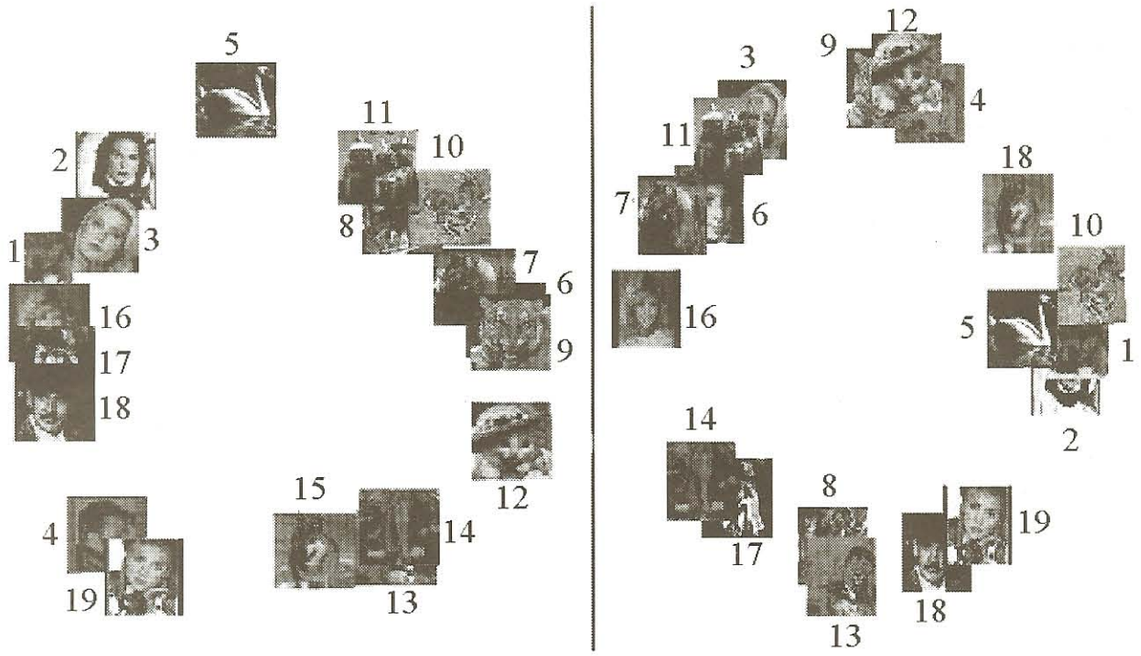

Figure 8. Shape space analysis of the scrambling experiment data. Left: Multidimensional scaling (MDS) was applied to the most significant voxels that responded preferentially to entire objects but not to scrambled versions of these objects. Data were taken from 7 subjects. Each image is labeled by its serial number in the epoch. Note that animals and faces form separate (in fact, nearly linearly separable) clusters. Right: The significance of the results depicted on the left was assessed by a bootstrap procedure, in which the MDS analysis was applied to randomly permuted time courses. Note that this plot reveals no clustering, indicating that the clusters on the left are statistically unlikely to be due to chance or to a data artifact.

in our stimulus set are, on the average, more similar to one another (mean Euclidean distance in the image pixel space of 27,600) than to cars (42,600), animals (42,000), fish (42,400), or planes (42,600). To control for this factor, the shape space experiment is currently being repeated with more than one image per object, the images having been taken over a wide range of orientations. For these stimuli, the pixel-space distances between images of the same object are not less than those between images of different objects presented in similar views, making the low-level feature explanation less likely.[5]

Second, part of the tendency of similar objects to cluster in the MDS rendition of the voxel activation pattern in the shape space experiment may stem from their order of presentation.[6] It has been shown that the presentation of a stimulus elicits a prolonged fMRI response in V1 voxels, which can be "smeared" over several seconds (Boynton, Engel, Glover, & Heeger, 1996). It is possible, therefore, that a signal measured at a specific time may be influenced by several preceding visual stimuli. We plan to control for this factor by revising the experimental procedure in such a manner that the images of various categories will be presented both in a random order and grouped by category in different phases of the experiment. Order effects can also be reduced by presenting the images in random order but in isolation, with blanks preceding and following each stimulus (Buckner et al., 1997).

The significance of the visualization of the shape space made possible by our method—provided that the results reported above withstand further empirical tests—is twofold. On the one hand, this method offers a peek into the internal representation of object shapes, including the philosophically exciting eventual possibility of "guessing" what the subject is looking at while his or her brain is being scanned (cf. Albright, 1991). On the other hand, the very amenability of cortical activation patterns to the kind of analysis we used here supports a particular variety of computational theories of object representation (Edelman, 1998; Edelman & Duvdevani-Bar, 1997b; Poggio, 1990) and offers an integrated theoretical interpretation for a range of empirical findings in monkeys (Logothetis et al., 1995; Sugihara, Edelman, & Tanaka, 1998; Tanaka, 1992, 1996), humans (Bülthoff et al., 1995; Cutzu & Edelman, 1996; Shepard & Chipman, 1970), and computers (Edelman & Duvdevani-Bar, 1997a).

REFERENCES

Albright, T. D. (1991). Motion perception and the mind–body problem. *Current Biology*, **1**, 391-393.

Allison, T., Ginter, H., McCarthy, G., Nobre, A., & Puce, A. (1994). Face recognition in human extrastriate cortex. *Journal of Neurophysiology*, **71**, 821-825.

Allison, T., McCarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers and colors. *Cerebral Cortex*, **5**, 544-554.

BAKER, J. R., HOPPEL, B. E., & STERN, C. E.(1993). Dynamic functional imaging of the complete human cortex using gradient-echo and asymmetric spin-echo echo planar magnetic resonance imaging. *Society for Magnetic Resonance in Medicine Abstracts*, **12**, 1400.

BARLOW, R. B. (1990). What the brain tells the eye. *Scientific American*, **262**, 66-70.

BIEDERMAN, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, **94**, 115-147.

BOYNTON, G. A., ENGEL, S. A., GLOVER, G. H., & HEEGER, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, **16**, 4207-4221.

BROWN, J. M., WEISSTEIN, N., & MAY, J. G. (1992). Visual search for simple volumetric shapes. *Perception & Psychophysics*, **51**, 40-48.

BUCKNER, R. L., BANDETTINI, P. A., O'CRAVEN, K., SAVOY, R. L., PETERSON, S., RAICHLE, M., & ROSEN, B. (1997). Detection of cortical activation during average single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, **93**, 14878-14883.

BÜLTHOFF, H. H., EDELMAN, S., & TARR, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, **5**, 247-260.

CHELAZZI, L., MILLER, E., & DESIMONE, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, **363**, 345-347.

CUTZU, F., & EDELMAN, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences*, **93**, 12046-12050.

CUTZU, F., & EDELMAN, S. (1998). Representation of object similarity in human vision: Psychophysics and a computational model. *Vision Research*, **38**, 2227-2257.

DEYOE, E. A., CARMAN, G. J., BANDETTINI, P., GLICKMAN, S., WIESER, J. R., COX, D. M., & NEITZ, J. (1996). Mapping striate and extrastriate visual areas in human cerebral cortex. *Proceedings of the National Academy of Sciences*, **93**, 2382-2386.

DUVDEVANI-BAR, S., EDELMAN, S., HOWELL, A. J., & BUXTON, H. (1998). A similarity-based method for the generalization of face recognition over pose and expression. In S. Akamatsu & K. Mase (Eds.), *Proceedings of the 3rd International Symposium on Face and Gesture Recognition (FG98)* (pp. 118-123). Washington, DC: IEEE Press.

EDELMAN, S. (1995). Representation of similarity in 3D object discrimination. *Neural Computation*, **7**, 407-422.

EDELMAN, S. (1997). Computational theories of object recognition. *Trends in Cognitive Science*, **1**, 296-304.

EDELMAN, S. (1998). Representation is representation of similarity. *Behavioral & Brain Sciences*, **21**, 449-498.

EDELMAN, S., & DUVDEVANI-BAR, S. (1997a). A model of visual recognition and categorization. *Philosophical Transactions of the Royal Society of London: Series B*, **352**, 1191-1202.

EDELMAN, S., & DUVDEVANI-BAR, S. (1997b). Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, **9**, 701-720.

EDELMAN, S., & WEINSHALL, D. (in press). Computational approaches to shape constancy. In V. Walsh & J. Kulikowski (Eds.), *Perceptual constancies: Why things look as they do*. Cambridge: Cambridge University Press.

EFRON, B., & TIBSHIRANI, R. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

FARAH, M. J. (1990). *Visual agnosia*. Cambridge, MA: MIT Press.

FILLENBAUM, S., & RAPOPORT, A. (1979). *Structures in the subjective lexicon*. New York: Academic Press.

FUJITA, I., TANAKA, K., ITO, M., & CHENG, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, **360**, 343-346.

GRILL-SPECTOR, K., KUSHNIR, T., HENDLER, T., EDELMAN, S., ITZCHAK, Y., & MALACH, R. (1998). A sequence of early object processing stages revealed by fMRI in human occipital lobe. *Human Brain Mapping*, **6**, 316-328.

HAXBY, J. V., HORWITZ, B., UNGERLEIDER, L. G., MAISOG, J. M., PIETRINI, P., & GRADY, C. L. (1994). The functional organization of human extrastriate cortex: A PET-rCBF study of selective attention to faces and locations. *Journal of Neuroscience*, **14**, 6336-6353.

HINTON, G. E. (1984). *Distributed representations* (Tech. Rep. No. CMU-CS 84-157). Carnegie Mellon University, Department of Computer Science.

HUMMEL, J. E., & BIEDERMAN, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, **99**, 480-517.

ISHAI, A., UNGERLEIDER, L. G., MARTIN, A., MAISOG, J., & HAXBY, J. (1997). fMRI reveals differential activation in the ventral object recognition pathway during the perception of faces, houses and chairs. *NeuroImage*, **5**, S149.

KANWISHER, N., CHUN, M. M., MCDERMOTT, J., & LEDDEN, P. J. (1996). Functional imaging of human visual recognition. *Cognitive Brain Research*, **5**, 55-67.

KANWISHER, N., MCDERMOTT, J., & CHUN, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, **17**, 4302-4311.

KANWISHER, N., WOODS, R. P., IACOBONI, M., & MAZZIOTA, J. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience*, **9**, 133-142.

KEELE, S. W., & NEILL, W. T. (1978). Mechanisms of attention. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 9, pp. 3-47). New York: Academic Press.

KOBATAKE, E., & TANAKA, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, **71**, 856-867.

KRUSKAL, J. B. (1964). Non-metric multidimensional scaling: A numerical method. *Psychometrika*, **29**, 115-129.

KRUSKAL, J. B., & WISH, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

LOGOTHETIS, N. K., PAULS, J., & POGGIO, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, **5**, 552-563.

LOGOTHETIS, N. K., PAULS, J., BÜLTHOFF, H. H., & POGGIO, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, **4**, 404-414.

LOGOTHETIS, N. K., & SCHEINBERG, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, **19**, 577-621.

MALACH, R., REPPAS, J. B., BENSON, R. R., KWONG, K. K., JIANG, J., KENNEDY, W. A., LEDDEN, P. J., BRADY, T. J., ROSEN, B. R., & TOOTELL, R. B. H. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, **92**, 8135-8139.

MARTIN, A., WIGGS, C. L., UNGERLEIDER, L. G., & HAXBY, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, **379**, 649-652.

MILLER, E. K., LI, L., & DESIMONE, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *Journal of Neuroscience*, **13**, 1460-1478.

MORAN, J., & DESIMONE, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, **229**, 782-784.

NISSEN, M. J. (1985). Accessing features and objects: Is location special? In M. I. Posner & O. S. M. Marin (Eds.), *Attention and performance XI* (pp. 205-219). Hillsdale, NJ: Erlbaum.

O'TOOLE, A. J., EDELMAN, S., & BÜLTHOFF, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, **38**, 2351-2363.

POGGIO, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, **LV**, 899-910.

POGGIO, T., & EDELMAN, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.

PUCE, A., ALLISON, T., ASGARI, M., GORE, J. C., & MCCARTHY, G. (1996). Differential sensitivity of human visual cortex to faces, letterstrings and textures: A functional magnetic resonance imaging study. *Journal of Neuroscience*, **16**, 5205-5215.

REYMENT, R., & JÖRESKOG, K. (1993). *Applied factor analysis in the natural sciences*. Cambridge: Cambridge University Press.

SAS INSTITUTE, INC. (1989). *SAS/STAT User's Guide, Version 6*. Cary, NC: Author.

SERENO, M. I., DALE, A. M., REPPAS, J. B., KWONG, K. K., BELLIVEAU, J. W., BRADY, T. J., ROSEN, R. B., & TOOTELL, R. B. H. (1995, May 12). Borders of multiple visual areas revealed by functional magnetic resonance. *Science*, **268**, 889-893.

SHEPARD, R. N. (1962). The analysis of proximities: Multidimensional scaling with unknown distance function: Part I. *Psychometrika*, **27**, 125-140.

SHEPARD, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, **210**, 390-397.

SHEPARD, R. N., & CERMAK, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology*, **4**, 351-377.

SHEPARD, R. N., & CHIPMAN, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, **1**, 1-17.

SIEGEL, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

SNIPPE, H. P., & KOENDERINK, J. J. (1992). Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, **66**, 543-551.

SPITZER, H., DESIMONE, R., & MORAN, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science*, **240**, 338-340.

SUGIHARA, T., EDELMAN, S., & TANAKA, K. (1998). Representation of objective similarity among three-dimensional shapes in the monkey. *Biological Cybernetics*, **78**, 1-7.

TANAKA, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, **2**, 502-505.

TANAKA, K. (1993a). Column structure of inferotemporal cortex: "Visual alphabet" or "differential amplifiers"? In *Proceedings of the International Joint Conference on Neural Networks* (pp. 1095-1099). Washington, DC: IEEE Press.

TANAKA, K. (1993b). Neuronal mechanisms of object recognition. *Science*, **262**, 685-688.

TANAKA, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, **19**, 109-139.

TANAKA, K. (1997). Mechanisms of visual object recognition: Monkey and human studies. *Current Opinion in Neurobiology*, **7**, 523-529.

TOMASI, C., & KANADE, T. (1992). Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, **9**, 137-154.

TOOTELL, R. B. H., DALE, A. M., SERENO, M. I., & MALACH, R. (1996). New images from human visual cortex. *Trends in Neurosciences*, **19**, 481-489.

ULLMAN, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, **32**, 193-254.

ULLMAN, S. (1995). Sequence-seeking and counter-streams: A model for information flow in the cortex. *Cerebral Cortex*, **5**, 1-11.

ULLMAN, S. (1996). *High level vision*. Cambridge, MA: MIT Press.

ULLMAN, S., & BASRI, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **13**, 992-1005.

WACHSMUTH, E., ORAM, M. W., & PERRETT, D. I. (1994). Recognition of objects and their component parts: Responses of single units in the temporal cortex of the macaque. *Cerebral Cortex*, **5**, 509-522.

WANG, G., TANAKA, K., & TANIFUJI, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. *Science*, **272**, 1665-1668.

NOTES

1. Although, according to the structural decomposition theory (Biederman, 1987), objects are represented in terms of generic parts and their spatial relationships, the expectation to find regions responsive selectively to such parts is not warranted by this theory or by models derived from it (Hummel & Biederman, 1992). Likewise, although the alignment theory (Ullman, 1989) calls for geometric information about object shapes to be represented, neither this theory nor the "sequence seeking" model derived from it (Ullman, 1995) specifies the form or the qualities of the representation that can be measured by fMRI.

2. In some cases, completely (Tomasi & Kanade, 1992; Ullman & Basri, 1991).

3. Current mathematical models of hyperacuity perception operate on the same principle: Several overlapping receptive fields carry high-precision spatial information, which cannot be recovered from the activity of each receptive field on its own (Snippe & Koenderink, 1992).

4. To relate these foci of activation to established human visual areas, we mapped in 7 subjects, during the same experimental sessions, the cortical representation of the vertical and horizontal meridians of the visual field (DeYoe et al., 1996; Sereno et al., 1995), using either objects or texture stimuli. A comparison of the mapped meridian and the three foci of activation indicated that the medial focus (blue in Figure 3) was confined to areas V1–V3, whereas the more lateral focus (yellow in Figure 3) overlapped ventrally with area V4v. The most lateral focus (red in Figure 3) lacked retinotopy and corresponded anatomically to the LO complex (Grill-Spector et al., 1998; Malach et al., 1995; Tootell, Dale, Sereno, & Malach, 1996).

5. Encouragingly, a recent psychophysical shape space experiment that involved several images per object revealed clustering by object shape, and not by viewpoint (Cutzu & Edelman, 1998).

6. This does not apply, however, to the reanalysis of the scrambled experiment data by MDS.

(Manuscript received December 4, 1997;
revision accepted for publication August 19, 1998.)