Similarity to reference shapes as a basis for shape representation

Shimon Edelman **Florin Cutzu** **Sharon Duvdevani-Bar**
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
edelman@wisdom.weizmann.ac.il

Abstract

We present a unified approach to visual representation, addressing both the needs of superordinate and basic-level categorization and of identification of specific instances of familiar categories. According to the proposed theory, a shape is represented by its similarity to a number of reference shapes, measured in a high-dimensional space of elementary features. This amounts to embedding the stimulus in a low-dimensional proximal shape space. That space turns out to support representation of distal shape similarities which is veridical in the sense of Shepard's (1968) notion of second-order isomorphism (i.e., correspondence between distal and proximal similarities among shapes, rather than between distal shapes and their proximal representations). Furthermore, a general expression for similarity between two stimuli, based on comparisons to reference shapes, can be used to derive models of perceived similarity ranging from continuous, symmetric, and hierarchical, as in the multidimensional scaling models (Shepard, 1980), to discrete and non-hierarchical, as in the general contrast models (Tversky, 1977; Shepard and Arabie, 1979).

Introduction

All but a few current theoretical treatments of visual representation still adhere to the Aristotelian doctrine of *representation by similarity*, according to which an internal entity represents an external object by virtue of resemblance between the two.[1] Simply put, the original version of that doctrine holds that the representation of a tomato has something of the redness and of the roundness of the real thing. The predominant theories of visual *shape* representation still speak about isomorphism: typically, it is assumed that structural (Biederman, 1987) or metric (Ullman, 1989) information stored in the brain reflects corresponding properties of shapes in the world. In comparison, no student of *color* vision seriously believes that representations of tomatoes are red, or even that the reflectance spectra of tomatoes are explicitly stored; this has been supplanted by the feature detector theory, according to which the response of internal mechanisms tuned to particular sensory stimuli constitute the basic representation for those stimuli. A major goal of the present paper is to show that shape too, over and above color or local orientation, can be encoded in a low-dimensional feature space.

An important step towards that goal has been made by Roger N. Shepard, who pointed out that instead of a first-order isomorphism between the shapes and their representations, it makes more sense to expect a second-order isomorphism between similarities of shapes and similarities of the internal representations they induce (Shepard, 1968). Essentially, this is a call for representation *of* similarity instead of representation *by* similarity.

A representation of a collection of shapes is veridical in Shepard's sense, if the mapping it implies between (some parameterization of) the distal shape space and the internal, or proximal, representation space preserves similarity ranks. Elsewhere, we show that a distal to proximal mapping realized by a bank of typical connectionist classifiers, each tuned to a particular shape, is likely to satisfy the requirements for similarity rank preservation generically, over appropriately limited regions of the distal space (Edelman, 1995b; Duvdevani-Bar and Edelman, 1995).

Here, we extend this theory of representation in two directions. First, we outline a common framework for treating categorization, recognition and identification as measurements of similarities to subspaces of the image space. Second, we show how similarity can be defined in such a manner as to form a bridge between theories of representation based on continuous feature spaces, and those based on lists of discrete-valued features. We conclude with a brief mention of some of the results supporting the theory, in areas ranging from psychophysics and physiology to computation and philosophy.

Representation = measurement + dimensionality reduction

In any cognitive system, the internal representations are constructed by subjecting the input to a set of measurements, whose aim is to provide an efficient description of the stimuli, e.g., as points in some low-dimensional parameter space. Because such a space is neither directly accessible nor known *a priori*, and because different tasks may call for different aspects of the stimuli to be represented, it is a good strategy to carry out as many measurements as possible, to increase the likelihood of correspondence between some subspace of the measurement space $\mathcal{M}$ and the relevant part of the parameter space. This makes $\mathcal{M}$ high-dimensional, and necessitates subsequent dimensionality reduction, whose aim is to recover the relevant subspace of $\mathcal{M}$. Likewise, the input to an object recognition system – an $n \times n$ image – can be considered as a point in a n^2-dimensional image or *raster* space $\mathcal{R} = R^{n^2}$, which we identify with the measurement space $\mathcal{M}$ (in biological vision, one may think of the space of patterns transmitted by the optic nerve to the brain). The task of recognition is, given $\mathbf{X} \in \mathcal{R}$, to determine whether $\mathbf{X}$ is an image of an

[1] "Representation of something is an image, model, or reproduction of that thing," (Suppes, Pavel, and Falmagne, 1994).

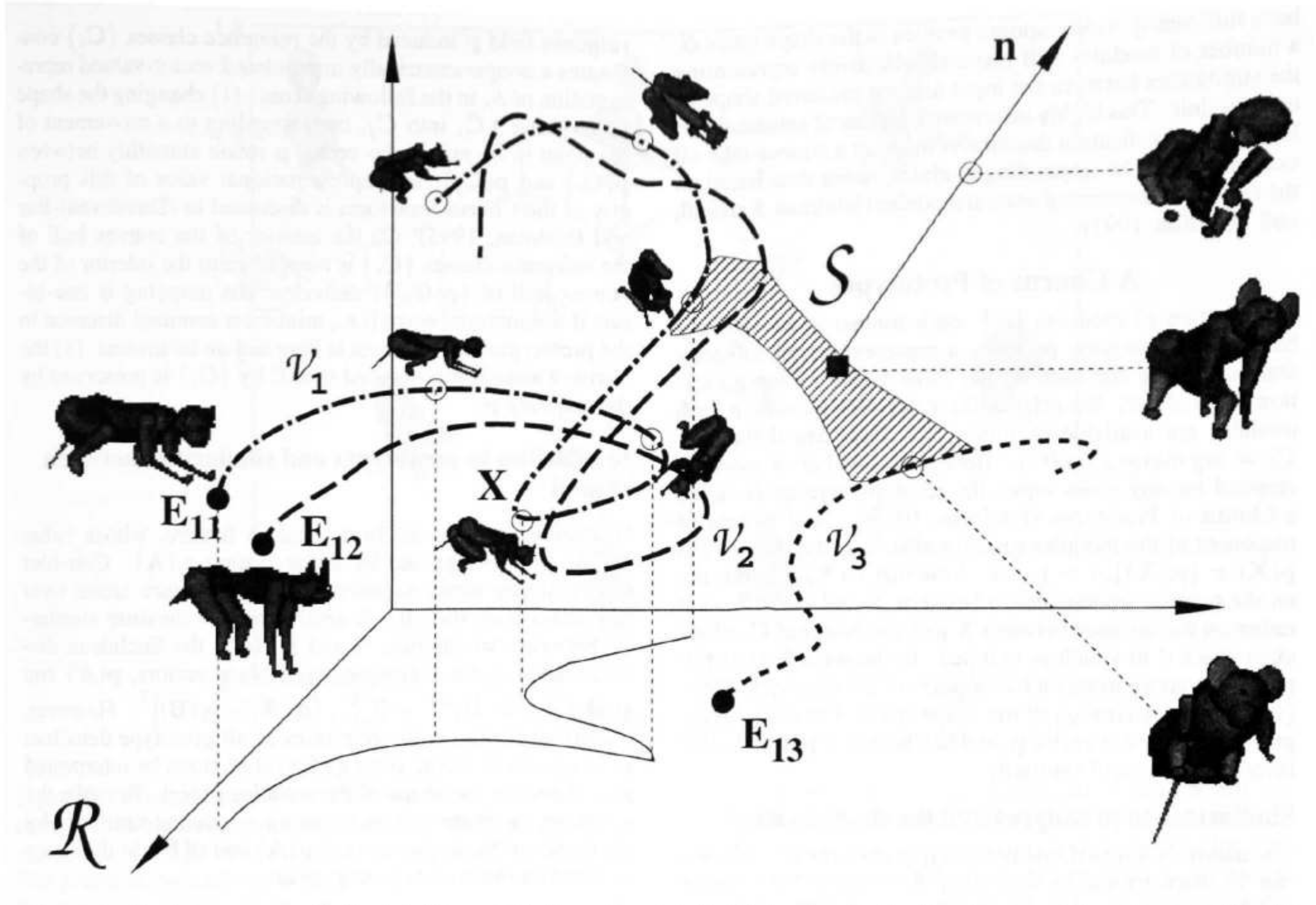

Figure 1: The image space, $\mathcal{R}$ (depicted here as 3-dimensional, to facilitate visualization), and some of its subspaces. The $\mathcal{V}_i$ (shown as dashed lines) are the view spaces for the three exemplars $\mathbf{E}_{1i}$ (marked by filled circles), all of which belong to the same class $\mathcal{C}_1$ (the class of 4-legged animal shapes). Some of the different views of $\mathbf{E}_{11}$ are shown (marked by open circles). The surface patch represents a part of the shape space $\mathcal{S}$, and the vector $\mathbf{n}$ – a normal to it. Movement along this direction in $\mathcal{R}$ corresponds to a reduction in the resemblance between the resulting image and the images of coherently looking objects. Image $\mathbf{X}$ should be classified as belonging to exemplar $\mathbf{E}_{11}$, class $\mathcal{C}_1$, and, of course, to the shape space $\mathcal{S}$.

object (a coherent entity, which, in intuitive terms, looks like something, rather than like random pixel noise), and, if it is, to establish the category to which the object belongs, and, if possible, the object's identity. It is convenient to cast this problem in terms of attributing to $\mathbf{X}$ a proper location, respectively, in the *shape* space $\mathcal{S}$, the *class* space $\mathcal{C}$, and the *exemplar* space $\mathcal{E}$, where $\mathcal{R} \supset \mathcal{S} \supset \mathcal{C}$, and $\mathcal{E} = \mathcal{E}(\mathcal{C})$ (see Figure 1). A complete characterization of an input calls for determining i, j, k, such that $\mathbf{X} \in \mathbf{E}_{jk}$, and $\mathbf{S}_i \subset \mathcal{S}$, $\mathbf{C}_j \subset \mathcal{C}$, $\mathbf{E}_{jk} \subset \mathcal{E}_j$ ($\mathcal{E}_j = \mathcal{E}(\mathbf{C}_j)$).

Basic level. Consider first the basic-level categorization problem: given $\mathbf{X}$, find j such that $\mathbf{X} \in \mathbf{C}_j$. The major obstacle to be overcome here is the dependence of the appearance of $\mathbf{X} \in \mathbf{C}_j$ on factors such as illumination and viewpoint, in addition to the category identity j. If $\mathbf{C}_j$ is taken to correspond to the image of a member of j in some canonical orientation, the viewing conditions can be seen to span a *view* space $\mathcal{V}_j$, which is, to a first approximation, orthogonal to the class space $\mathcal{C}$, and pierces it at $\mathbf{C} = \mathbf{C}_j$. By training a general-purpose function approximation module to perform the mapping $T(j) : \mathcal{V}_j \rightarrow \mathbf{C}_j$, one can largely eliminate the dependence of categorization on viewing conditions (Poggio and Edelman, 1990). The normalizing transformation $T(j)$ can work even for inputs not previously encountered by the system (that is, for different instances $\mathbf{E}_{jk}$), provided that they belong to the class j (Lando and Edelman, 1995).

Subordinate levels. The central problem in determining the identity k lies in the fine resolution that must be attained within the instance space $\mathcal{E}_j$, in the face of the residual misalignment left over from the action of the normalizing transformation T. This problem can be approached by learning hyperacuity in the instance space, as it is done in other hyperacuity-related tasks (Poggio, Fahle, and Edelman, 1992); experience shows that hyperacuity can be attained despite considerable misalignment of the stimulus as a whole, relative to its "home" or training pose.

Superordinate levels. The most challenging problem arises when the system encounters an unfamiliar shape, belonging to none of the classes for which specially trained categorization modules are available. The key to a solution here lies in considering the *population response* at the basic categorization level (Edelman, 1995b). If the existing modules have sufficiently wide response profiles in the shape space $\mathcal{S}$, a number of modules will respond, effectively representing the similarities between the input and the preferred shape of each module. This highly informative pattern of similarities is lost if the classification decision is made in a winner-take-all manner among the responding modules, rather than based on the ensemble response of several modules (Edelman, Reisfeld, and Yeshurun, 1992).

A Chorus of Prototypes

A collection of modules $\{p_i\}$, each trained to recognize a basic shape category, provides a representational substrate that is suitable for each of the three levels of categorization listed above. We refer to the categories for which such modules are available as *prototypes*; these are defined as $\mathbf{C}_i = \arg\max_{\mathbf{C} \subset \mathcal{C}} p_i(\mathbf{C})$. Because a number of modules respond for any given input, the resulting scheme is called a Chorus of Prototypes (Edelman, 1995b). The pattern of responses of the modules to a stimulus $\mathbf{X}$ is the ordered list $\mathbf{p}(\mathbf{X}) = \{p_i(\mathbf{X})\}, i = 1 \ldots k$. Note that $p_i(\mathbf{X})$ depends not on the point-to-point distance between $\mathbf{X}$ and some $\mathbf{X}_i$, but rather on the distance between $\mathbf{X}$ and that member $\mathbf{C}_i$ of the class space $\mathcal{C}$ to which p_i is tuned. In the remainder of this paper, we concentrate on two aspects of the Chorus scheme: (1) the characterization of the shape space $\mathcal{S}$ in terms of the prototype response vector $\mathbf{p}$, and (2) the use of $\mathbf{p}$ in tasks that involve judgment of similarity.

Similarities to prototypes and the shape space $\mathcal{S}$

The nature of dimensionality reduction performed by Chorus can be characterized by describing the relationship between the shape space $\mathcal{S}$ and the vector of responses of the prototype modules, $\mathbf{p}$. One way to do that is by viewing the action of Chorus as interpolation: intuitively, one would expect the shape space to be a (hyper)surface that passes through the reference classes $\{\mathbf{C}_i\}$ and behaves reasonably in between (see Figure 1). Now, different tasks carry with them different notions of reasonable behavior. Consider first the least specific level in a hierarchy of recognition tasks: deciding whether $\mathbf{X}$ is the image of some (familiar) object. For this purpose, it would suffice to represent $\mathcal{S}$ as a scalar field over the image space $S(\mathbf{X}) : \mathcal{R} \rightarrow R$, which would express for each $\mathbf{X}$ its degree of membership in $\mathcal{S}$. For example, we may set $S = \max_i\{p_i\}$ (the activity of the strongest-responding prototype module), or shape $= \sum_i p_i$ (the total activity; cf. Nosofsky, 1988). We remark that it should be possible to characterize a superordinate-level category of the input image, and not merely decide whether it is likely to be the image of a familiar object, by determining the identities of the prototype modules that respond above some threshold (i.e., if, say, the cat, the sheep and the cow modules are the only ones that respond, the stimulus is probably a four-legged animal).

At the basic and the subordinate category levels, we are interested in the location of the input *within* $\mathcal{S}$, which, therefore, can no longer be considered a scalar. Note that parametric interpolation is not possible in this case, as the intrinsic dimensionality of $\mathcal{S}$ is not given *a priori*.[2] Now, the prototype response field $\mathbf{p}$ induced by the reference classes $\{\mathbf{C}_i\}$ constitutes a nonparametrically interpolated vector-valued representation of $\mathcal{S}$, in the following sense: (1) changing the shape ("morphing") $\mathbf{C}_i$ into $\mathbf{C}_j$, corresponding to a movement of the point in $\mathcal{S}$, makes the vector $\mathbf{p}$ rotate smoothly between $\mathbf{p}(\mathbf{C}_i)$ and $\mathbf{p}(\mathbf{C}_j)$; the representational value of this property of the Chorus transform is discussed in (Duvdevani-Bar and Edelman, 1995); (2) the interior of the convex hull of the reference classes $\{\mathbf{C}_i\}$ is mapped onto the interior of the convex hull of $\{\mathbf{p}(\mathbf{C}_i)\}$; moreover, the mapping is one-to-one if a minimum-norm (i.e., minimum summed distance to the prototypes) requirement is imposed on its inverse; (3) the Voronoi tessellation induced over $\mathcal{C}$ by $\{\mathbf{C}_i\}$ is preserved by the mapping $\mathbf{p}$.

Similarities to prototypes and similarities between stimuli

In Chorus, each p_i is, in a sense, a feature, whose value for $\mathbf{A} \in \mathcal{R}$ is signified by the activation $p_i(\mathbf{A})$. Consider the similarity structure induced by this feature space over the universe of stimuli. A natural way to measure similarity between two stimuli, $\mathbf{A}$ and $\mathbf{B}$, is by the Euclidean distance between the corresponding feature vectors, $\mathbf{p}(\mathbf{A})$ and $\mathbf{p}(\mathbf{B})$: $s_E(\mathbf{A}, \mathbf{B})^{-1} \sim \sum_{i=1}^{k} \left[p_i(\mathbf{A}) - p_i(\mathbf{B})\right]^2$. However, a uniform scaling in the responses of all prototype detectors $\mathbf{p} \rightarrow c\,\mathbf{p}$ (as in seeing through fog) should not be interpreted as a change in the shape of the stimulus object. To make the similarity insensitive to such scaling, we define similarity by the cosine of the angle between $\mathbf{p}(\mathbf{A})$ and $\mathbf{p}(\mathbf{B})$, in the space spanned by the prototype responses:

$$s_a(\mathbf{A}, \mathbf{B}) \sim \sum_{i=1}^{k} p_i(\mathbf{A}) p_i(\mathbf{B}) \doteq \langle \mathbf{p}(\mathbf{A}), \mathbf{p}(\mathbf{B}) \rangle \qquad (1)$$

This definition of similarity must, however, be further modified, for two reasons. First, s_a is independent of context, whereas perceived similarity depends on the "contrast set" against which it is to be judged. Second, s_a is symmetric, whereas human perception of similarity appears to be asymmetric in many cases (Tversky, 1977). To make s_a depend on the context, we introduce a vector of weights, one per prototype, such that $w_i = w_i\left(\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots\}\right)$. Thus, comparing $\mathbf{A}$ and $\mathbf{B}$ in two contexts, $\{\mathbf{A}, \mathbf{B} \mid \mathbf{C}, \mathbf{D}, \mathbf{E}\}$ and $\{\mathbf{A}, \mathbf{B} \mid \mathbf{F}, \mathbf{G}, \mathbf{H}\}$, may result in different values of similarity between $\mathbf{A}$ and $\mathbf{B}$. To model the asymmetry which frequently arises when subjects are required to estimate the similarity of some stimulus $\mathbf{A}$ to another stimulus $\mathbf{B}$, we observe, following Mumford, that subjects in this case behave as if they take "$\mathbf{A}$ is similar to $\mathbf{B}$" to mean "$\mathbf{B}$ is some kind of prototype in a category which includes $\mathbf{A}$. Thus, the stimulus input $\mathbf{A}$ being analyzed is treated differently from the memory benchmark $\mathbf{B}$" (Mumford, 1991). To give $\mathbf{B}$ the required distinction, each feature $p_i(\mathbf{B})$ can be weighted in proportion to its long-term saliency $\mathsf{sal}(p_i, \mathbf{B})$ in distinguishing between $\mathbf{B}$ and the other stimuli. The resulting expression for similarity, which provides for the effects of context and for asymmetry, is

[2] This is unlike the case of the three-dimensional view space, parameterized by the Euler angles. However, it may still be possible to estimate the dimensionality of $\mathcal{S}$ by examining the neighbor structure of the reference points; see, e.g., (Tversky and Hutchinson, 1986).

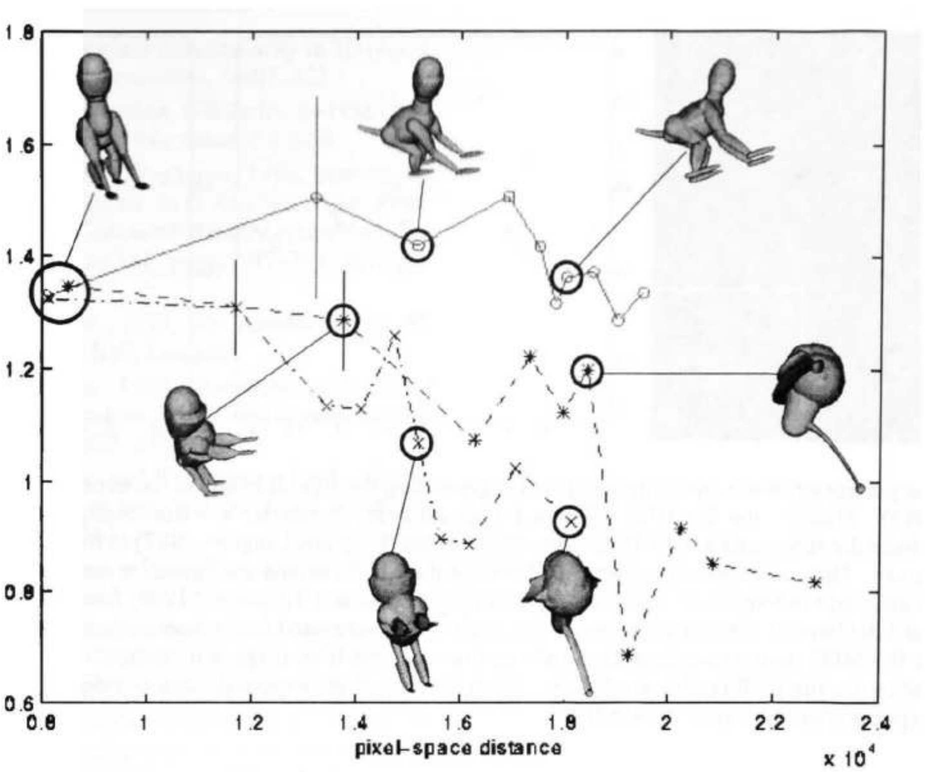

Figure 2: The response of a radial basis function module, trained on 10 random views of a parametrically defined object, to stimuli differing from a reference view of that object (marked by the big circle), in three ways: (1) by progressive view change, marked by ∘'s; (2) by progressive shape change, marked by ×'s; (3) by combined shape and view change, marked by $*$'s. The points along each curve have been sorted by pixel-space distance between the test and the reference stimuli (shown along the abscissa). Points are means over 10 repetitions with different random view-space and shape-space directions of change; a typical error bar (± standard error of the mean) is shown for each curve. Note the insensitivity of the module's output to view-space changes, relative to shape-space changes. Thus, the output can be interpreted as signalling the proximity of the stimulus to the view space of the reference object.

$$s(\mathbf{A},\mathbf{B}) \sim \sum_{i=1}^{k} w_i p_i(\mathbf{A}) \left(\frac{p_i(\mathbf{B})}{\mathsf{sal}(p_i,\mathbf{B})} \right) \qquad (2)$$

Note that this definition has the same form as the additive clustering (ADCLUS) similarity measure of (Shepard and Arabie, 1979), which, in turn, instantiates Tversky's (1977) discrete contrast model of feature-based similarity. At the same time, it is built on top of a continuous metric representational substrate – the shape space spanned by proximities to prototypes. The degree of compromise between these two approaches to similarity may depend on the demands of the task at hand, via the parameters of equation 2. At the one extreme, a Chorus-based system may behave as if it maps the stimuli pertaining to a task into a metric space, with the ensuing symmetric similarity and possible interaction among different dimensions; the other extreme may involve discrete all-or-none features, as in the examples surveyed by Tversky (1977).

Similarities to prototypes as a basis for veridical perception

The veridicality of representation of parametrically defined shapes in human subjects has been tested in two recent studies (Edelman, 1995a; Cutzu and Edelman, 1995). In each of a series of experiments, which involved pairwise similarity judgment and delayed matching to sample, subjects were confronted with several classes of computer-rendered 3D animal-like shapes, arranged in a complex pattern in a common parameter space. Response time and error rate data were combined into a measure of subjective shape similarity, and the resulting proximity matrix was submitted to non-metric multidimensional scaling (MDS; Shepard, 1980). In the resulting solution, the relative geometrical arrangement of the points corresponding to the different objects invariably reflected the complex low-dimensional structure in parameter space that defined the relationships between the stimuli classes (see Figure 3).

Computer simulations showed that the recovery of the low-dimensional structure from *image-space* distances between the stimuli was impossible, as expected. In comparison, the

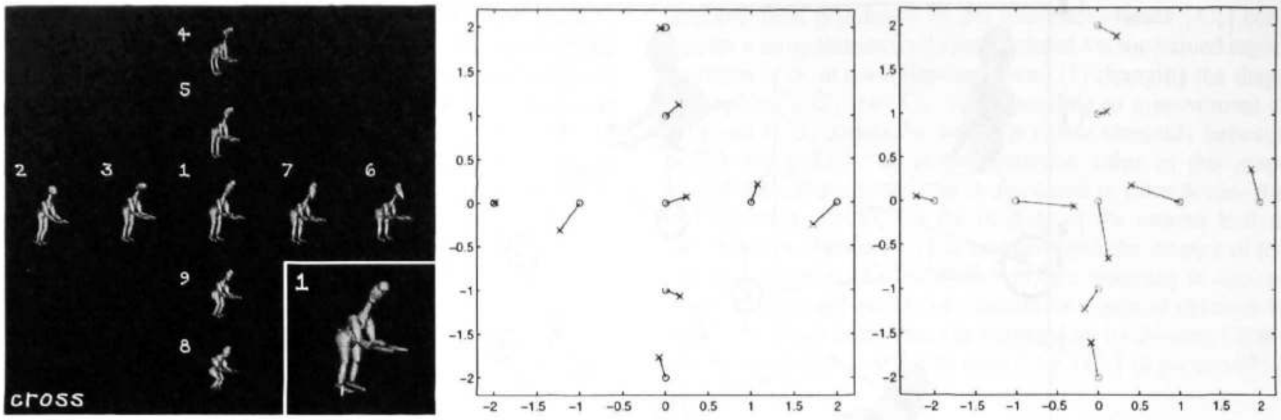

Figure 3: *Left:* the parameter-space configuration used for generating the stimuli in one of the experiments described in (Cutzu and Edelman, 1995). *Middle:* the 2D MDS solution for all subjects. Symbols: ○ – true configuration; × – configuration derived by MDS from the subject data, then Procrustes-transformed (Borg and Lingoes, 1987) to fit the true one. Lines connect corresponding points. The coefficient of congruence between the MDS-derived configuration and the true one was 0.99. In comparison, the expected random value, estimated by bootstrap (Efron and Tibshirani, 1993) from the data, was 0.86 ± 0.03 (mean and standard deviation); 100 permutations of the point order were used in the bootstrap computation. The Procrustes distance between the MDS-derived configuration and the true one was 0.66 (expected random value: 3.14 ± 0.15). *Right:* the 2D MDS solution for the RBF model; coefficient of congruence: 0.98 (expected random value: 0.86 ± 0.03); Procrustes distance: 1.11 (expected random value: 3.14 ± 0.17).

psychophysical results were fully replicated by a model patterned after a higher stage of object processing, in which nearly viewpoint-invariant representations of individual objects are available; a rough analogy is to the inferotemporal visual area IT; see, e.g., (Tanaka, 1993; Logothetis, Pauls, and Poggio, 1995). Such a representation of a 3D object can be easily formed, if several views of the object are available, by training a radial basis function (RBF) network to interpolate a characteristic function for the object in the space of all views of all objects (Poggio and Edelman, 1990). Following the Chorus approach, we chose a number of reference objects (in Figure 3, the corners of the parameter-space CROSS), and trained an RBF network to recognize each such object (i.e., to output a constant value for any of its views, encoded by the activities of the underlying receptive field layer). At the RBF level, the similarity between two stimuli was defined as the cosine of the angle between the vectors of outputs they evoked in the RBF modules trained on the reference objects (equation 1). The MDS-derived configurations obtained with this model showed significant resemblance to the true parameter-space configurations (see Figure 3, right).

Conclusion

Because the reference shapes can be considered complex features, Chorus effectively extends the notion of representation by feature detection from simple "primary" perceptual qualities such as color to all visual dimensions, including shape. This makes it possible to use multidimensional feature spaces in which different dimensions correspond to radically different qualities, not all of which need even be visual. Moreover, the system can maintain a high degree of plasticity, as new complex features can be learned by memorization, without paying for versatility by the need for dynamic binding, as in structural representation involving generic features.

The ensemble of feature detectors responds (J. J. Gibson would say, resonates) to the environment (while extracting task-specific information), without reconstructing it internally. By merely mirroring proximally the similarity structure of a distal shape space, Chorus embodies the ideas of those philosophers who argued that "meaning ain't in the head" (Putnam, 1988) and that "cognitive systems are largely in the world" (Millikan, 1995), circumvents the severe difficulties encountered by the reconstructionist approaches in computer vision, and may explain the impressive performance of biological visual systems, which, in any case, appear to be too sloppy to do a good job of reconstructing the world (O'Regan, 1992). Thus, in an important sense, Chorus lets the world be its own representation.

Acknowledgments

Thanks to Fiona Newell, Dan Roth, Josh Tenenbaum, Alice O'Toole, and Yair Weiss for useful comments. SE is an incumbent of the Sir Charles Clore Career Development Chair at the Weizmann Institute. FC is now with the Dept. of Cognitive and Linguistic Sciences, Brown University, Providence RI 02192.

References

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

Borg, I. and J. Lingoes. 1987. *Multidimensional Similarity Structure Analysis*. Springer, Berlin.

Cutzu, F. and S. Edelman. 1995. Explorations of shape space. CS-TR 95-01, Weizmann Institute of Science.

Duvdevani-Bar, S. and S. Edelman. 1995. On similarity to prototypes in 3D object representation. CS-TR 95-11, Weizmann Institute of Science.

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

Edelman, S. 1995b. Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines*, 5:45–68.

Edelman, S., D. Reisfeld, and Y. Yeshurun. 1992. Learning to recognize faces from examples. In G. Sandini, editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 787–791. Springer Verlag.

Efron, B. and R. Tibshirani. 1993. *An introduction to the bootstrap*. Chapman and Hall, London.

Lando, M. and S. Edelman. 1995. Receptive field spaces and class-based generalization from a single view in face recognition. *Network*, 6:551–576.

Logothetis, N. K., J. Pauls, and T. Poggio. 1995. Shape recognition in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563.

Millikan, R. 1995. *White Queen Psychology and other essays for Alice*. MIT Press, Cambridge, MA.

Mumford, D. 1991. Mathematical theories of shape: do they model perception? In *Geometric methods in computer vision*, volume 1570, pages 2–10, Bellingham, WA. SPIE.

Nosofsky, R. M. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:700–708.

O'Regan, J. K. 1992. Solving the real mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology*, 46:461–488.

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

Poggio, T., M. Fahle, and S. Edelman. 1992. Fast perceptual learning in visual hyperacuity. *Science*, 256:1018–1021.

Putnam, H. 1988. *Representation and reality*. MIT Press, Cambridge, MA.

Shepard, R. N. 1968. Cognitive psychology: A review of the book by U. Neisser. *Amer. J. Psychol.*, 81:285–289.

Shepard, R. N. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–397.

Shepard, R. N. and P. Arabie. 1979. Additive clustering: representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86:87–123.

Suppes, P., M. Pavel, and J. Falmagne. 1994. Representations and models in psychology. *Ann. Rev. Psychol.*, 45:517–544.

Tanaka, K. 1993. Neuronal mechanisms of object recognition. *Science*, 262:685–688.

Tversky, A. 1977. Features of similarity. *Psychological Review*, 84:327–352.

Tversky, A. and J. W. Hutchinson. 1986. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93:3–22.

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