Reprinted from

Current Opinion in NEUROBIOLOGY

Volume 1, 1991

CB
CURRENT
BIOLOGY

Models of object recognition

Shimon Edelman and Tomaso Poggio

The Weizmann Institute of Science, Rehovot, Israel and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Progress in the understanding of visual recognition in the past year has been signified by the demonstration of computational feasibility of and psychophysical support for two-dimensional view-interpolation methods.

Current Opinion in Neurobiology 1991, **1**:270–273

Introduction

Much of the visual activity of a mobile intelligent being endowed with sight may be classified as recognition. Loosely speaking, such diverse uses of vision as the identification of prey or predator, navigation, and non-verbal social function all require that the visual system be capable of object recognition. Thus recognition is a rather broad topic to which it would be difficult to do justice within the limitations of a short review.

The present discussion is therefore largely confined to model-based subordinate-level [1] recognition of isolated objects. It will not deal with the different problem of categorization, that is recognizing a chair as such, even if that particular chair has never been seen before. Model-based recognition means that information derived from the image of an object to be recognized is compared with a set of internal models that represent objects known to the system. This definition stresses the issue of internal representation, that is common to vision and other cognitive faculties [2]. As we shall see, representation has indeed been the central issue in the recent developments in the field of recognition. In particular, the nature of representation in human vision appears to differ enough in subordinate-level, as compared with basic-level, recognition to warrant a separate treatment of the two cases.

Recognition in computer vision

To succeed in recognition, a system (natural or artificial) must cope with two major sources of variability in object appearance. The first problem is the variability due to differing light conditions. Proposed solutions to this problem in computer vision (e.g. the use of intensity and other discontinuities as a base representation [3]), as well as accounts of human recognition performance under varying light conditions, are outside the scope of this article. The second major problem in recognition — changes in object appearance due to a potentially infinite variety of positions relative to the viewer — has in recent years been found to have a satisfactory solution.

Recall that model-based recognition calls for a comparison between the input and the 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. Surprisingly, it turns out that one need not know the identity of the object to compensate for the viewpoint transformation. Specifically, one can match a small number of key features between the image and a model, compute and carry out the transformation implied by their correspondence, and verify the match by assessing the similarity between the 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 [4], must be carried out (possibly in parallel) for all known object models.

Although alignment and related approaches [4–6,7••] have proved successful in the recognition of a variety of object classes under realistic conditions, they all share the significant limitation of a dependence on externally supplied three-dimensional representations of known objects, without which viewpoint transformations cannot be properly compensated for. This handicap affects the suitability of approaches based on three-dimensional representations both for machine vision and for the modeling of recognition in human vision, as true three-dimensional reconstruction of the world is computationally difficult and is probably not implemented fully in the human visual system (for a review, see [8]).

Model-based recognition without three-dimensional models

The past year has seen developments in the theory of object recognition that have effectively disposed with the

Abbreviation

HyperBF—hyper basis function.

ISSN 0959-4388

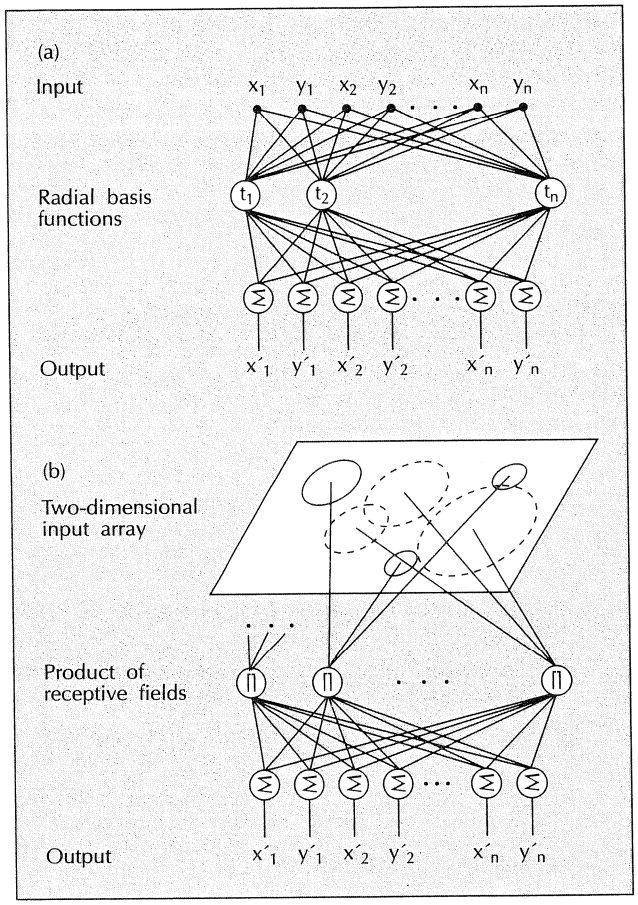

Fig. 1. (a) A schematic illustration of the object recognition module implemented as a Hyper basis function (HyperBF) network which is capable of recognizing three-dimensional objects from two-dimensional views. Each basis unit in the middle layer computes the distance of the input view from a fixed standard view, which is represented by the centre of the basis unit, and applies to it the weighted distance function. The resulting value can be regarded as the activity of the unit. A basis unit attains maximum activity when the input view exactly matches the prototypical view. The output of the module is a linear superposition of the activities of all the basis units in the network; $x_n y_n$ are the components of the vector (t) associated with a two-dimensional view of the object, $x'_n y'_n$ represent the output 'standard view'. Σ, summation. (b) Shows an equivalent interpretation of (a) for the case of Gaussian basis functions. A multidimensional Gaussian can be synthesized as the product (Π) of two-dimensional Gaussian receptive fields operating on retinotopic maps of features. The solid circles in the image plane represent the two-dimensional Gaussians associated with the first radial basis function (RBF), which corresponds to the first view of the object. The broken circles represent the two-dimensional receptive fields that synthesize the Gaussian radial function associated with another view. The receptive fields transduce positions of features, represented implicitly as activity in a retinotopic array, and their product computes the basis function without the need of calculating norms and exponentials explicitly.

need for three-dimensional representations. In one such development, Ullman and Basri [9••] have shown that under orthographic projection any view of a three-dimensional object can be represented (up to a multiplicative factor) in a point-by-point fashion as a linear combination of at most six fixed views of the same objects. Furthermore, it has been shown that a few views of an object are sufficient to synthesize a linear operator for the recognition of that object from a single two-dimensional image.

Another recently developed method [10••], similarly capable of learning to recognize three-dimensional objects from two-dimensional views, is based on the observation that the problem of recognition can be formulated in terms of approximation of a smooth mapping (e.g. the mapping from any viewpoint of an object to a fixed 'standard' view). This method uses an extension of the results from the function approximation theory, namely, approximation with Hyper basis functions (or HyperBFs) [11,12], and can be implemented as a network (see Fig. 1) [13] that learns to recognize three-dimensional objects given a sample set of their two-dimensional views. This approach is compatible with perspective projection and in addition can use a variety of representational primitives to image coordinates of object features. The performance of the HyperBF network on both synthetic and real images of three-dimensional objects is encouraging [13].

Experimental evidence in favor of the view-interpolation theory of recognition

It is now becoming increasingly clear that the HyperBF network is a viable model of at least one pathway to object recognition in human vision. Psychophysical evidence to that effect is provided by two different classes of data. The classes correspond to the two major phenomena in the psychology of subordinate-level recognition: first, the existence of canonical views [14] and their development with practice (notably, the disappearance of orderly dependence of recognition time on misorientation relative to a canonical view) [15,16]; and second, limited generalization to novel views, even in the presence of full three-dimensional information on the structure of the stimulus [17,18••].

Support of a different kind for the notion of viewpoint-specific representation has been provided by simulated 'psychophysical' experiments, in which computer implementations of the HyperBF and a related model [19], have been presented with the same computer-generated stimuli seen by the human subjects. In these experiments, the models replicated key characteristics of human performance mentioned above (for a review, see [18••]).

Yet another source of support for the idea that objects are represented in the brain by sets of predominantly two-dimensional views can be found in single-cell recordings from the inferotemporal cortex (the target area of the major shape-processing stream in the brain [20]) in the monkey [21,22]. The discovery in that area of cells that respond preferentially to specific aspects (see [23]) of human faces has long been quoted by the proponents of the 'grandmother cell' doctrine [24] in support of their theory. This view seems to have gained plausibility recently following consistent replication of 'face cell' findings (for a review, see [25]), the emergence of evidence

that face cells 'learn' by modifying their response profile with experience [26], and the demonstration of the role of single cells in mediating perceptual behavior [27]. On the theoretical side, this is paralleled by the development of view-interpolation models, which can be interpreted as a computational rationale for the grandmother cell theory [10••,13].

Modeling the variety of recognition

Viewpoint-dependent performance is exhibited by just one of the many pathways to recognition apparently active within the human visual system. The development of a comprehensive model of recognition that would offer a unified account of all psychophysical data depends on a better understanding of these pathways and their interactions. Apart from the general observation that basic-level but not subordinate-level recognition tends to be viewpoint-invariant [28], not much is known about the details of the conditions under which different pathways are dominant (a similar situation is found in the study of another visual system that excels in recognition, that of the pigeon [29,30]). One relevant result states that viewpoint-dependent recognition prevails when the objects differ along more than one spatial dimension [31••]. Other characterizations of viewpoint dependence of recognition are related to details of object presentation [32]. Finally, basic-level recognition of line drawings of common objects seems to be consistently invariant to viewpoint, as well as to translation and reflection [33,34]. A somewhat involved model of this aspect of recognition is described by Hummel and Biederman [35]. It should be noted that this model, based on a theory known as recognition by components (see [36]), glosses over important issues in low-level vision and has not yet been applied to the real-world stimuli.

Perspectives

The operation of the viewpoint-dependent pathway to recognition is currently the subject of intensive research. In particular, the generalization capability of this pathway, to transformations other than rotation, has recently been investigated. It was found that the factor most closely correlated with recognition error rate was the amount of two-dimensional (image) shape change, even when the stimulus was transformed non-rigidly (i.e. when it was deformed), as predicted by the HyperBF model of recognition [37]. As to the dependence of recognition on the position of the stimulus within the visual field, the findings range from very limited invariance for dot patterns [38] to translationally-invariant but viewpoint-dependent performance for pictures of real objects [32]. More research is clearly required to characterize the phenomena of position and size [39] invariance in recognition.

One possible approach to a computational understanding of these phenomena is through the integration of existing models of shape processing in early vision [40,41] with the HyperBF techniques mentioned above. Any attempt at such synthesis would have to address what seems to be the major remaining problem in subordinate-level recognition, namely, phenomenological and computational characterization of the basic features that are the primitives of form representation.

References and recommended reading

Papers of special interest, published within the annual period of review, have been highlighted as:

- • of interest
- •• of outstanding interest

1. ROSCH E, MERVIS CB, GRAY WD, JOHNSON DM, BOYES-BRAEM P: **Basic Objects in Natural Categories.** *Cogn Psychol* 1976, 8:382–439.

2. MARR D: *Vision* [book]. San Francisco: WH Freeman 1982.

3. POGGIO T, GAMBLE EB, LITTLE JJ: **Parallel Integration of Vision Modules.** *Science* 1988, 242:436–440.

4. ULLMAN S: **Aligning Pictorial Descriptions: an Approach to Object Recognition.** *Cognition* 1989, 32:193–254.

5. LOWE DG: **Three-Dimensional Object Recognition from Single Two-Dimensional Images.** *Artif Intell* 1987, 31:355–395.

6. THOMPSON DW, MUNDY JL: **Three-Dimensional Model Matching from an Unconstrained Viewpoint.** In *Proceedings of IEEE Conference on Robotics and Automation.* Raleigh, North Carolina: 1987, pp 208–220.

7. •• LOWE DG: **Stabilized Solution for 3D Model Parameters.** In *Proceedings of European Conference on Computer Vision,* edited by Faugeras O. New York: Springer 1990, pp 408–412.
A good example of a computer vision approach to the task of model-based object recognition.

8. EDELMAN S, WEINSHALL D: **Computational Vision: a Critical Review.** In *Vision and Visual Dysfunction* [book], edited by Watt R. Boca Raton, Florida: CRC Press 1991, volume 14, chapter 4.

9. •• ULLMAN S, BASRI R: **Recognition by Linear Combinations of Models.** *A.I. Memo No. 1152.* Cambridge: Artificial Intelligence Laboratory, Massachusetts Institute of Technology 1990.
A surprising computational result regarding the geometry of imaging. Under orthographic projection, an appropriately defined two-dimensional view of a three-dimensional object is the linear combination of two views of the same object. This observation provides insights into the complexity of shape-based and model-based object recognition. It also suggests a specific algorithm for three-dimensional recognition.

10. •• POGGIO T, EDELMAN S: **A Network that Learns to Recognize Three-Dimensional Objects.** *Nature* 1990, 343:263–266.
The original paper describing the application of the HyperBF technique to object recognition.

11. BROOMHEAD DS, LOWE D: **Multivariable Functional Interpolation and Adaptive Networks.** *Complex Syst* 1988, 2:321–355.

12. POGGIO T, GIROSI F: **Regularization Algorithms for Learning that Are Equivalent to Multilayer Networks.** *Science* 1990, 247:978–982.

13. EDELMAN S, POGGIO T: **Bringing the Grandmother Back into the Picture: a Memory-Based View of Object Recognition.** *A.I. Memo No.1181.* Cambridge: Artificial Intelligence Laboratory, Massachusetts Institute of Technology 1990.

14. PALMER SE, ROSCH E, CHASE P: **Canonical Perspective and the Perception of Objects.** In *Attention and Performance*

[book], edited by Long J, Baddeley A. Hillsdale, New Jersey: Erlbaum 1981, pp 135–151.

15. TARR M, PINKER S: **Mental Rotation and Orientation-Dependence in Shape Recognition.** *Cogn Psychol* 1989, 21:233–?82.

16. EDELMAN S, BULTHOFF H, WEINSHALL D: **Stimulus Familiarity Determines Recognition Strategy for Novel 3D Objects.** *A.I. Memo No.1138.* Cambridge: Artificial Intelligence Laboratory, Massachusetts Institute of Technology 1989.

17. ROCK I, DIVITA J: **A Case of Viewer-Centered Object Perception.** *Cogn Psychol* 1987, 19:280–293.

18. •• EDELMAN S, BULTHOFF HH: **Viewpoint-Specific Representations in 3D Object Recognition.** *A.I. Memo No. 1239.* Cambridge: Artificial Intelligence Laboratory, Massachusetts Institute of Technology 1990.

Preliminary results on the psychophysics of object recognition support schemes based on a set of two-dimensional views rather than three-dimensional models. These appear to be more directly consistent with interpolation schemes such as HyperBF, rather than with strict versions of linear combination schemes.

19. EDELMAN S, WEINSHALL D: **A Self-Organizing Multiple-View Representation of 3D Objects.** *Biol Cybern* 1991, 64:209–219.

20. MISHKIN M, UNGERLEIDER LG, MACKO KA: **Object Vision and Spatial Vision: Two Cortical Pathways.** *Trends Neurosci* 1983, 4:414–417.

21. GROSS CG, ROCHA-MIRANDA CE, BENDER DB: **Visual Properties of Cells in Inferotemporal Cortex of the Macaque.** *J Neurophysiol* 1972, 35:96–111.

22. PERRETT DI, ROLLS ET, CAAN W: **Visual Neurones Responsive to Faces in the Monkey Temporal Cortex.** *Exp Brain Res* 1982, 47:329–342.

23. SEIBERT M, WAXMAN AM: **Learning Aspect Graph Representations from View Sequences.** In *Natural Information Processing Systems* [book], edited by Touretzky D. San Mateo: Morgan Kaufmann 1990, volume 2, pp 258–265.

24. BARLOW HB: **The Role of Single Neurons in the Psychology of Perception.** *Q J Exp Psychol* [A] 1985, 37:121–145.

25. PERRETT DI, MISTLIN AJ, CHITTY AJ: **Visual Neurones Responsive to Faces.** *Trends Neurosci* 1989, 10:358–364.

26. ROLLS ET, BAYLIS GC, HASSELMO ME, NALWA V: **The Effect of Learning on the Face Selective Responses of Neurons in the Cortex in the Superior Temporal Sulcus of the Monkey.** *Exp Brain Res* 1989, 76:153–164.

27. SALZMAN CD, BRITTEN KH, NEWSOME WT: **Cortical Microstimulation Influences Perceptual Judgements of Motion Direction.** *Nature* 1990, 346:174–177.

28. EDELMAN S: **A Network Model of Object Recognition in Human Vision.** In *Networks for Vision* [book], edited by Wechsler H. New York: Academic Press 1991.

29. LOMBARDI CM, DELIUS JD: **Size Invariance in Visual Pattern Recognition by Pigeons.** In *Quantitative Analyses of Behavior* [book], edited by Commons ML, Herrnstein RJ, Kosslyn SM, Mumford DB. Hillsdale, New Jersey: Erlbaum 1990, volume VIII.

30. CERELLA J: **Pigeon Pattern Perception: Limits on Perspective Invariance.** *Perception* 1990, 19:141–159.

31. •• TARR M, PINKER S: **When Does Human Object Recognition Use a Viewer-Centered Reference Frame?** *Psychol Sci* 1990, 1:253–256.

This paper is one of the most recent attempts to address the question of mental rotation in the context of object recognition.

32. ELLIS R, ALLPORT DA, HUMPHREYS GW, COLLIS J: **Varieties of Object Constancy.** *Q J Exp Psychol* [A] 1989, 41:775–796.

33. BIEDERMAN I: **Human Image Understanding: Recent Research and a Theory.** *Comput Vis Graph Image Processing* 1985, 32:29–73.

34. BIEDERMAN I, COOPER EE: **Evidence for Complete Translational and Reflectional Invariance in Visual Object Priming.** *Perception* 1991, 20: in press.

35. HUMMEL JE, BIEDERMAN I: **Dynamic Binding: a Basis for the Representation of Shape by Neural Networks.** In *Proceedings of 12th Annual Conference of the Cognitive Science Society.* Hillsdale, New Jersey: Erlbaum 1990, pp 614–621.

36. BIEDERMAN I: **Recognition by Components: a Theory of Human Image Understanding.** *Psychol Rev* 1987, 94:115–147.

37. EDELMAN S, BULTHOFF HH: **Generalization of Object Recognition in Human Vision Across Stimulus Transformations and Deformations.** In *Proceedings of 7th Israeli AICV Conference,* edited by Feldman Y, Bruckstein A. Elsevier 1990, pp 479–487.

38. NAZIR T, O'REGAN JK: **Some Results on Translation Invariance in the Human Visual System.** *Spatial Vis* 1990, 5:81–100.

39. LARSEN A: **Pattern Matching: Effects of Size Ratio, Angular Difference in Orientation and Familiarity.** *Percept Psychophys* 1985, 38:63–68.

40. CAVANAGH P: **Local Log Polar Frequency Analysis in the Striate Cortex as a Basis for Size and Orientation Invariance.** In *Models of the Visual Cortex* [book], edited by Rose D, Dobson VG. New York: Wiley 1985, pp 146–157.

41. MALLOT HA, VON SEELEN W, GIANNAKOPOULOS F: **Neural Mapping and Space-Variant Image Processing.** *Neural Networks* 1990, 3:

S Edelman, Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel.

T Poggio, Department of Brain and Cognitive Science, Room E 25–201, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.