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

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

[16] H. B. Barlow. Cerebral cortex as model builder. In D. Rose and V. G. Dobson, editors, *Models of the visual cortex*, pages 37–46. Wiley, New York, 1985.

[17] H. B. Barlow. The role of single neurons in the psychology of perception. *Quart. J. Exp. Psychol.*, 37A:121–145, 1985.

[18] W. Pitts and W. S. McCulloch. How we know universals: the perception of auditory and visual forms. In *Embodiments of mind*, pages 46–66. MIT Press, Cambridge, MA, 1965.

[19] S. Edelman and H. H. Bülthoff. Generalization of object recognition in human vision across stimulus transformations and deformations. In Y. Feldman and A. Bruckstein, editors, *Proc. 7th Israeli AICV Conference*, pages 479–487. Elsevier, 1990.

[20] S. Edelman. Features of recognition. In *Proc. Intl. Workshop on Visual Form, Capri, Italy*, New York, 1991. Plenum Press.

[21] 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.

theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

[3] I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293, 1987.

[4] M. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282, 1989.

[5] S. Edelman, H. Bülthoff, and D. Weinshall. Stimulus familiarity determines recognition strategy for novel 3D objects. A.I. Memo No. 1138, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, July 1989.

[6] 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.

[7] S. E. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In J. Long and A. Baddeley, editors, *Attention and Performance IX*, pages 135–151. Erlbaum, Hillsdale, NJ, 1981.

[8] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

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

[10] I. Rock, D. Wheeler, and L. Tudor. Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21:185–210, 1989.

[11] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2D interpolation theory of object recognition, 1990. submitted.

[12] R. N. Shepard and L. A. Cooper. *Mental images and their transformations*. MIT Press, Cambridge, MA, 1982.

[13] D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

4 Discussion

The CLF approach to the modeling of recognition, described in this article, has three key characteristics. A version of the first one, which states that conjunctions of features are important, has been advocated in the past by Barlow, who also has been a long-time proponent of the grandmother cell dogma [16,17]. The second ingredient of CLF, namely, the achievement of constancy over a group of transformations (such as rotations in 3D) through exhaustive coverage of the resulting configuration space, can be traced back to Pitts and McCulloch [18]. Both these ideas used to draw criticism on the grounds of excessive memory requirements. It is not too surprising, therefore, that the third key ingredient of CLF, blurred template matching, is designed to permit it to store relatively few views, while maintaining adequate generalization performance. Poggio and Girosi have recently used techniques from approximation theory to show why such an approach works [14].

To date, the CLF model has achieved a degree of success in replicating several basic findings in the psychology of three-dimensional object recognition (see [9,11,19,6]). Some of the issues currently under investigation are modeling of the influence of depth cues, and extension to recognition and classification of complex objects on various categorical levels. Already at this stage, however, the available simulation results prompt one to consider seriously the possibility that a major recognition pathway in human vision relies on two-dimensional view-specific representations (see also [20]). The amenability of the model to implementation in an adaptive network architecture which complies with general rules of cortical organization (see the discussion in [9,21]) lends further support to this possibility.

References

[1] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[2] I. Biederman. Human image understanding: Recent research and a

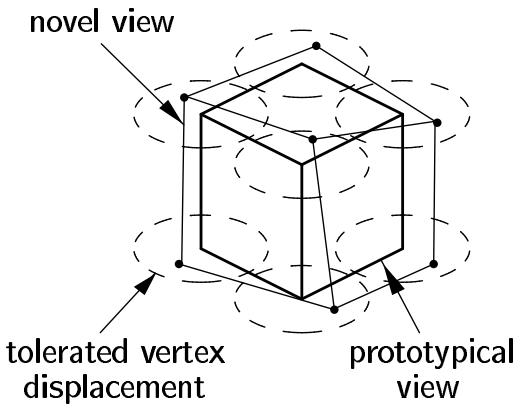

Figure 7: Generalization capability of the CLF model may be attributed to the fact that it tolerates feature displacements caused, e.g., by the object's rotation because of the diffuse projections from the feature to the representation layer. The anisotropy of generalization can be accounted for by asymmetries in the shape of the tolerance regions centered at the average positions of features. Here the tolerance regions are elongated in the horizontal direction, to replicate the better generalization in the horizontal plane found in human data.

when the correlation between its representation and the proper footprint is computed. Thus, test views that are close to more than one familiar view are easier to recognize, because of superposition of the contributions of the feature detectors corresponding to those views, achieved by blurring the input. This may explain why human subjects are better at interpolation than at extrapolation to a novel view. Furthermore, the anisotropy of generalization with respect to the horizontal/vertical distinction may be accounted for by postulating an asymmetrical point spread function (see Figure 7). An analysis of the generalization capability of the CLF model, along with a discussion of functional similarities between blurred template matching and nonlinear interpolation by regularization networks [14,15], can be found in [9].

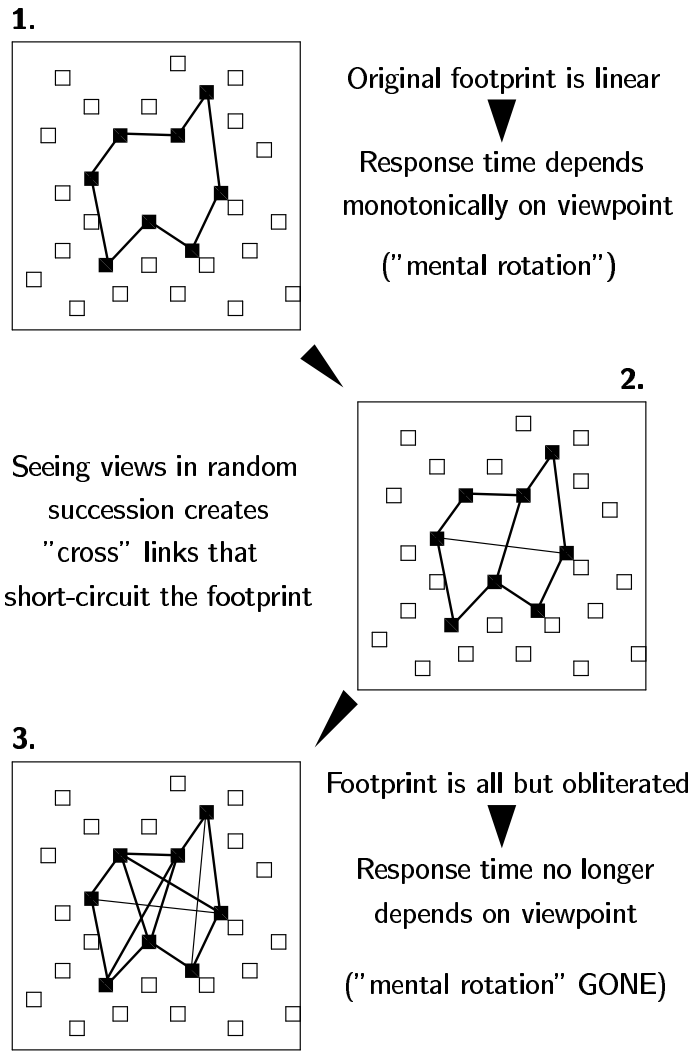

Figure 6: Three stages in the development of object representation with practice, as implemented in the CLF model. *Top:* Immediately after training with an orderly sequence of views that arises, e.g., when the object is rotating, its footprint (see Figure 4) is highly structured. The sequential spread of activation through the footprint following exposure to one view creates a semblance of mental rotation. *Middle:* Two shortcuts across the footprint are created, e.g., because of practice-induced association between non-neighbor views. *Bottom:* The dependency of "response time" on viewpoint is lost due to the weakening of the original footprint structure.

two patterns of activity, measured by their (2D) correlation, is then interpreted as the model's analog of response time (see Figure 5). The variation of this measure with viewpoint (that is, with the initial locus of activation) is the counterpart of the canonical views phenomenon.

3.4 Replicating Mental Rotation and Its Disappearance with Practice

The simulated response time not only varies with viewpoint: because of the sequential structure of the footprint, it depends on the viewpoint in an orderly fashion, resembling the typical pattern of mental rotation (see Figure 6, top). When the same views on which the model has been trained appear in a different order, the original sequential structure of the footprint is weakened, because of the emergence of new lateral links between different R-units that are not necessarily adjacent to each other in the footprint (see Figure 6, middle). Eventually, the interconnection pattern of the participating R-units becomes amorphous, causing mental rotation signs, which are epiphenomenal to the structure of the footprint, to disappear (see Figure 6, bottom).

3.5 Replicating Limited Anisotropic Generalization

The generalization capability of the CLF model is explored by training it on sets of views of several objects, presented separately in succession. Quite understandably, the model performs perfectly when tested on any of the training views, provided that the footprints do not overlap (that is, if the representation capacity is not exceeded). Even in that simple situation, the model yields a useful insight into possible mechanisms of generalization of recognition.

Generalization in the CLF model is made possible by the bell shape of the point spread function that governs the pattern of projection from first-layer to second-layer units (see section 3.2). Intuitively, the blurring of the input activity distribution caused by the point spread function increases the chances that moderate distortion of the input view (due, e.g., to a rotation of the object away from a training attitude) will be tolerated

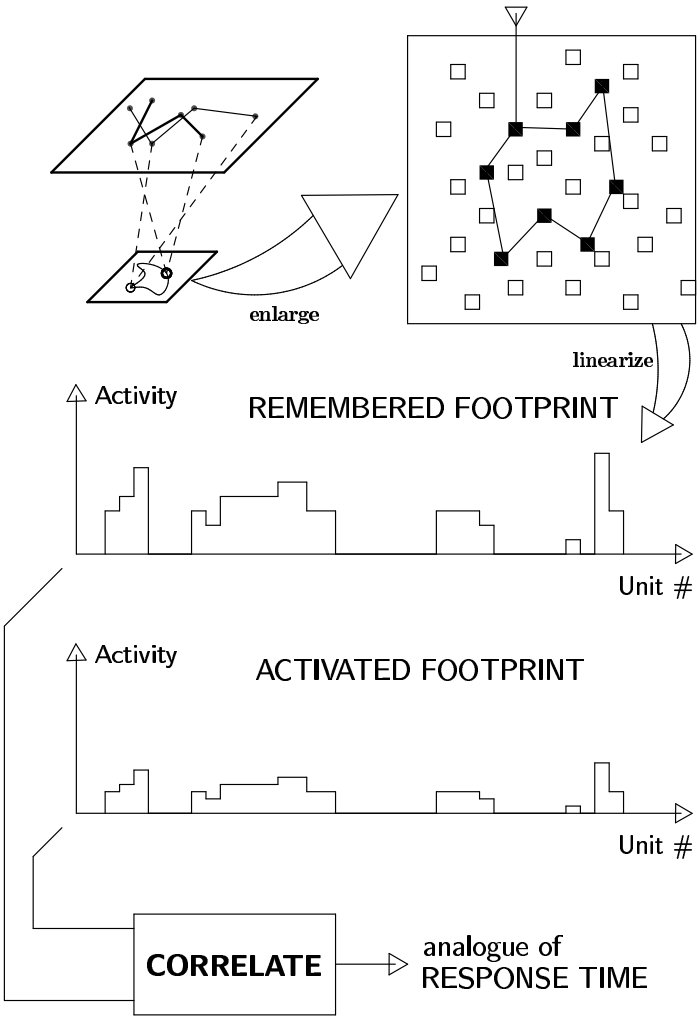

Figure 5: *Top:* a schematic illustration of a footprint. Solid and hollow squares stand for allocated and free R-units, respectively. *Bottom:* R-units comprising a footprint become active when the system is exposed to one of the familiar views, at a level that decreases with their distance along the footprint from the excitation point (because of the imperfect efficacy of the lateral links). The gradation in the spread of activity causes the correlation between actual and stored footprint snapshots to be less than ideal. The dependency of this correlation on the excitation point parallels a similar dependency of response time on viewpoint, known as mental rotation.

together more efficiently to form representations of entire objects.

3.2 Learning Object Representations

The CLF model acquires the representation of a novel object as follows. The very first view of the object is allocated a representation unit in the second layer through projection convergence, followed by non-maximum suppression. First, each input unit projects activation to an area in the second layer defined by a bell-shaped point spread function, with many inputs converging on the same representation unit. Next, a non-maximum suppression or winner-take-all mechanism selects the most active representation unit and allocates it. Once an R-unit is allocated, its input weights and threshold are adjusted according to a Hebbian rule to ensure future selectivity to the view it encodes.

When a new view is shown to the system, it attempts to recognize it by looking at the activation levels of allocated R-units. If the new view is sufficiently different from any of the old ones (i.e., none of the allocated R-units passes the threshold), a new R-unit is recruited from the pool of free units. At the same time, a lateral link is established between the two R-units in the representation layer, again by a Hebbian rule. Eventually, a chain of R-units standing for the entire object — the object's *footprint* — is formed in this fashion. By definition, a snapshot of the activity of all the units participating in a footprint (rather than mere connection pattern of these units) constitutes the representation of the object.

3.3 Replicating the Canonical Views Phenomenon

This is what happens when activation is injected at a specific point of a footprint as a result of exposing the system to a test view (assume for the moment that the test view is familiar to the system from the training period; the question of generalization to novel views will be treated later). First, activation is allowed to spread through the lateral links to the footprint-neighbors of the R-unit corresponding to the test view. After a fixed period of time, the activity of the entire footprint is compared to the snapshot stored during training. The degree of similarity between the

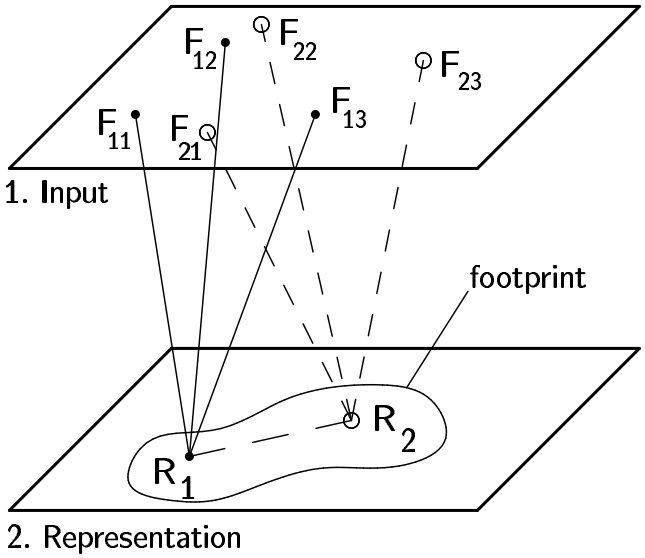

Figure 4: The CLF model represents an object by a collection of its views. Each view is encoded as a conjunction of several features, occuring at well-defined locations in the 2D image. The views are tied together in the order of their original presentation to the system (e.g., in the order of appearance during rotation of the object), forming a characteristic *footprint* of the object. In this schematic example, the first view activates feature units F_{11}, F_{12} and F_{13} in the input layer, and is represented by unit R_1 in the second layer. The second view activates F_{21}, F_{22} and F_{23}, and is represented by R_2.

can use. While the lower levels of recognition are assumed to rely on simple visual events such as individual edge elements or corners, progressively more complex features may be built from these in a hierarchical fashion (for example, a CLF recognizer for a face may use eyes, nose and mouth as features).

The main requirement imposed on the representation of an individual view in the CLF model is that of compactness. In principle, there is no reason why a view should not be jointly represented by a substantial proportion of the second-layer units (see Figure 4). In practice, however, views are better represented by grandmother units, since these can be linked

lation conditions, not all directions of rotation away from a familiar view are equivalent: subjects tolerate about three times as much misorientation in the horizontal than in the vertical plane before recognition is reduced to guessing. Note that this anisotropy is ecologically understandable: creatures confined to the horizontal plane have more use for information about what an object looks like from the side than from above.

2.4 Summary of Human Performance

From the preceding review it appears that at least one of the routes to recognition available to the human visual system can be jointly characterized by a cluster of phenomena — canonical views, mental rotation and limited anisotropic generalization — whose common denominator is *viewpoint dependency*. As the following section shows, accepting viewpoint dependency as the basic premise in computational modeling of recognition allows one to replicate all three central characteristics of human performance discussed above.

3 The CLF Model

3.1 An Overview

Computational accounts of vision describe recognition in terms of a *comparison* between an appropriately encoded and processed input image and an internal *representation* [13]. Different representations thus require different comparison procedures and are bound to result in different recognition performance. In particular, viewpoint-dependent performance can be rather easily obtained with viewpoint-specific representations and a simple comparison method based on template matching. The model proposed in this section does precisely that. What follows is an intuitive description; details can be found in [9].

The model, called CLF (standing for Conjunctions of Localized Features), encodes specific views of objects by recording the co-occurence of arbitrary features at certain locations in (two-dimensional) input images. The CLF framework places no constraints on the type of features that it

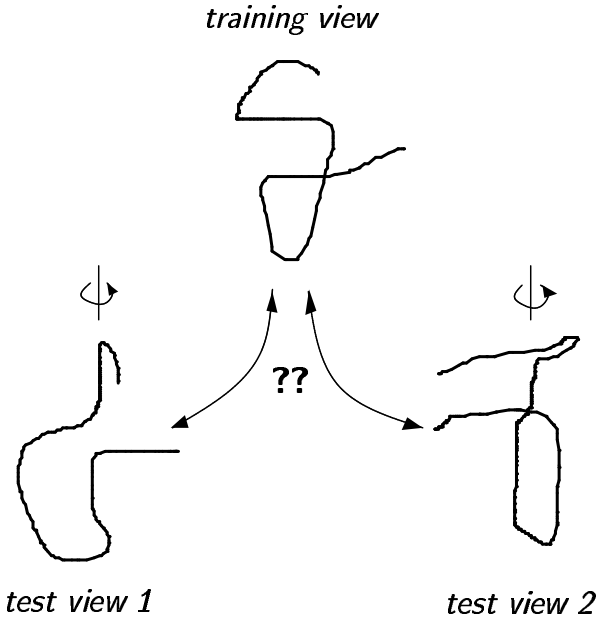

Figure 3: Limited generalization to novel views: error rate for views never seen before by the subject deteriorates rapidly with misorientation relative to a familiar view. If asked which of the bottom two images of wire-like objects matches the one at the top, subjects perform essentially at chance level when the rotation in depth is as small as 40°.

severely foreshortened; see [2]). In comparison, when the task can only be solved through relatively precise shape matching, the error rate reaches chance level already at misorientation of about 40° relative to a familiar attitude ([3,6]; see Figure 3). The detrimental effect of misorientation persists in the presence of depth cues such as binocular disparity, which reduces somewhat the mean error rate, but does not cancel the dependency of error rate on viewpoint [6].

The increase in the error rate depends on the arrangement of familiar views with respect to each other, and not just on the distance between the test view and the nearest familiar view. Specifically, interpolation among familiar views obtained by rotating the object in a fixed plane appears to be easier than extrapolation, which, in turn, is easier than recognition of views that lie outside the plane of rotation [11,6]. Furthermore, under interpo-

The explanation of mental rotation in terms of an analog process involving continuous transformation of internal representations, offered by Shepard and his coworkers, has been incorporated into the foundations of the current paradigm in vision [13]. At present, monotonic dependency of response time on orientation is still widely accepted as evidence for 3D object-centered representations that can be subjected to analog transformations such as rotation, at will.

Caution regarding such an interpretation of the mental rotation phenomena is well-advised in view of recent findings that show the dependency of mental rotation phenomena in recognition on the subject's familiarity with the stimuli. For example, Tarr and Pinker [4] have found that repeated exposure to the same stimulus causes an apparent shift in the subject's strategy: while naming time for novel test views grows monotonically with misorientation relative to the nearest training view, familiar test views yield essentially constant response times (this is consistent with a changeover from time-consuming rotation-based strategy to a faster memory-intensive approach that saves time by storing all frequently occuring views). A similar effect has been reported by Edelman et al. [5,9], who show how both the initial manifestation of mental rotation and its disappearance with exposure can be replicated by a model that does not rely on 3D object-centered representations and, a fortiori, has no means for rotating such representations (see section 3).

2.3 Limited Anisotropic Generalization

The limited ability of the visual system to generalize recognition to novel views of a stimulus (previously seen from a narrow range of viewpoints) is perhaps the most counterintuitive characteristic of human performance in recognition. When asked to give a relatively broad classification of an object seen from an odd viewpoint (that is, when the task requires basic-level categorization), people virtually never err (except when the object appears

shown images were projections of the same 3D object, or of different objects related by a mirror transformation. In this respect, classical mental rotation is different from recognition, where the comparison is made, presumably, between an image of an object and its internal representation.

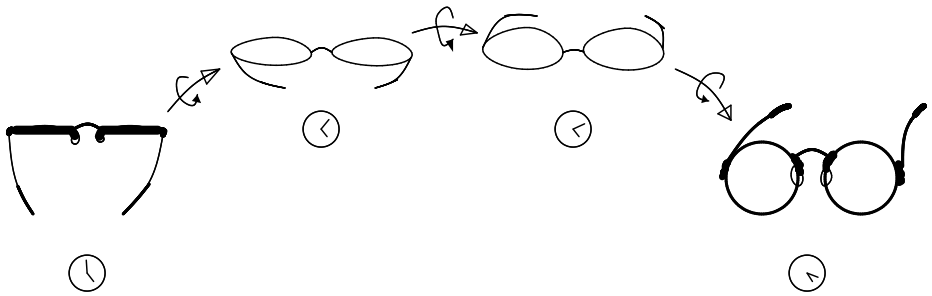

Figure 2: Recognition time for an object grows monotonically with its misorientation relative to a canonical view, as if the object is mentally rotated to match an internal representation. Rates of "rotation" range between 40 and 550 degrees per second, depending on the stimuli and the task. This effect tends, however, to disappear with practice.

are found for synthetic novel objects under controlled exposure conditions, when each view is shown equally often [5].

While uniform initial exposure does not preclude the formation of canonical views, repeated presentation of the same stimulus eventually increases the uniformity of response time over different views of the stimulus. Thus, practice affects the response time aspect of the canonical views phenomenon: after only a few trials, the differences in response time between canonical and random views diminish significantly, even in the absence of any feedback to the subject [5]. Notably, the differences in error rate remain fairly constant.

2.2 Mental Rotation

Transition from a canonical to a non-canonical view of an object does not merely increase the expected recognition time: response latency depends on the viewpoint in an orderly fashion, growing monotonically with misorientation relative to the nearest canonical view ([4,6]; see Figure 2). This dependency of response time on misorientation resembles the celebrated finding by Shepard and Metzler [8] of a class of phenomena that became known as mental rotation (see [12] for an overview).[1]

[1]In Shepard's experiments the task was to determine whether two simultaneously

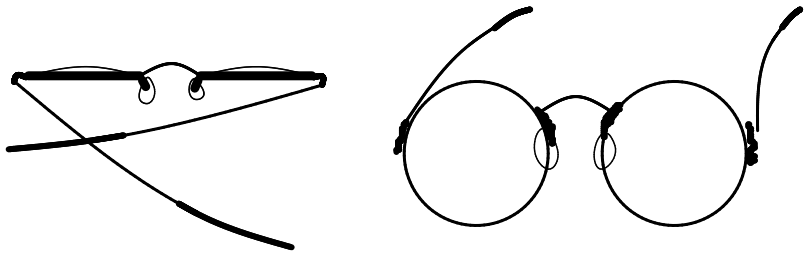

Figure 1: Canonical views: certain views of 3D objects are consistently easier to recognize or process in a variety of visual tasks. For example, a front view of a pair of spectacles is bound to yield lower response time and error rate and to receive higher subjective "goodness" score than a top view of the same object. Such differences may exist even among views that are seen equally often.

2 Shape-Based Recognition Performance in Human Vision

Three basic characteristics of human performance in tasks in which recognition is viewpoint-dependent are illustrated schematically in figures 1 through 3 (detailed accounts of the relevant experiments can be found in the references cited below). These are the phenomena of canonical views [7,5], mental rotation (analogous to the "classical" mental rotation of [8]; see [4,9]), and limited generalization [3,10,6,11]. Following is a brief account of the relevant psychophysical findings.

2.1 Canonical Views

Three-dimensional objects are more easily recognized when seen from certain viewpoints, called canonical, than from other, random, viewpoints (Figure 1). The advantage of canonical views is manifested in consistently shorter response time, lower error rate and higher subjective "goodness" rating [7]. Moreover, this advantage cannot depend solely on the variation in the subject's prior exposure to the different views, since canonical views

work model of recognition. The success of the model in replicating central features of human performance supports the notion that at least one of the available pathways to recognition in the human visual system relies on viewpoint-specific representations.

1 Introduction

The human visual system excels at recognizing three-dimensional objects despite wide variation in the appearance of their retinal projections, caused by changes in illumination and vantage point. For many object classes (for example, human figures and faces) recognition does not break down even when the shape of the object undergoes nonrigid deformation. To a large extent, this performance is made possible by the extreme versatility of vision. In addition to the shape-based pathway to recognition, the existence of which is apparent, e.g., in our ability to identify objects in line drawings and cartoons, there are many other pathways, some of which rely on cues such as characteristic color or texture, others on top-down influences of prior scene knowledge and reasoning. Ready availability of these cues in everyday situations tends to mask certain peculiarities of the shape-based pathway, the study of which in isolation can yield insights into mechanisms of vision.

The same stimulus can engage different processes within the shape-based pathway, depending on the precise specification of the task at hand. For example, if asked to classify a stimulus at the basic category level (see [1]), subjects' performance is essentially viewpoint-invariant [2]. In contrast, at the subordinate levels recognition is markedly dependent on the viewpoint of the observer relative to the object [3,4,5,6]. The present article deals with viewpoint effects in recognition. The next section reviews three major psychophysical characterizations of the shape-based viewpoint-dependent pathway to recognition. The rest of the paper offers a computational account of the psychophysical findings and describes an implemented model whose performance in simulated experiments parallels that of human subjects.

A Network Model of Object Recognition in Human Vision

S. EDELMAN

The Weizmann Institute of Science
Rehovot, Israel

Abstract

Unlike basic-level categorization, which is largely viewpoint-invariant, object recognition at the subordinate levels depends on the observer's point of view in several ways. The first part of this article surveys three viewpoint-dependent aspects of human performance in recognition: canonical views, mental rotation, and limited anisotropic generalization to novel views. The second part offers a detailed but informal computational account of these phenomena, obtained by analyzing the functioning of an implemented net-