PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

A model of visual recognition and categorization

Shimon Edelman and Sharon Duvdevani-Bar

Phil. Trans. R. Soc. Lond. B 1997 **352**, 1191-1202
doi: 10.1098/rstb.1997.0102

A model of visual recognition and categorization

SHIMON EDELMAN[1] AND SHARON DUVDEVANI-BAR[2]

[1]*Center for Biological and Computational Learning, Massachusetts Institute of Technology E25-201, Cambridge, MA 02142, USA*
(edelman@ai.mit.edu)
[2]*Department of Applied Math and Computer Studies, The Weizmann Institute of Science, Rehovot 76100, Israel*
(sharon@wisdom.weizmann.ac.il)

SUMMARY

To recognize a previously seen object, the visual system must overcome the variability in the object's appearance caused by factors such as illumination and pose. Developments in computer vision suggest that it may be possible to counter the influence of these factors, by learning to interpolate between stored views of the target object, taken under representative combinations of viewing conditions. Daily life situations, however, typically require categorization, rather than recognition, of objects. Due to the open-ended character of both natural and artificial categories, categorization cannot rely on interpolation between stored examples. Nonetheless, knowledge of several representative members, or prototypes, of each of the categories of interest can still provide the necessary computational substrate for the categorization of new instances. The resulting representational scheme based on similarities to prototypes appears to be computationally viable, and is readily mapped onto the mechanisms of biological vision revealed by recent psychophysical and physiological studies.

1. INTRODUCTION

To perceive shapes as instances of object categories that persist over time, a visual system—biological or artificial—must combine the capacity for internal representation and for the storage of object traces with the ability to compare these against the incoming visual stimuli, namely, images of objects. The appearance of the latter is determined by (i) the shape and the surface properties intrinsic to the object, (ii) its disposition with respect to the observer and the illumination sources, (iii) the optical properties of the intervening medium and the imaging system, and (iv) the presence and location of other objects in the scene (Ullman 1996). Thus, to detect that two images, which may be taken seconds or years apart, belong, in fact, to the same three-dimensional object, the visual system must overcome the influence of a number of extrinsic factors that affect the way objects look.

Possible approaches to separating information on the intrinsic shape of an object from the extrinsic factors affecting its appearance depend on the nature of the task faced by the system. One of these tasks, *recognition* (knowing a previously seen object as such), appears now to require little more than storing information concerning earlier encounters with the object, as suggested by the success of view-based recognition algorithms recently developed in computer vision (Ullman 1996). As we shall see, it is surprisingly easy to extend such a memory-based strategy to deal with *categorization*, a task that requires the system to make sense of novel shapes. Thus, familiarity with a relatively small selection of objects can be used as a foundation for processing (i.e., representing and categorizing) other objects, never seen before.

The theory of representation outlined in the present paper is based on the idea of describing objects in terms of their similarities to a relatively small number of reference shapes (Edelman 1995*c*; Edelman 1997*b*). The theoretical underpinnings of this approach are discussed elsewhere (Edelman & Duvdevani-Bar 1997); here, we demonstrate its viability on a variety of objects and object classes, and discuss the implications of its successful implementation for understanding object representation and categorization in biological vision.

(*a*) *Visual recognition*

If the appearance of visual objects were immutable and unaffected by any extrinsic factors, recognition would amount to simple comparison by template matching, a technique in which two patterns are regarded as the same if they can be brought into perfect register. As things stand, the effects of the extrinsic factors must be mitigated to ensure that the comparison is valid. Theories of recognition, therefore, tend to have two parts: one concentrating on the form of the internal representation into which images of objects are cast, and the other on the details of the comparison process.

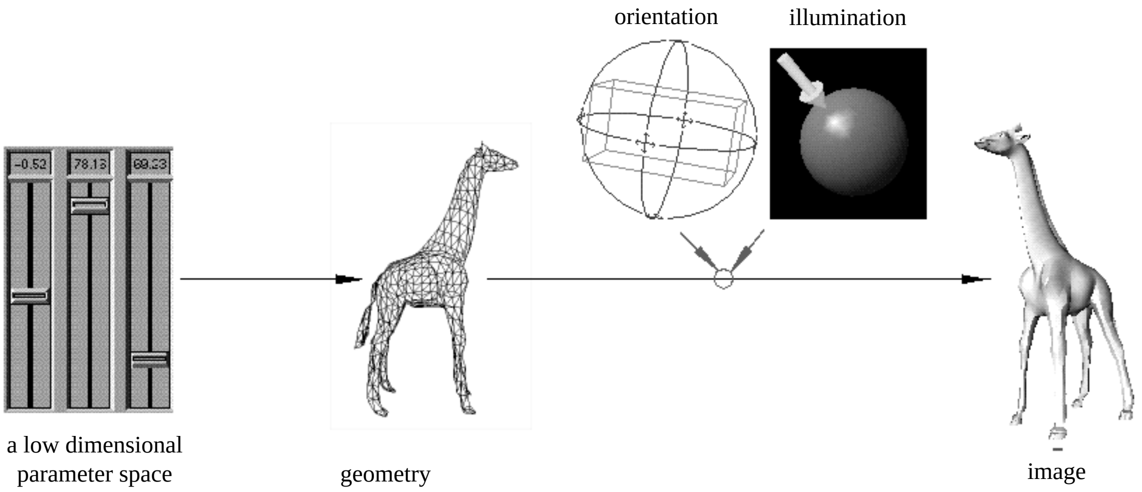

Figure 1. The process of image formation. A family of shapes (e.g. four-legged animal-like objects) can be defined parametrically, using a small number of variables (Edelman & Duvdevani-Bar 1997), illustrated symbolically on the left by the three 'sliders' that control the values of the shape variables. These, in turn, determine the geometry of the object, e.g. the locations of the vertices of a triangular mesh that approximates the object's shape. Finally, intrinsic and extrinsic factors (geometry and viewing conditions) together determine the appearance of the object.

A theory of recognition that is particularly well-suited to the constraints imposed by a biological implementation has been described in Poggio & Edelman (1990). This theory relies on the observation that the views of a rigid object undergoing transformation such as rotation in depth reside in a smooth low-dimensional manifold (a 'surface') embedded in the space of coordinates of points attached to the object (Ullman & Basri 1991; Jacobs 1996). One may observe, further, that the properties of smoothness and low dimensionality of this manifold, which may be called the *view space* of the object, are likely to be preserved in whatever measurement space is used by the front-end of the visual system. The operational consequence of this observation is that a new view of an object may be recognized by interpolation among its selected stored views, which together represent the object. A criterion that indicates the quality of the interpolation can be formed by comparing the stimulus view to the stored views, by passing the ensuing proximity values through a Gaussian nonlinearity, and by computing a weighted sum of the results (this amounts to a basis-function interpolation of the view manifold, as described in § 3 *a*). The outcome of this computation is an estimate of the measurement-space distance between the point that encodes the stimulus and the view manifold. If a sufficient number of views is available to define that manifold, this distance can be made arbitrarily independent of the pose of the object, one of the extrinsic factors that affect the appearance of object views. The influence of the other extrinsic factors (e.g. illumination) can be minimized in a similar manner, by storing examples that span the additional dimensions of the view manifold, corresponding to the additional degrees of freedom of the process of image formation.

In the recognition scenario, the tacit assumption is that the stimulus image is either totally unfamiliar, or, in fact, corresponds to one of the objects known to the system. A sensible generic decision strategy under this assumption is *nearest-neighbour* (Cover & Hart 1967), which assigns to the stimulus the label of the object that matches it optimally (modulo the influence of the extrinsic factors, and, possibly, measurement noise). In the view-interpolation scheme, the decision can be based on the value of the distance-to-the-manifold criterion that reflects the quality of the interpolation (a low value signifies an unfamiliar object). As we argue next, this approach, being an instance of the generic nearest-neighbour strategy, addresses only a small part of the problem of visual object processing.

(*b*) *Visual categorization*

Bound by the assumption that variability in object appearance is mainly due to factors such as illumination and pose, the standard approach to recognition calls for a comparison between the intrinsic shape of the viewed object (separated from the influence of the extrinsic factors) and the stored representation of that shape. According to this view, a good representation is one that makes explicit the intrinsic shape of an object in great detail and with high fidelity.

A reflection on the nature of everyday recognition tasks prompts one to question the validity of this view of representation. In a normal *categorization* situation (Rosch 1978; Smith 1990), human observers are expected to *ignore* many of the shape details (Price & Humphreys 1989). Barring special (albeit behaviourally important) cases such as face recognition, entry-level (Jolicoeur *et al.* 1984) names of objects (that is, names spontaneously produced by observers) correspond to categories rather than to individuals, and it is the category of the object that the visual system is required to determine. Thus, the observer is confronted with potential variation in the intrinsic shape of an object, because objects called by the same name do not, generally, have exactly the same shape. This variability in the shape (and not merely in the appearance) of objects must be adequately represented, so that it can be treated properly at the categorization stage.

Different gradations of shape variation call for different kinds of action on the part of the visual system. On the one hand, moderately novel objects can be handled by the same mechanism that processes familiar ones, insofar as such objects constitute variations on familiar themes. Specifically, the nearest-neighbour strategy around which the generic recognition mechanism is built can be allowed to handle shape variation that does not create ambiguous situations in which two categories vie for the ownership of the current stimulus. On the other hand, if the stimulus image belongs to a radically novel object—e.g. one that is nearly equidistant, in the similarity space defined by the representational system, to two or more familiar objects, or very distant from any such object—a nearest neighbour decision no longer makes sense, and should be abandoned in favour of a better procedure.

To provide a basis for a categorization procedure that does not break down when faced with novel shapes, we adopt the concept of a *shape space*—a representational tool that treats all shapes, familiar or novel, equivalently. As we noted above, views of any object span a manifold (a smooth, continuous surface) in the space of measurements carried out by the front end of a visual system. Measuring the proximity of a stimulus (i.e. a point in the measurement space) to such a manifold, corresponding to some known object, yields the similarity between the stimulus and that object.

Similarities measured with respect to several previously seen objects can then be used to categorize shapes, even those seen for the first time. In particular, the currently viewed object can be classified as being the same shape as some previous one, or as being similar in shape to several previously seen objects. The details of this procedure, which is suitable for representing both familiar and novel shapes, are described in the next section.

2. THE SHAPE SPACE

To put familiar and novel shapes on an equal footing, it is useful to describe shapes as points in a common parameter space. A common parameterization is especially straightforward for shapes that are sampled at a preset resolution, then defined by the coordinates of the sample points (cf. figure 1). For instance, a family of shapes each of which is a 'cloud' of k points spans a $3k$-dimensional shape space (Kendall 1984); moving the k points around in 3D (or, equivalently, moving around the single point in the $3k$-dimensional shape space) amounts to changing one shape into another.

By defining similarity between shapes via a distance function in the shape space, clusters of points are made to correspond to classes of shapes (i.e. sets of shapes whose members are more similar to each other than to members of other sets). To categorize a (possibly novel) shape, then, one must first find the corresponding point in the shape space, then determine its location with respect to the familiar shape clusters. Note that while a novel shape may fall in between the clusters, it will in any case possess a well-defined representation. This representation may be then acted upon, e.g. by committing it to memory, or by using it as a seed for establishing a new cluster.

(*a*) *The high-dimensional measurement space*

Obviously, a visual system has no direct access to whatever shape space in which the geometry of distal objects may be defined (in fact, the notion of a unique geometrical shape space does not even make sense: the same physical object can be described quantitatively in many different ways). The useful and intuitive notion of a space in which each point corresponds to some shape can, however, be put to work by introducing an intermediary concept: *measurement space*.

A system that carries out a large number of measurements on a visual stimulus effectively maps that stimulus into a point in a high-dimensional space; the diversity and the large number of independent measurements increase the likelihood that any change in the geometry of the distal objects ends up represented at least in some of the dimensions of the measurement space. Indeed, in primate vision, the dimensionality of the space presented by the eye to the brain is roughly one million, and is determined by the number of fibres in each optic nerve.

Most of this high-dimensional space is empty: a randomly chosen combination of pixel values in an image is extremely unlikely to form a picture of a coherent object. The locus of the measurement-space points that do represent images of coherent objects depends on all the factors that participate in image formation, both intrinsic (the shapes of objects) and extrinsic (e.g. their pose). These points together define the *proximal* (or subjective, as opposed to distal, or objective) shape space. Note that smoothly changing the shape of the imaged object causes the corresponding point to ascribe a manifold in the measurement space. The dimensionality of this manifold depends on the number of degrees of freedom of the shape changes; for example, simple morphing of one shape into another produces a one-dimensional manifold (a curve). Likewise, rotating the object in depth (a transformation with two degrees of freedom) gives rise to a two-dimensional manifold which we call the view space of the object. It turns out that the proximal shape space, produced by the joint effects of deformation and transformation, can be safely considered a locally smooth low-dimensional manifold embedded in the measurement space (Edelman & Duvdevani-Bar 1997).

(*b*) *Dimensionality reduction and the proximal shape space*

In the above formulation, the categorization problem becomes equivalent to determining the location of the measurement-space representation of the stimulus within the proximal shape space. Our approach to this problem is inspired by the observation that the location of a point can be precisely defined by specifying its distance to some prominent reference points, or *landmarks* (Edelman & Duvdevani-Bar 1997). Here distance is meant to capture difference in shape (i.e. the amount of deformation), therefore its estimation must exclude (i) components of measurement-space distance that are orthogonal to the shape space, as well as (ii) components of shape transformation, such as rotation. As we shall see, a convenient computational mechanism for distance estimation that satisfies these two requirements is a module tuned to a particular shape, that is, designed to respond selectively to that shape, irrespective of its transformation. A few such modules, tuned to different reference shapes, effectively reduce the dimensionality of the representation from that of the measurement space to a small number, equal to the number of modules (figure 3). In the next section, we describe a system for shape categorization based on a particular implementation of this approach, which we call the 'chorus of prototypes' (Edelman 1995*b*); its relevance as a model of shape processing in biological vision is discussed in § 5.

3. THE IMPLEMENTED MODEL

A module tuned to a particular shape will fulfil the first of the two requirements stated above—ignoring the irrelevant components of the measurement-space distance—if it is trained to discriminate among objects all of which belong to the desired shape space. Such training imparts to the module the knowledge of the relevant measurement-space directions, by making it concentrate on the features that help discriminate between the objects. To fulfil the second requirement—insensitivity to shape transformations—the module must be trained to respond equally to different views of the object to which it is tuned. A trainable computational mechanism capable of meeting these two requirements is a radial basis function (RBF) interpolation module.

(*a*) *The RBF module*

When stated in terms of an input–output relationship, our goal is to build a module that would output a non-zero constant for all views of a certain target object, and zero for any view of all the other objects in the training set. Because only a few target views are usually available for training, the problem is, in fact, to *interpolate* the view space of the target object, given some examples of its members. With basis function interpolation, this problem is easily solved by a distributed network, whose structure can be learned from

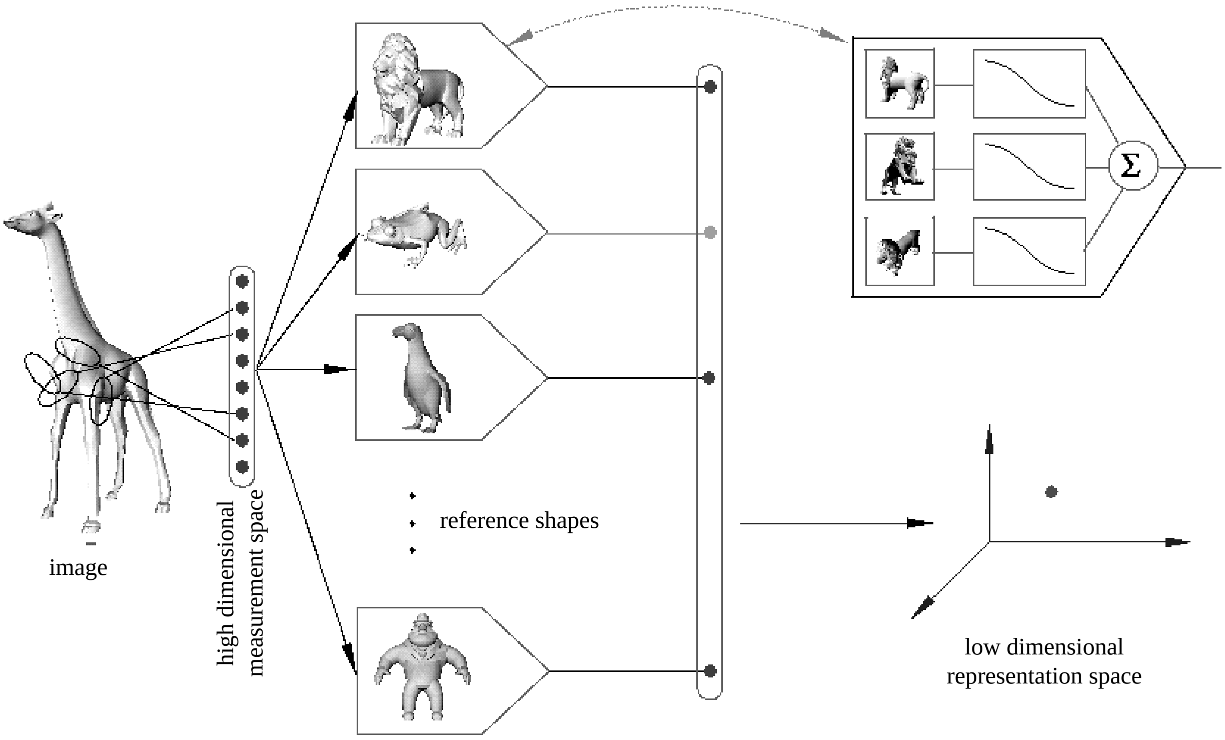

Figure 2. The 'chorus' scheme (§ 3). The stimulus is first projected into a high-dimensional measurement space, spanned by a bank of receptive fields. Second, it is represented by its similarities to reference shapes. In this illustration, only three modules respond significantly, spanning a shape space that is nominally three-dimensional (in the vicinity of the measurement-space locus of giraffe images). The *inset* shows the structure of each module. Each of a small number of training views, $\boldsymbol{v}_{\mathrm{i}}$, serves as the centre of a Gaussian basis function $G(\boldsymbol{a},\boldsymbol{b};\sigma)=\exp(\|\boldsymbol{a}-\boldsymbol{b}\|^2/\sigma^2)$; the response of the module to an input vector $\boldsymbol{x}$ is computed as $y=\Sigma_t w_t G(x;v_t)$. The weights w_t and the spread parameter σ are learned as described in (Poggio & Girosi 1990). It is important to realize that the above approach, which amounts to an interpolation of the view space of the training object using the radial basis function (RBF) method, is not the only one applicable to the present problem. Other approaches, such as interpolation using the multilayer perceptron architecture, may be advantageous, e.g., when the measurement space is 'crowded', as in face discrimination (Edelman & Intrator 1997).

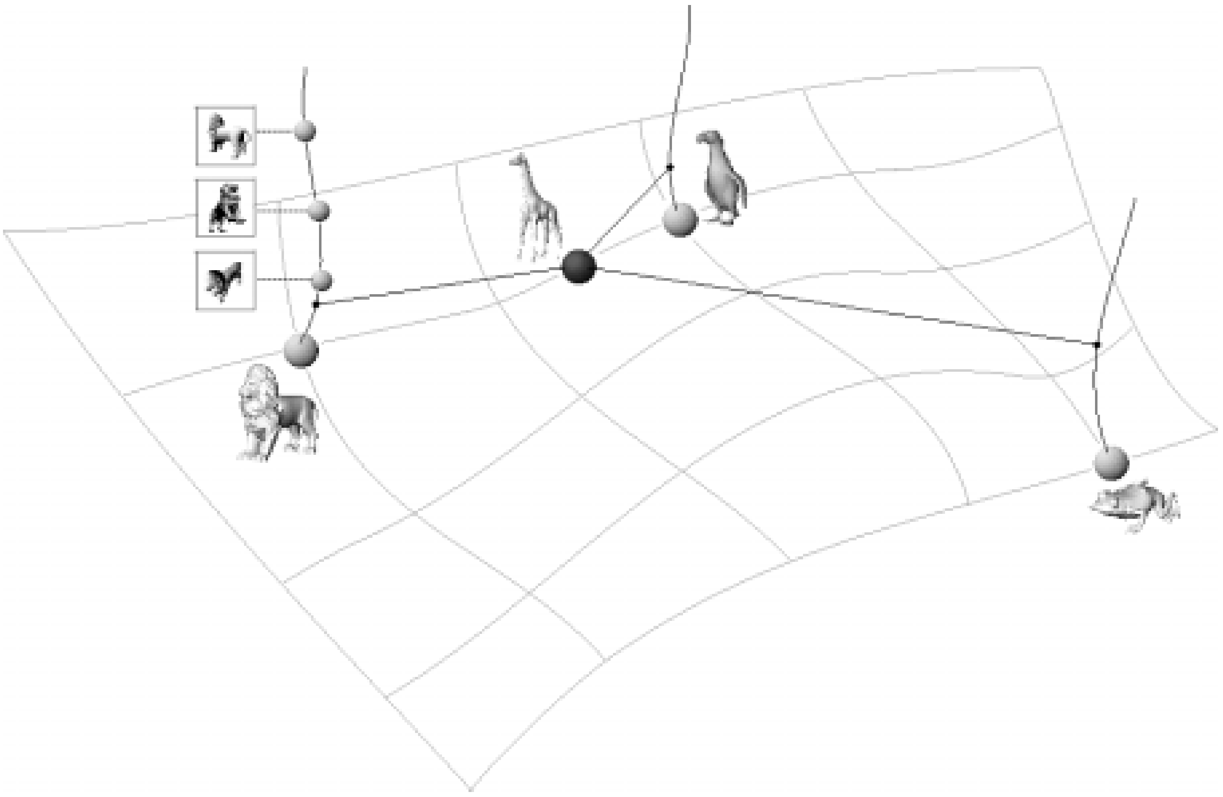

Figure 3. A schematic illustration of the shape-space manifold defined by a chorus of three active modules (lion, penguin, frog). Each of the three reference-shape modules is trained to ignore the viewpoint-related factors (the view space dimension, spanned by views that are shown explicitly for lion), and is thus made to respond to shape-related differences between the stimulus (here, the giraffe}) and its 'preferred' shape. The actual dimensionality of the space spanned by the outputs of the modules (Edelman & Intrator 1997), can be lower than its nominal dimensionality (equal to the number of modules); here the space is shown as a two-dimensional manifold.

examples (Broomhead & Lowe 1988; Poggio & Girosi 1990).

According to this method, the interpolating function is constructed out of a superposition of basis functions, whose shape reflects the prior knowledge concerning the change in the output as one moves away from the data point. In the absence of evidence to the contrary, all directions of movement are considered equivalent, making it reasonable to assume that the basis function is radial (that is, it depends only on the distance between the actual input and the original data point, which serves as its centre). The resulting scheme is known as radial basis function (RBF) interpolation. Once the basis functions have been placed, the output of the interpolation module for any test point is computed by taking a weighted sum of the values of all the basis functions at that point.

An application of RBF interpolation to object recognition has been described in (Poggio & Edelman 1990); the RBF model was subsequently used to replicate a number of central characteristics of the process of recognition in human vision (Bülthoff & Edelman 1992). In its simple version, one basis function is used for (the measurement-space representation of) each familiar view. The appropriate weight for each basis is then computed by an algorithm that involves matrix inversion (a closed-form solution exists for this case). This completes the process of training the RBF network. To determine whether a test view belongs to the object on which the network has been trained, this view (that is, its measurement-space representation) is compared to each of the training views. This step yields a set of distances between the test view and the training views that serve as the centres of the basis functions. In the next step, the values of the basis functions are combined linearly to determine the output of the network (see figure 2).

(*b*) *Training individual modules*

In the computational experiments described below, ten different reference objects were chosen at random from a commercially available database of several hundreds of shapes (see figure 4). To minimize the memory requirements of the scheme, we trained the modules on a few views per object, strategically placed to optimize the coverage of its view space. The choice of the training views was guided by the following three requirements: (i) approximately constant output of the module for different test views of the same object, (ii) tight clustering of the views of each object in the space of the outputs of the modules, and (iii) wide separation between clusters corresponding to the different objects in that space. These three criteria were combined into a single canonical distortion measure (Baxter 1995), used to guide the optimal choice of views in a procedure

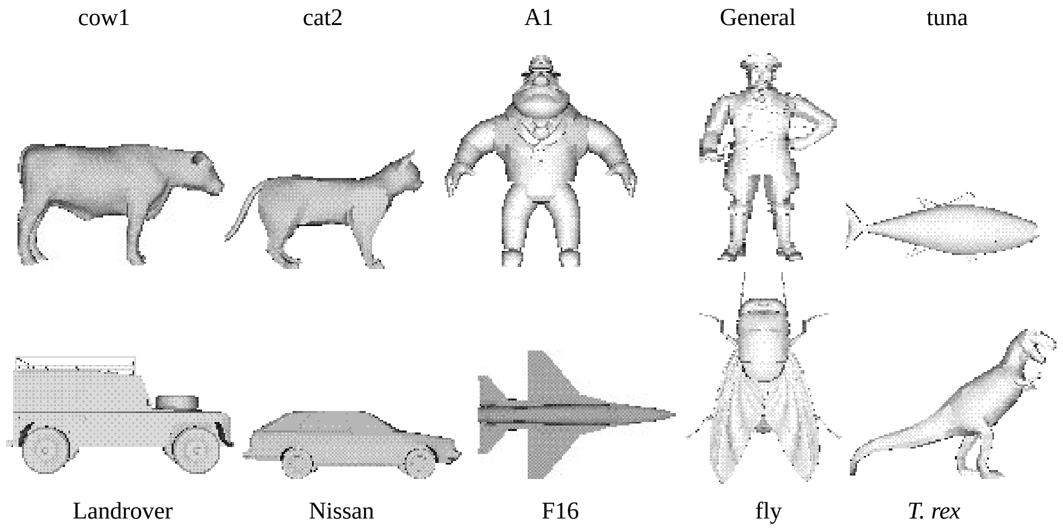

Figure 4. The ten training objects used as reference shapes in the computational experiments described in the text. The objects were chosen at random from a database available from Viewpoint Datalabs, Inc. (http://www.viewpoint.com/).

known as vector quantization (Linde *et al*. 1980). This resulted in a choice of about 15 views per object, which were then used to train the RBF modules, following the standard algorithm described in Poggio & Girosi (1990). The only other free parameter of the RBF scheme, the width of the Gaussian basis functions, was chosen so as to minimize the categorization error of the resulting system over the ten training objects.

4. EXPERIMENTAL RESULTS

We assessed the performance of the 'chorus' scheme in three different tasks: (i) recognition of novel views of the ten objects on which the system had been trained; (ii) categorization of 20 novel objects belonging to categories of which at least one exemplar was available in the training set; and (iii) representation of ten radically novel objects, picked at random from the database.

(*a*) *Recognition*

To test the ability of the system to generalize recognition to novel views of the trained objects, we experimented with three different recognition algorithms. The performance of each algorithm was evaluated on a set of 169 views of each of the trained objects, taken around the canonical orientation (Palmer *et al*. 1981) over a range of $\pm 60^\circ$ in azimuth and elevation, at 10° increments.

First, we computed the performance of each of the ten RBF modules, using individually determined thresholds (set to the mean activity of the module on trained views minus one standard deviation). The generalization error rate (defined as the mean of the miss and the false alarm error rates) for this algorithm was 8%.

We next considered the 'winner-takes-all' (WTA) algorithm, according to which the outcome of the recognition step is the label of the module that gives the strongest response to the current stimulus (in the patterns of module activation shown in table 1, entries for modules that responded, on average, the strongest are marked by bold typeface). The error rate of the WTA method was 12%.

Finally, we trained a second-level RBF module to map the ten-element vectors of the first-level RBF units into ten-element vectors in which the single proper element (signifying the identity of the stimulus) was set to 1, and the others to 0. This approach takes advantage of the observation that important information concerning the shape of the stimulus is contained in the entire pattern of activities that it induces over the reference-object modules (cf. table 1), and not merely in the identity of the strongest-responding module (Edelman *et al*. 1992). Indeed, the WTA algorithm applied to the second-level RBF output resulted in an error rate of 6% (computed over all 169 views of each of the ten objects).

(*b*) *Categorization*

Our second experiment tested the ability of the 'chorus' scheme to categorize 20 'moderately' novel objects, each of which belonged to one of the categories present in the original training set of ten objects. To visualize the utility of representation by similarity to the training objects, we used multidimensional scaling (Shepard 1980) to embed the ten-dimensional layout of points corresponding to various views of test objects into a two-dimensional space (figure 6). An examination of the resulting plot reveals a number of satisfying properties, such as clustering of views by object identity, and grouping of view clusters by similarity between the corresponding objects.

To support this impression by quantitative data, we used the ten-dimensional representation of the 20 novel objects in two tasks: categorization and recognition. The same arrangement of 169 test views per object as before was used here. In the categorization task, an error was counted for each view that was attributed to

Table 1. *RBF activities (averaged over all 169 test views) for the trained objects*

(Each row shows the average activation pattern induced by views of one of the objects over the ten reference-object RBF modules; boldface indicates the largest entry (see § 4 *a*).)

	cow 1	cat 2	A1	General	tuna	Land-rover	Nissan	F16	fly	*T. Rex*
cow 1	**2.84**	0.76	0.08	0.27	0.82	0.24	0.49	0.45	0.79	0.19
cat 2	1.15	**1.80**	0.06	0.20	1.14	0.23	0.64	0.38	0.83	0.23
A1	0.22	0.10	**2.23**	0.24	0.24	0.07	0.05	0.10	0.90	0.05
General	1.09	0.41	0.52	**2.45**	0.56	0.09	0.23	0.54	1.96	0.44
tuna	0.55	0.56	0.02	0.08	**2.96**	0.08	0.54	0.44	0.47	0.21
Landrover	0.65	0.57	0.11	0.05	0.97	**1.51**	0.64	0.24	0.58	0.08
Nissan	0.96	1.07	0.04	0.13	1.86	0.58	**2.46**	0.90	0.81	0.28
F16	0.74	0.50	0.06	0.21	0.99	0.12	0.65	**1.99**	0.74	0.24
fly	0.45	0.27	0.17	0.27	0.33	0.07	0.14	0.19	**2.86**	0.20
T. Rex	0.36	0.30	0.03	0.14	0.56	0.03	0.16	0.13	0.88	**3.04**

an incorrect category by a *k*-nearest neighbour (*k*-NN) algorithm (Duda & Hart 1973), as explained below.

First, we assigned a category label to each of the ten training objects (for instance, 'cow' and 'cat' were both labelled as 'quadrupeds'). Second, we represented the stimulus view as a ten-element vector of RBF-module responses. Third, we determined the labels of the $k=16$ nearest neighbours of the stimulus among the 169×10 vectors corresponding to all the views of the reference objects (other values of *k*, ranging from 2 to over 100, yielded essentially the same results). Fourth, we let the majority of those *k* votes decide the category label of the stimulus view. This procedure resulted in a misclassification rate of 21% (see table 2).

In the recognition task (i.e. when all 20 object identity labels were used instead of the seven category labels), the error rate was 17%. When only 25 views spanning the range of $\pm 20°$ around the canonical orientation of each object were considered, the recognition error rate dropped to 1.5%.

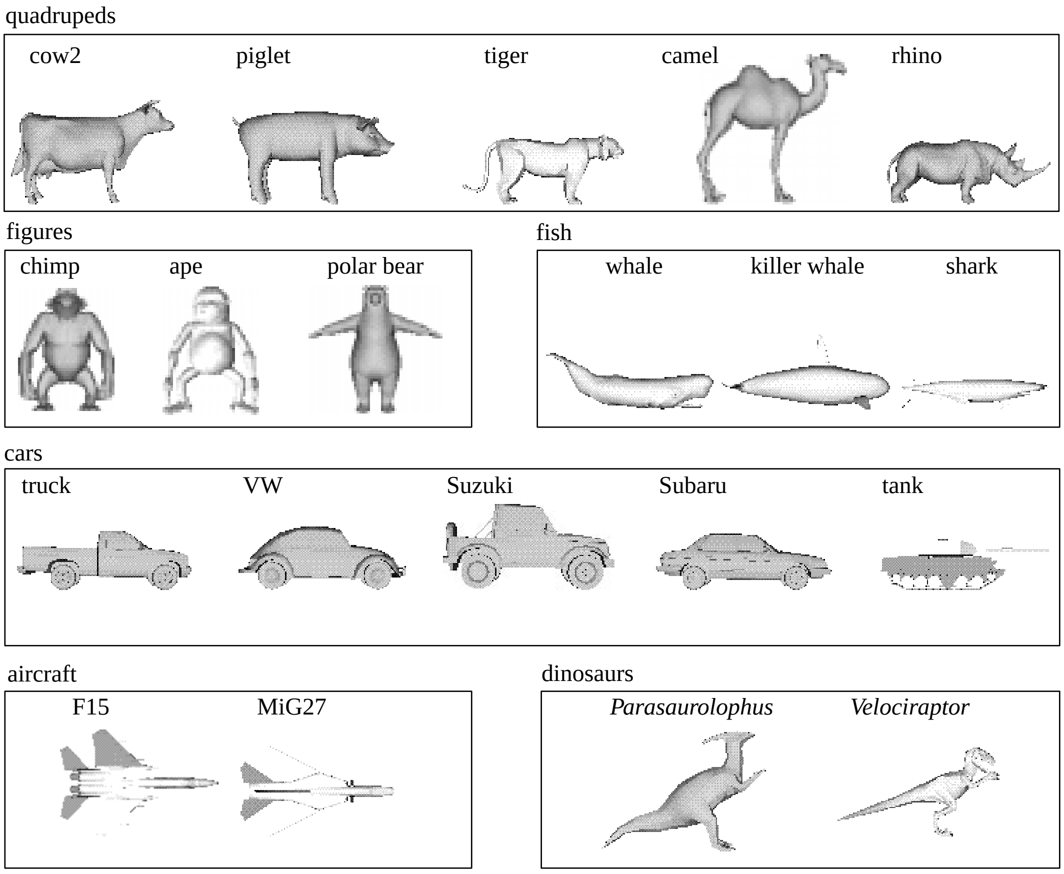

Figure 5. The 20 novel objects used to test the categorization ability of the model (see § 4 *b*); objects are grouped by shape category.

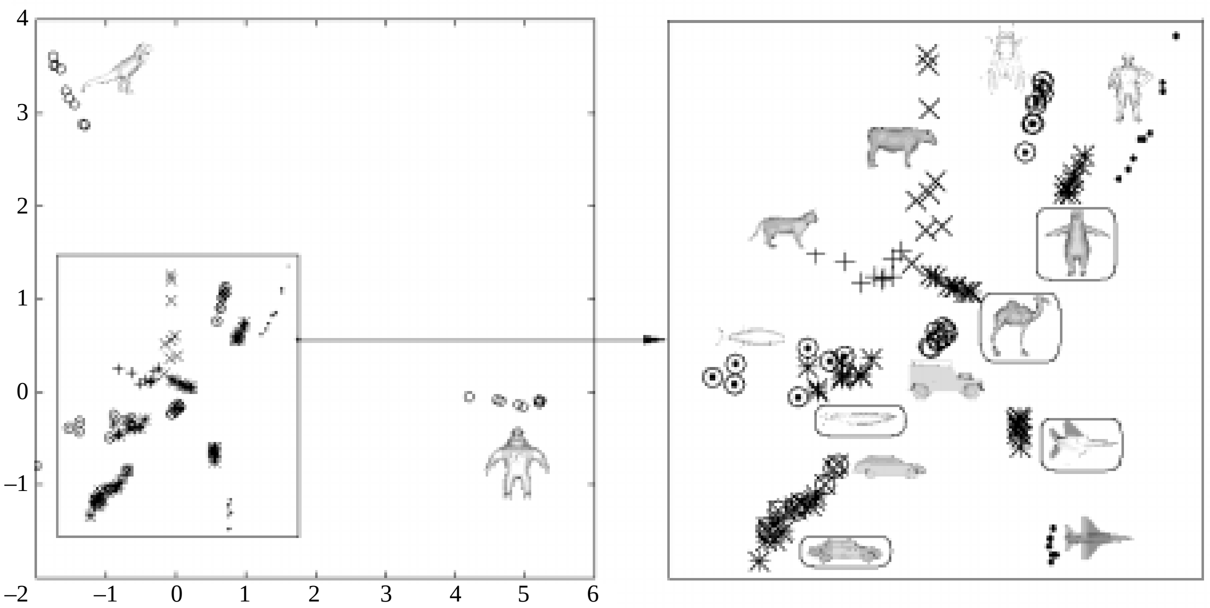

Figure 6. A 2D plot of the ten-dimensional shape space spanned by the outputs of the RBF modules (same-category test objects); multidimensional scaling (MDS) was used to render the 10D space in 2D, while preserving as much as possible distances in the original space (Shepard 1980). Each point corresponds to a test view of one of the objects; nine views of each of the ten training and five novel objects (camel, polar bear, shark, Subaru, F15, denoted by *) are shown. *Left*: the layout of the test views of all 15 objects. *Right*: an enlargement of the central portion of the plot. Note that views belonging to the same object tend to cluster (the residual spread of each cluster can be attributed to the constraint, imposed by MDS, of fitting the two dimensions of the viewpoint variation *and* the dimensions of the shape variation into the same 2D space of the plot). Note also that clusters corresponding to similar objects (e.g. the quadrupeds) are near each other. The icons of the objects appear near the corresponding view clusters; those of five novel objects are drawn in cartouche.

(*c*) *Representation*

Our third experiment tested the ability of 'chorus' to represent and discriminate ten novel objects, picked at random from the database. We used again the same arrangement of 169 test views per object as before. The representation of the test objects is summarized in table 3, which shows the activation of the ten reference-shape RBF modules, produced by each object. It is instructive to consider the patterns of similarities revealed in this distributed ten-dimensional representation of the test objects. For instance, the 'giraffe' turns out to be similar to the two quadrupeds present in the training set (cow and cat), as well as to the dinosaur (*Tyrannosaurus rex*), for obvious reasons (it is also similar to the tuna and to the fly, for reasons which are less obvious, but immaterial: both these shapes are similar to most test objects, which makes their contribution to the representation uninformative). Thus, in the spirit of figure 3, the giraffe can be represented by the vector [1.40 0.99 1.17] of similarities to three reference objects which turn out to be informative in this discrimination context (cow, cat, and *T. rex*).

To visualize the representation of the novel objects, we used again multidimensional scaling (figure 8). As for the same-category objects, the model clustered views by object identity, and grouped view clusters by similarity between the corresponding objects. In a quantitative estimate of recognition performance, the *k*-NN algorithm yielded an error rate of 10% over the 169 test views of the ten novel objects. When only 25 views spanning the range of $\pm 20^\circ$ around the canonical orientation of each object were considered, the error rate dropped to 0.5%. This improvement may be attributed in part to the exclusion of non-representative views, e.g., the head-on view of the manatee, which is easily confused with the top view of the pawn.

5. DISCUSSION

We have described a computational model of shape-based recognition and categorization, which encodes stimuli by their similarities to a number of reference shapes, themselves represented by specially trained dedicated modules. The performance of the model suggests that this principle may allow for efficient representation, and, in most cases, correct categorization, of shapes never before encountered by the observer—a goal which we consider of greater importance than mere recognition of previously seen objects, and which so far has eluded the designers of computer vision systems.

(*a*) *Implications for theories of visual representation*

In computer vision, one may discern three main theoretical approaches to object representation: pictorial representations, structural descriptions, and feature spaces (Ullman 1989). According to the first approach, objects are represented by the same kind of

Table 2. *Categorization results for the 20 test objects shown in figure 5*

(Each row corresponds to one of the test objects; the proportion of the 169 test views of that object attributed to each of the seven categories present in the training set appears in the appropriate column. Each row sums to somewhat less than 1.0, because near-zero entries were omitted for clarity; boldface indicates the largest entry in each row. The mean misclassification rate over all 169 views of all objects is 21%; when only 13 views are considered (spaced at 10° around the equator of the viewing sphere, centred on the canonical view), the misclassification rate drops to 15%.)

		quadrupeds		figures		fish	cars		aircraft	fly	dinosaurs
categories	objects	cow 1	cat 2	Al	Gen-eral	tuna	Land-rover	Nissan	F16	fly	*T. Rex*
quadrupeds	cow 2	**0.66**	0.29	—	—	—	0.03	—	—	—	—
	piglet	**0.76**	0.05	—	0.01	—	0.10	—	—	0.04	—
	tiger	0.18	**0.52**	0.01	0.01	0.01	0.21	—	0.01	0.01	—
	camel	**0.61**	—	—	—	—	0.37	—	—	—	—
	rhino	**0.62**	0.08	—	—	—	0.04	0.14	0.10	—	—
figures	chimp	0.22	0.04	**0.57**	—	—	—	0.02	—	0.11	—
	ape	0.02	0.05	**0.52**	0.02	0.01	—	—	0.03	0.31	—
	polar-bear	0.02	0.07	—	**0.83**	—	—	—	0.01	0.05	—
fish	whale	0.05	0.04	—	—	**0.53**	0.3	—	0.06	—	—
	killer-whale	0.04	0.02	0.01	—	**0.83**	0.02	—	0.02	0.02	—
	shark	0.01	0.02	—	—	**0.84**	0.04	0.05	0.01	—	—
cars	truck	0.01	0.03	—	—	—	—	**0.93**	—	—	—
	VW	—	0.01	—	—	0.01	0.01	**0.92**	—	0.01	—
	Suzuki	0.05	0.11	—	—	0.13	**0.41**	0.28	—	—	—
	Subaru	—	—	—	—	—	0.01	**0.96**	—	—	—
	tank	0.02	0/02	—	—	0.03	—	**0.71**	0.19	—	—
aircraft	F15	—	0.01	—	0.01	—	—	—	**0.92**	—	0.03
	MiG27	—	—	—	—	0.04	0.17	—	**0.78**	—	—
dinosaurs	*Parasaurolophus*	—	—	—	—	0.04	—	—	—	—	**0.95**
	Velociraptor	—	—	—	—	0.21	—	—	0.10	—	**0.67**

information one finds in a picture: coordinates of primitive elements (which may be as simple as intensity values of pixels in an image). Because of the effects of factors extrinsic to shape, this mode of representation can be used for recognition only if it is accompanied by a method for normalizing the appearance of objects. The normalization may be carried out explicitly, as in recognition by alignment (Ullman 1989), or implicitly, as in recognition by linear combination of images (Ullman & Basri 1991).

It is not easy to adapt the pictorial approach to carry out categorization rather than recognition. One reason for that is the excessive amount of detail in pictures: much of the information in a snapshot of an object is unnecessary for categorization, as attested by the ability of human observers to classify line drawings of common shapes (Biederman & Ju 1988; Price & Humphreys 1989). Although a metric over images that would downplay within-category differences may be defined in some domains, such as classification of stylized 'clip art' drawings (Ullman 1996, p. 173), attempts to classify pictorially represented 3D objects (vehicles) met with only a limited success (Shapira & Ullman 1991).

We believe that extension of alignment-like approaches from recognition to categorization is problematic for a deeper reason than mere excess of information in images of objects. Note that both stages in the process of recognition by alignment (normalization and comparison; see Ullman 1989) are geared towards pairing the stimulus with a *single* stored representation (which may be the average of several actual objects, as in Basri's algorithm (Basri 1996)). As we pointed out in the introduction, this strategy, designed to culminate in a winner-takes-all decision, is inherently incompatible with the need to represent radically novel objects.

The ability to deal with novel objects has been considered so far the prerogative of structural approaches to representation (Marr & Nishihara 1978; Biederman 1987). The structural approach employs a small number of generic primitives (such as the thirty-odd geons postulated by Biederman), along with spatial relationships defined over sets of primitives, to represent a very large variety of shapes. The classification problem here is addressed by assigning objects that have the same structural description to the same category.

In principle, even completely novel shapes can be given a structural description, because the extraction of primitives from images and the determination of spatial relationships is supposed to proceed in a purely bottom-up, or image-driven fashion. In practice, however, both these steps have so far proved to be impossible to automate. State-of-the-art computer vision systems either ignore the challenge posed by the problems of categorization and of representation of novel objects (Murase & Nayar 1995), or treat categorization as a byproduct of recognition (Mel 1996).

Table 3. *RBF activities (averaged over all 169 test views) for the ten test objects shown in figure 7*

(In each row, corresponding to a different test object, entries within 50% of the maximum for that row are marked by boldface. These entries constitute a low-dimensional representation of the test object whose label appears at the head of the row, in terms of similarities to some of the ten reference objects. Unlike in table 1, the test objects here are *not* familiar to the system; their representation is made possible by the significant response of at least some of the reference-object modules to novel stimuli. The utility of this representation is illustrated by the sensible layout of its equivalent shape space (figure 8): objects whose shapes are similar are indeed grouped together. For instance, the 'manatee' (an aquatic mammal known as the sea cow) turns out to be like (in decreasing order of similarity), a 'tuna', a 'cow', and, interestingly, but perhaps not surprisingly, a 'Nissan' wagon.)

	cow 1	cat 2	Al	General	tuna	Landrover	Nissan	F16	fly	*T. Rex*
butterfly	**1.19**	**0.81**	0.05	0.18	**1.03**	0.35	**0.74**	0.53	**0.88**	0.29
frog	0.19	0.12	0.29	0.09	0.20	0.08	0.08	0.08	**0.99**	0.10
tennis shoe	0.25	0.31	0.05	0.06	**0.79**	0.15	**0.40**	0.27	**0.55**	0.09
pump	**0.77**	**0.58**	0.02	0.09	**1.12**	0.13	**0.75**	0.46	**0.65**	0.12
Beethoven	0.04	0.02	0.12	0.01	0.04	0.02	0.00	0.01	**0.39**	0.00
giraffe	**1.40**	**0.99**	0.02	0.28	**1.64**	0.07	0.68	0.78	**1.28**	**1.17**
pawn	0.24	0.08	0.21	0.16	0.08	0.02	0.01	0.02	**1.08**	0.03
manatee	**0.84**	0.71	0.07	0.17	**1.49**	0.13	**0.76**	0.61	0.71	0.16
Fiat	0.89	0.80	0.00	0.07	**1.98**	0.17	**1.61**	0.72	0.59	0.17
Toyota	1.17	1.06	0.08	0.12	**1.63**	0.87	**1.67**	0.66	0.71	0.18

In comparison to all these approaches, the 'chorus' model is designed to treat both familiar and novel objects equivalently, as points in a shape space spanned by similarities to a handful of reference objects (according to Ullman's taxonomy, this makes it an instance of the feature-based approach, the features being similarities to entire objects). The minimalistic implementation of 'chorus' described in the preceding sections achieved recognition performance on a par with that of the state-of-the-art computer vision systems, despite relying only on shape cues where other systems use colour and/or texture together with shape (Murase & Nayar 1995; Mel 1996; Schiele & Crowley 1996). Furthermore, this performance was achieved with a low-dimensional representation (ten nominal dimensions), whereas the other systems typically employ about a hundred dimensions; for a discussion of the importance of low dimensionality in this context, see (Edelman & Intrator 1997). Finally, our model also exhibited significant capabilities for shape-based categorization and for useful representation of novel objects; it is reasonable to assume that its performance in these tasks can be improved, if more lessons from biological vision are incorporated into the system.

(*b*) *Implications for understanding object representation in primate vision*

The architecture of 'chorus' reflects our belief that a good way to achieve progress in computer vision is to follow examples set by biological vision. Each of the building blocks of 'chorus', as well as its general layout, can be readily interpreted in terms of well-established properties of the functional architecture of the primate visual system. The basic mechanism in the

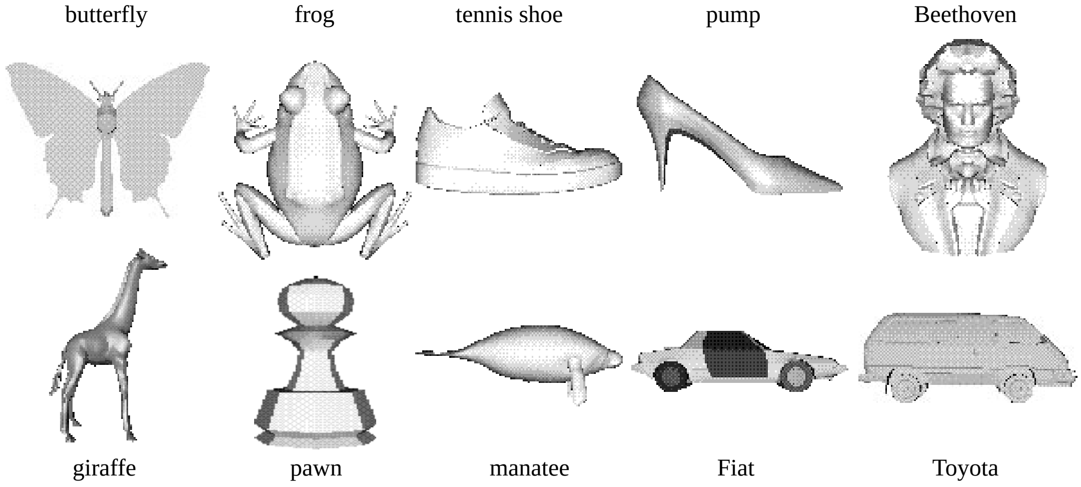

Figure 7. The ten novel objects, picked at random from the object database, which we used to test the representational abilities of the model (see § 4 *c*).

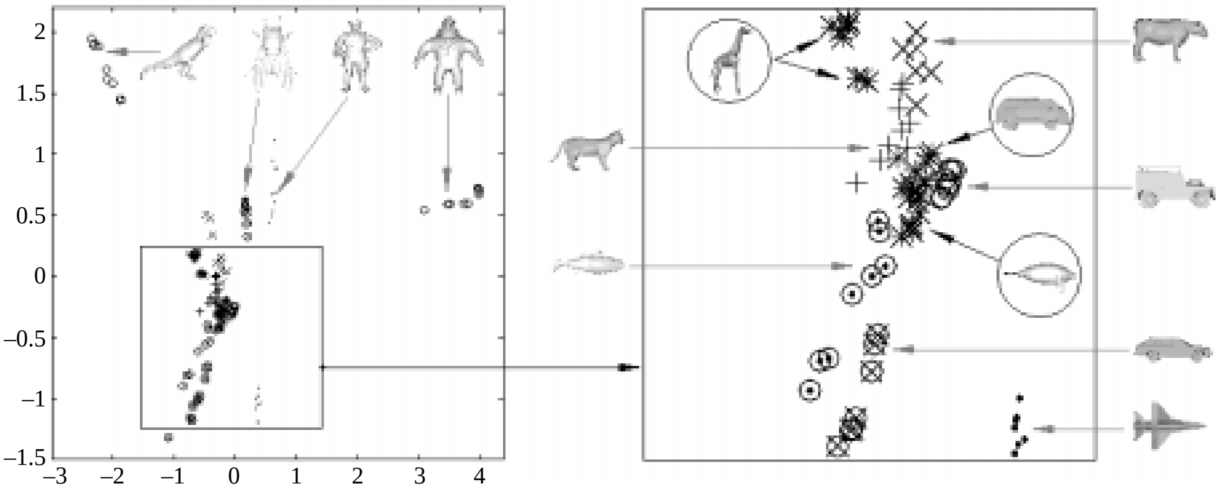

Figure 8. A 2D MDS plot of the ten-dimensional shape space spanned by the outputs of the RBF modules (novel test objects). As in figure 6, each point corresponds to a test view of one of the objects; views of the ten training objects and of three novel objects (giraffe, Toyota, manatee, denoted by *) are shown. *Left*: the layout of the test views of all 13 objects. *Right*: an enlargement of the central portion of the plot. As for the same-category test objects, views belonging to the same object cluster together, and clusters corresponding to similar objects are near each other. The three novel objects are circled.

implementation of this scheme is a *receptive field*—probably the most ubiquitous functional abstraction of the physiologist's tuned unit, widely used in theories of biological information processing (Edelman 1997*a*). The receptive fields at the front end of 'chorus' are intended to parallel those of the photoreceptors (the use of centre-surround receptive fields at a variety of orientations and scales, as in the primate visual cortex, should improve the present results). Furthermore, an RBF module of the kind used in the subsequent stage of 'chorus' can be seen also as a receptive field, tuned both to a certain location in the visual field (defined by the extent of the front-end receptive fields) and to a certain location in the shape space (corresponding to the shape of the object on which the module has been trained).

Functional counterparts both of individual components (basis functions) of RBF modules and of entire modules have been found in a recent electrophysiological study of the inferotemporal (IT) cortex in awake monkeys (Logothetis *et al.* 1995). The former correspond to cells tuned to particular views of objects familiar to the animal; the latter to cells that respond nearly equally to a wide range of views of the same object. It is easy to imagine how an ensemble of cells of the latter kind, each tuned to a different reference object, can span an internal shape space, after the manner suggested above.

While a direct test of this conjecture still awaits experimental confirmation, indirect evidence suggests that a mechanism not unlike the 'chorus of prototypes' is deployed in the IT cortex. This evidence is provided by the work of K. Tanaka and his collaborators, who studied object representation in the cortex of anaesthetized monkeys (Tanaka 1992; Tanaka 1996). These studies revealed cells tuned to a variety of simple shapes, arranged so that units responding to similar shapes were clustered in columns running perpendicular to the cortical surface; the set of stimuli that proved effective depended to some extent on the monkey's prior visual experience. If further experimentation reveals that a given object consistently activates a certain possibly disconnected subset of the columns, and if that pattern of activation smoothly changes in response to a continuous change in the shape or the orientation (Wang *et al.* 1996) of the stimulus, the principle of representation of similarity that serves as the basis of 'chorus' would be implicated also as the principle behind shape representation in the cortex.

The results of several recent psychophysical studies of object representation in primates support the above conjecture. In each of a series of experiments, which involved subjective judgement of shape similarity and delayed matching to samples, human subjects (Edelman 1995*a*; Cutzu & Edelman 1996) and monkeys (Sugihara *et al.* 1996) have been confronted with several classes of computer-rendered 3D animal-like shapes, arranged in a complex pattern in a common parameter space (cf. Shepard & Cermak 1973). In each experiment, processing of the subject data by multidimensional scaling (used to embed points corresponding to the stimuli into a 2D space for the purpose of visualization) invariably revealed the low-dimensional parametric structure of the set of stimuli. In other words, the proximal shape space internalized by the subjects formed a faithful replica of the distal shape space structure imposed on the stimuli. Furthermore, this recovery was reproduced by a 'chorus'-like model, trained on a subset of the stimuli and subsequently exposed to the same test images shown to the subjects. As we argue elsewhere, these findings may help understand the general issue of cognitive representation, and, in particular, the manner in which representation can conform (be faithful), to its object (Edelman & Duvdevani-Bar 1997; Edelman 1997*b*); their full integration will require a coordinated effort in the fields of behavioural physiology, psychophysics, and computational modelling.

We thank M. Dill, N. Intrator, P. Sinha, and two anonymous reviewers for helpful suggestions concerning the manuscript.

REFERENCES

Basri, R. 1996 Recognition by prototypes. *Int. J. Computer Vision* **19**, 147–168.

Baxter, J. 1995 The canonical metric for vector quantization. NeuroCOLT technical report NC-TR-95-047, University of London.

Biederman, I. 1987 Recognition by components: a theory of human image understanding. *Psychol. Rev.* **94**, 115–147.

Biederman, I. & Ju, G. 1988 Surface versus edge-based determinants of visual recognition. *Cognitive Psychol.* **20**, 38–64.

Broomhead, D.-S. & Lowe, D. 1988 Multivariable functional interpolation and adaptive networks. *Complex Syst.* **2**, 321–355.

Bülthoff, H.-H. & Edelman, S. 1992 Psychophysical support for a 2D view interpolation theory of object recognition. *Proc. Natn. Acad. Sci. USA* **89**, 60–64.

Cover, T. & Hart, P. 1967 Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **IT-13**, 21–27.

Cutzu, F. & Edelman, S. 1996 Faithful representation of similarities among three-dimensional shapes in human vision. *Proc. Natn. Acad. Sci. USA* **93**, 12046–12050.

Duda, R.-O. & Hart, P.-E. 1973 *Pattern classification and scene analysis*. New York: Wiley.

Edelman, S. 1995*a* Representation of similarity in 3D object discrimination. *Neural Computation* **7**, 407–422.

Edelman, S. 1995*b* Representation, similarity, and the chorus of prototypes. *Minds & Machines* **5**, 45–68.

Edelman, S. 1997*a* Receptive fields for vision: from hyperacuity to object recognition. In *Vision* (ed. R. Watt). Cambridge, MA: MIT Press. (In the press.)

Edelman, S. 1997*b* Representation is representation of similarity. *Behav. Brain Sci.* (In the press.)

Edelman, S. & Duvdevani-Bar, S. 1997 Similarity, connectionism, and the problem of representation in vision. *Neural Computation* **9**, 701–720.

Edelman, S. & Intrator, N. 1997 Learning as extraction of low-dimensional representations. In *Mechanisms of perceptual learning* (ed. D. Medin, R. Goldstone, & Schyns, P.). San Diego, CA: Academic Press. (In the press.)

Edelman, S., Reisfeld, D. & Yeshurun, Y. 1992 Learning to recognize faces from examples. In *Proc. 2nd European Conf. Computer Vision, Lecture Notes in Computer Science, vol. 588* (ed. G. Sandini), pp. 787–791. Berlin: Springer.

Jacobs, D.-W. 1996 The space requirements of indexing under perspective projections. *IEEE Trans. Pattern Analysis & Machine Intelligence* **18**, 330–333.

Jolicoeur, P., Gluck, M. & Kosslyn, S.-M. 1984 Pictures and names: making the connection. *Cognitive Psychol.* **16**, 243–275.

Kendall, D.-G. 1984 Shape manifolds, Procrustean metrics and complex projective spaces. *Bull. Lond. Math. Soc.* **16**, 81–121.

Linde, Y., Buzo, A. & Gray, R. 1980 An algorithm for vector quantizer design. *IEEE Trans. Communications* **COM-28**, 84–95.

Logothetis, N.-K., Pauls, J. & Poggio, T. 1995 Shape recognition in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563.

Marr, D. & Nishihara, H.-K. 1978 Representation and recognition of the spatial organization of three-dimensional structure. *Proc. R. Soc. Lond.* B **200**, 269–294.

Mel, B. 1996 *SEEMORE: combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition*. Technical report, University of South California, Los Angeles, CA.

Murase, H. & Nayar, S. 1995 Visual learning and recognition of 3D objects from appearance. *Int. J. Computer Vision* **14**, 5–24.

Palmer, S.-E., Rosch, E. & Chase, P. 1981 Canonical perspective and the perception of objects. In *Attention and performance,* vol. IX (ed. J. Long & A. Baddeley), pp. 135–151. Hillsdale, NJ: Erlbaum.

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

Poggio, T. & Girosi, F. 1990 Regularization algorithms for learning that are equivalent to multilayer networks. *Science* **247**, 978–982.

Price, C.-J. & Humphreys, G.-W. 1989 The effects of surface detail on object categorization and naming. *Q. Jl Exp. Psychol.* A **41**, 797–828.

Rosch, E. 1978 Principles of categorization. In *Cognition and categorization* (ed. E. Rosch, & B. Lloyd), pp. 27–48. Hillsdale, NJ: Erlbaum.

Schiele, B. & Crowley, J.-L. 1996 Object recognition using multidimensional receptive field histograms. In *Proc. European Conf. Computer Vision 1996, vol. 1, Lecture Notes in Computer Science* (ed. B. Buxton & R. Cipolla), pp. 610–619. Berlin: Springer.

Shapira, Y. & Ullman, S. 1991 A pictorial approach to object classification. *Proc. IJCAI,* 1257–1263.

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 Psychol.* **4**, 351–377.

Smith, E.-E. 1990 Categorization. In *An invitation to cognitive science: thinking, vol. 2* (ed. D.-N. Osherson, & Smith, E.-E.), pp. 33–53. Cambridge, MA: MIT Press.

Sugihara, T., Edelman, S. & Tanaka, K. 1996 Representation of objective similarity among 3D shapes in the monkey. In *Proc. ARVO, Invest. Ophthalm. Vis. Sci. Suppl.* **37**.

Tanaka, K. 1992 Inferotemporal cortex and higher visual functions. *Curr. Opinion Neurobiol.* **2**, 502–505.

Tanaka, K. 1996 Inferotemporal cortex and object vision. *A. Rev. Neurosci.* **19**, 109–139.

Ullman, S. 1989 Aligning pictorial descriptions: an approach to object recognition. *Cognition* **32**, 193–254.

Ullman, S. 1996 *High level vision*. Cambridge, MA: MIT Press.

Ullman, S. & Basri, R. 1991 Recognition by linear combinations of models. *IEEE Trans. Pattern Analysis & Machine Intelligence* **13**, 992–1005.

Wang, G., Tanaka, K. & Tanifuji, M. 1996 Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* **272**, 1665–1668.