[13] N. Intrator, J. I. Gold, H. H. Bülthoff, and S. Edelman. Three-dimensional object recognition using an unsupervised neural network: understanding the distinguishing features. In J. Moody and S. J. Hanson, editors, *Neural Information Processing Systems*, volume 4. Morgan Kaufmann, San Mateo, CA, 1992. to appear.

[14] D. G. Lowe. *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston, MA, 1986.

[15] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395, 1987.

[16] D. G. Lowe. Stabilized solution for 3D model parameters. In O. Faugeras, editor, *Proc. European Conference on Computer Vision*, pages 408–412, New York, 1990. Springer.

[17] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[18] J. L. Mundy and A. J. Heller. The evolution and testing of a model-based object recognition system. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 268–282. IEEE, Washington, DC, Osaka, 1990.

[19] T. Poggio. A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910, 1990.

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

[21] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

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

[23] H. Wechsler, editor. *Neural Networks in Perception.* Academic Press, New York, NY, 1991.

[24] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222:960–962, 1969.

and diffuse — can be detected by computational mechanisms resembling receptive fields, and can be considered, therefore, as a natural extension of a basis-function classification network [9,12].

A concrete example of the potential tradeoff between discrimination power and viewpoint invariance of a feature set is provided by recent experimental data [3] shown in Figure 1. The plot on the left suggests that human recognize 3D wire-like objects nearly independently of their image-plane orientation (but not of the orientation in depth). A similar behavior is exhibited by a view-interpolation model which includes lengths of segments between connected vertices in the object representations (in addition to the coordinates of individual vertices). This relative insensitivity to rotation in the image plane is expected to cause the model to be more prone to confuse objects that have similar projected segment lengths, but different 3D structure.

In summary, networks models of 3D object recognition based on interpolation among specific stored views behave in several respects similarly to human observers in a number of recognition tasks. Even closer replication of human performance in recognition should be expected, once the issue of the features used to represent object views is resolved.

References

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

[2] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64, 1992.

[3] H. H. Bülthoff, S. Edelman, and E. Sklar. Generalizing object recognition over 2d and 3d transformations. A. I. Memo, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1992. in preparation.

[4] P. Cavanagh. Local log polar frequency analysis in the striate cortex as a basis for size and orientation invariance. In D. Rose and V. G. Dobson, editors, *Models of the visual cortex*, pages 146–157. Wiley, New York, NY, 1985.

[5] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[6] S. Edelman. Features of recognition. CS-TR 10, Weizmann Institute of Science, 1991.

[7] S. Edelman and H. H. Bülthoff. Viewpoint-specific representations in 3D object recognition. A.I. Memo No. 1239, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[8] S. Edelman and H. H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 1991. submitted.

[9] S. Edelman and T. Poggio. Bringing the Grandmother back into the picture: a memory-based view of object recognition. A.I. Memo No. 1181, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990. to appear in Int. J. Pattern Recog. Artif. Intell.

[10] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3D objects. *Biological Cybernetics*, 64:209–219, 1991.

[11] D. Forsyth, J. L. Mundy, A. Zisserman, and C. M. Brown. Invariance - a new framework for vision. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 598–605, Osaka, 1990. IEEE, Washington, DC.

[12] J. E. Hummel and I. Biederman. Dynamic binding: a basis for the representation of shape by neural networks. In *Proc. 12th Annual Conference of the Cognitive Science Society*, pages 614–621, Hillsdale, NJ, 1990. Erlbaum.

corresponding to the familiar views. The approximation can be performed by a two-stage network (see Figure 2). In the first stage intermediate responses are formed by a collection of nonlinear "receptive fields" (shaped, e.g., as multidimensional Gaussians), centered at the familiar views. The output of the second stage is a linear combination of the intermediate receptive field responses. If the regularization network is trained to output the value 1 for various views of a given object, the decision criterion for recognition can be stated as

$$|\sum_k c_k G\left(\|X^{(2D)} - X_k^{(2D)}\|\right) - 1| < \theta \tag{1}$$

where $X^{(2D)}$ is the input image, represented as a vector of feature locations (e.g., vertices for a polyhedral object), $X_k^{(2D)}$ are the familiar or prototypical views stored in the system, c_k are the linear coefficients, and the function $G(\cdot)$ represents the shape of the receptive field. A recognition system based on this method performs well when the novel view is close to the stored ones, and progressively worse on views that are far from the familiar ones.

4 An open issue: features of recognition

In the preceding section we have surveyed a successful replication of central characteristics of human performance in 3D object recognition by a computational model based on interpolation among a number of stored views of each object. This success suggests that what is commonly regarded as the main problem in 3D recognition — the dependence of object appearance on the viewpoint — can be effectively reduced to a classical problem in pattern recognition, namely, classification in a multidimensional vector space [5]. Notably, basis function interpolation, which is one of the best methods for such classification, also happens to be equivalent to a class of three-layer networks that can be regarded as computing weighted sums of localized receptive fields and are, therefore, attractive from the standpoint of biological modeling [21,19].

A major issue that remains at this point unresolved is whether a recognition scheme based on interpolation among specific views is capable of the degree of viewpoint-invariant recognition exhibited by human subjects under certain circumstances (especially in tasks requiring basic-level classification, rather than the identification of individual objects; see [1]). Evidence of viewpoint-invariant recognition has served in the past as an argument against multiple-view representation of objects, and has presented a serious challenge to the developers of network models of recognition. We propose that such evidence can be accommodated within the framework of multiple-view representation, by allowing for an appropriate encoding of the stored views. In other words, we propose to capture the varying degree of viewpoint invariance found in human recognition performance by endowing the model with an extensive repertoire of feature detectors, whose output (and not the raw input image) is fed into the basis-function classifier [6].

Those of the detected features that are well-localized in the image (e.g., polyhedral vertices, as mentioned in the preceding section; see also [13] would allow fine distinctions among objects, at the expense of relatively strong sensitivity to viewpoint (the location of a corner in the projected image is highly dependent on the object's attitude with respect to the observer). On the other hand, the so-called non-accidental features [14,1] offer relative insensitivity to viewpoint, at the expense of reduced power of discrimination among objects. An example of such a feature is the presence of near-parallel lines in the image, which is highly unlikely to be caused by an accident of a viewpoint, but at the same time only allows to discriminate between objects that possess such parallel lines and those that do not. Finally, "diffuse" features such as surface color or texture may support recognition performance that is basically viewpoint-invariant and is exact to the extent that the surface markings are distinctive for each object under consideration. It is important to note that all three kinds of features — localized, non-accidental,

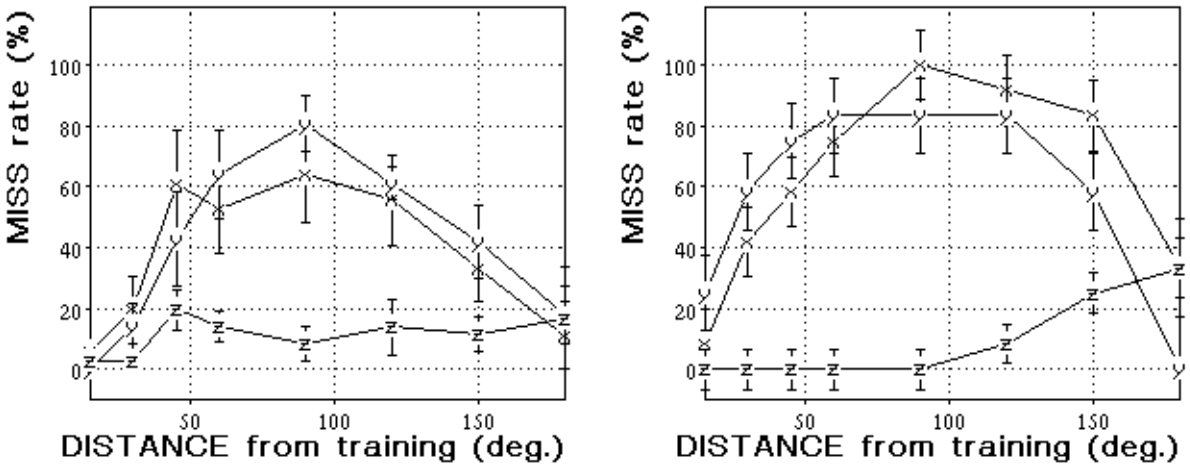

Figure 1: *Left:* human performance in the recognition of rotated wire-like 3D objects [2]. Error rate of subjects trained on single view is plotted vs. distance between training and test views. Note poor generalization across rotations in depth (curves marked by x, y), compared to rotation in the image plane (curve marked by z; see text). *Right:* performance of the HyperBF network model [20,9,2] in the same task.

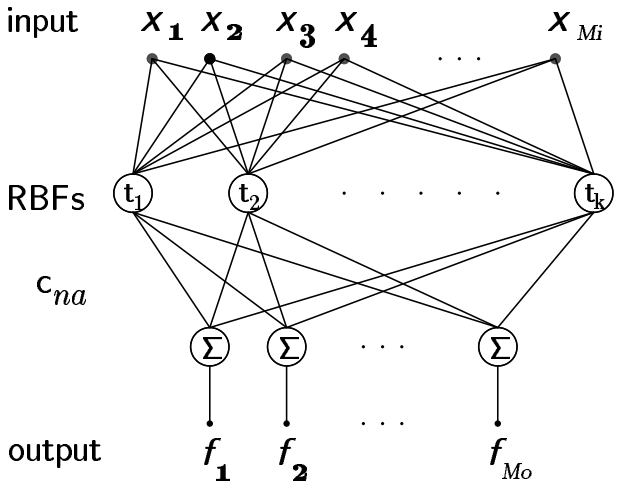

Figure 2: A generic HyperBF network. The input to the network is a vector of real numbers. The output is a weighted sum of real-valued basis functions of limited support. In the context of vision a biologically plausible implementation of the network is possible if the responses of the intermediate-layer units are calculated by multiplying activities of retinal receptive fields [20,10].

must be discounted by the recognition process. An example of such an approach, based roughly on the known architectural properties of the early stages of the human visual system, calls for preprocessing that combines complex logarithmic and Fourier transformations to achieve recognition that is invariant to scaling and to shifts and rotations in the image plane [4]. Obviously, this approach does not guarantee invariance with respect to rotations in depth. The search for appropriately invariant descriptions of shape is currently the subject of intensive research (e.g., [11]). The resulting invariants, so far, do not seem to be especially amenable to network implementation, and are not claimed to be biologically plausible.

2 Computational theory of recognition

Setting aside temporarily architectural constraints and the requirement of biological realism allows one to consider the problem of recognition on an abstract or computational level. Computationally, recognition calls for a comparison between the input and set of internal representations (models). Because of viewpoint dependency of object appearance, such a comparison cannot be made unless the transformation associated with the particular viewpoint is compensated for. It turns out that one need not know the identity of the object beforehand to compensate for the viewpoint transformation. Specifically, one can match a small number of key features in the image and in a model, compute and carry out the transformation implied by their correspondence, and verify the match by assessing the similarity between two representations. If the input image indeed belongs to that particular model, similarity will be high, and the image may be considered recognized. In principle, this process, referred to as recognition by alignment [22], must be carried out, possibly in parallel, for all known object models. This and related approaches [15,18,16] proved successful in the recognition of a variety of object classes under realistic conditions.

3 Psychophysical support for a view-interpolation model

The development of network models of recognition in the human visual system — our main interest in the present survey — was thus faced with a dilemma. On one hand, powerful computational approaches to recognition became available that could in principle support reliable and viewpoint-independent performance under a variety of conditions. On the other hand, such performance could not be achieved with network models, whose reliance on memory of specific images leads, in the absence of suitable invariants, to inherently viewpoint-dependent performance.

Fortunately, a series of psychophysical experiments have recently shown that the human performance in recognition is limited just as it would be if its underlying computational mechanism were memory lookup. The main findings were as follows (see [7,8,2] for details). First, when subjects had to recognize previously seen views of objects that appeared at arbitrary 3D orientations, some of the views yielded shorter response times and lower error rates than others. This happened even when each view was shown for the same number of times during training. Second, generalization to novel views was severely limited, with performance dropping to chance level already at misorientations of about 40° relative to familiar views (see Figure 1). Third, adding binocular disparity to provide the subjects with an additional and reliable cue to the third dimension reduced the mean error rate, but did not lead to near-invariant performance. Importantly, the availability of depth information did not change the basic feature of generalization to novel views, namely, the increase in the error rate with misorientation relative to a familiar view.

These performance characteristics were replicated by a model of recognition that employed representation by multiple specific views and implemented in a regularization network architecture [21]. Under this approach, which includes as a special case approximation by radial basis functions (RBFs), generalization from familiar to novel views is regarded as a problem of approximating a smooth hypersurface in the space of all possible views, with the "height" of the surface known only at a sparse set of points

Modeling human visual object recognition

Shimon Edelman **Heinrich H. Bülthoff**

Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100
Israel
edelman@wisdom.weizmann.ac.il

and

Dept. of Cognitive and Linguistic Sciences
Brown University
Providence, RI 02192
USA
hhb@holmes.cog.brown.edu

Abstract

Networks models of three-dimensional object recognition based on interpolation among specific stored views behave in several respects similarly to human observers in a number of recognition tasks. We expect that the resolution of the issue of features to be used to represent object views should lead to an even closer replication of human performance in recognition.

1 Network models of object recognition

Recognizing three-dimensional objects from single two-dimensional images is difficult, because the appearance of an object may vary widely, depending on the illumination conditions, and on the relative position of the object and the imaging system (eye or camera). Consequently, artificial neural network models of object recognition are successful insofar as they are able to generalize across changes in illumination and viewpoint, or, alternatively, to use object representations that are invariant to such changes. Currently, the majority of networks that perform recognition are designed to handle limited classes of inputs (e.g., handwritten characters), and their ability to handle transformed inputs is limited to the Euclidean group in the image plane. In comparison, surprisingly few network models that attempt to recognize 3D objects when these are allowed to undergo rotation in depth can be found in the literature.

Some of the reasons for the scarcity of network models of 3D recognition have to do with the historical distribution of research effort across the different network computation paradigms[1]. For example, the study of the capabilities of perceptrons [17], initially expected to lead to the development of universal recognition devices, became sidetracked, as far as 3D recognition is considered, into issues such as image connectivity, which is not directly relevant to recognition. Network models of associative memory, which were developed at about the same time [24], were widely regarded as inadequate for carrying out "real" object recognition, because of a claim that the achievement of acceptable performance irrespective of the object's attitude would require storing an inordinate number of the object's views (see a discussion in [22]).

An alternative to a brute-force memory-based approach is to precede associative lookup by a preprocessing stage, in which the input is given a description that is invariant to the transformations that

[1] References to previous work on network models of recognition are given to illustrate specific points, and do not constitute a review of the field. For a representative collection of articles on networks in the context of perception see, e.g., [23]