11

Evaluating Object Recognition Theories by Computer Graphics Psychophysics

H.H. BÜLTHOFF[1] and S. EDELMAN[2]
[1]Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912, U.S.A.
[2]Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT

Computational or information-processing theories of vision describe object recognition in terms of a comparison between the input image and a set of stored models that represent known objects. The nature of these representations is reflected in the performance of the visual system and may be studied experimentally by presenting subjects with computer graphics simulations of three-dimensional objects (with precisely controlled shape cues), and by analyzing the ensuing patterns of response time and error rate.

We discuss a series of psychophysical experiments that explore different aspects of the problem of subordinate-level object recognition and representation in human vision. Contrary to the paradigmatic view, which holds that the representations are three-dimensional and object-centered, the results consistently support the notion of view-specific representations that include at most partial depth information. In simulated experiments that involved the same stimuli being shown to human subjects, computational models built around two-dimensional, multiple-view representations replicated psychophysical results concerning the observed pattern of generalization errors. We argue that extensions of the multiple-view theory, based on the notion of a hierarchy of spatial and nonspatial features, could lead to a unification of theoretical accounts of a wide range of phenomena in human object recognition.

INTRODUCTION

How does the human visual system represent three-dimensional (3D) objects for recognition? Object recognition is carried out by the human visual system with such expediency that upon introspection it normally appears to be immediate and effortless

Exploring Brain Functions: Models in Neuroscience
Edited by T.A. Poggio and D.A. Glaser

(Fig. 11.1, canonical). Computationally, recognition of a 3D object seen from an arbitrary viewpoint is difficult because its appearance may vary considerably depending on its pose relative to the observer (Fig. 11.1, noncanonical). Because of this variability, simple 2D template matching is hardly a plausible approach to 3D object recognition, since it would require that a template be stored for each view that will ever have to be recognized. Consequently, until recently, most computational theories of object recognition (for a survey, see Ullman 1989) rejected the notion of view-specific representations. Some approaches, rooted in pattern recognition theory, postulated that objects are represented by lists of viewpoint-invariant properties or by points in abstract multidimensional feature spaces (Duda and Hart 1973). Others suggested that the representations are 3D and object-centered (Marr and Nishihara 1978; Biederman 1987), much like the solid geometrical models used in computer-aided design.

Surprisingly, classical theories that rely on object-centered 3D representations fail to account for a number of important characteristics of human performance in recognition. We describe the results of a series of experiments that provide converging evidence in favor of an alternative theory of recognition, based on viewpoint-specific, largely 2D representations. Most of the psychophysical results are accompanied by data from simulated experiments, in which central characteristics of human performance were replicated by computational models based on viewpoint-specific 2D representations. Before proceeding to describe and interpret our computational and psychophysical findings, we discuss briefly several theoretical issues relevant to the understanding of the logic behind the experiments and of the predictions they were designed to test. More about these theories and about the implemented computational models of recognition used in our simulations can be found in Lowe (1986), Biederman (1987), Ullman (1989), Ullman and Basri (1990), Poggio and Edelman (1990), Edelman et al. (1990), and Edelman and Weinshall (1991).

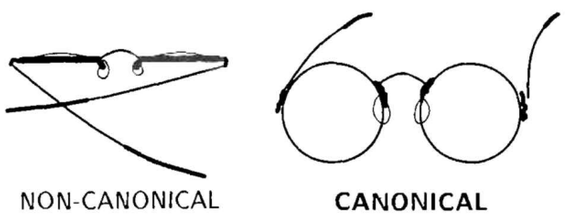

Figure 11.1 Canonical views: certain views of 3D objects are consistently easier to recognize or process in a variety of visual tasks. Once this object is identified as a pair of spectacles seen from above, we find it difficult to believe its recognition was anything less than immediate. Nevertheless, recognition is at times prone to error, and even familiar objects take longer to recognize if they are seen from unusual (noncanonical) viewpoints (Palmer et al. 1981). Exploring this and other related phenomena can help elucidate the nature of the representation of 3D objects in the human visual system.

COMPUTATIONAL THEORIES OF OBJECT RECOGNITION

Explicit computational theories of recognition serve as good starting points for inquiry into the nature of object representation, by providing concrete hypotheses that may be refuted or refined through appropriately designed experiments. Two examples of concrete questions pertinent to the representation issue are:

1. Do the representations include depth or 3D information, or are they, in a sense, flat or 2D?
2. Are the representations object-centered (coded in an object-based coordinate system) or viewer-centered?

One can use the 3D/2D and the object-centered/viewer-centered distinctions among the various theories to generate specific psychophysical predictions. Intuitively, if the representation is 3D, then depth information (available, e.g., through stereopsis) should facilitate recognition. Furthermore, if the representation is object-centered, then neither recognition time nor error rate should depend on the attitude of the object with respect to the observer. To obtain more detailed predictions, one must have a closer look at the different theories.

Theories that Attempt to Achieve Full Object Constancy

Theories of the first kind we mention here attempt to achieve a computer vision equivalent of object constancy,[1] a central characteristic in the human perception of 3D shape (Ellis et al. 1989). Two major approaches to object constancy can be discerned. The first approach uses fully 3D object-centered representations and requires that a similar representation of the input be achieved before it is matched to memory. The second approach, however, represents objects by selected views that include depth information, and attempts to normalize the appearance of the input by applying an appropriate spatial transformation.

Theories that Use Viewpoint-independent 3D Representations

The notion that the processing of the visual input culminates in a full restoration of its 3D structure, which may then be matched to 3D object-centered representations in memory, was popularized by Marr and Nishihara (1978). This theory was never fully implemented due to severe difficulties faced by the attempts to solve the problem of reconstructing the input in 3D in the general case. However, a variant of this approach became widely adopted as a psychological model of recognition, following the work

[1] The tendency of human subjects to perceive and recognize 3D shapes irrespective of factors such as the viewpoint.

of Biederman and his associates. Biederman's theory, known as recognition by components, postulates that the human visual system represents objects by 3D structural relationships between generic volumetric primitives called geons (Biederman 1987).

Theories that Use Viewpoint-specific 3D Representations in Conjunction with Normalization

As a representative of this class of theories, we consider recognition by viewpoint normalization, of which Ullman's method of alignment is an example (Ullman 1989). In the alignment approach the 2D input image is compared with the projection of a stored model, much like in template matching, but only after the two are brought into register. The transformation necessary to achieve alignment is computed by matching a small number of features in the image with the corresponding features in the 3D model. The aligning transformation is computed separately for each of the models stored in the system. The outcome of the recognition process is the model that fits the input most closely after the two are aligned. Related schemes (Lowe 1986; Thompson and Mundy 1987) choose the best model using viewpoint consistency constraints that relate the projected locations of the features of a model to its 3D structure, given a hypothesized viewpoint.

Ullman (1989) distinguishes between full alignment that uses 3D models and attempts to compensate for 3D transformations of objects, such as rotation in depth, and the alignment of pictorial descriptions that decomposes objects into (nongeneric) parts and uses multiple views rather than a single object-centered description. Ullman also notes (Ullman 1989, p. 228) that this multiple-view version of alignment involves representation that is "view-dependent, since a number of different models of the same object from different viewing positions will be used," but at the same time "view-insensitive, since the differences between views are partially compensated by the alignment process." Thus, view-independent performance (e.g., error rate) can be considered the central distinguishing feature of both versions of this theory, which subsequently will be referred to simply as alignment.

Theories that Use 2D Representations

Linear Combination (LC) of Views

Three recently proposed approaches to recognition dispense with the need to store 3D models. The first of these, recognition by LC of views (Ullman and Basri 1990), is built on the observation that, under orthographic projection, the 2D coordinates of an object point can be represented as a LC of the coordinates of the corresponding points in a small number of fixed 2D views of the same object. The required number of views depends on the allowed 3D transformations of the objects and on the representation of an individual view. For a polyhedral object to undergo a general linear transformation, three views are required if separate linear bases are used to represent the x and

the y coordinates of a new view. Two views suffice if a mixed x, y basis is used (Ullman and Basri 1990; Edelman and Poggio 1990). A system that relies solely on the LC approach should achieve uniformly high performance on those views that fall within the space spanned by the stored set of model views, and should perform poorly on views that belong to an orthogonal space.

Two-dimensional View Interpolation (HyperBF)

Another approach that represents objects by sets of 2D views is view interpolation by hyper basis functions (HyperBF) implemented as regularization networks (Broomhead and Lowe 1988; Moody and Darken 1989; Poggio and Edelman 1990; Poggio and Girosi 1990). In this approach, generalization from stored to novel views is regarded as a problem of multivariate function interpolation in the space of all possible views. The interpolation is performed in two stages. In the first, intermediate responses are formed by a collection of nonlinear receptive fields (these can be, e.g., multidimensional Gaussians). The output of the second stage is a LC of the intermediate receptive field responses.

More explicitly, a Gaussian-shaped basis function is placed at each of the prototypical stored views of the object so that an appropriately weighted sum of the Gaussians approximates the desired characteristic function for that object over the entire range of possible views (for details see Poggio and Edelman 1990; Edelman and Poggio 1990). Recognition of the object represented by such a characteristic function amounts to a comparison between the value of the function computed for the input image and a threshold situated between 0 and 1.

Conjunction of Localized Features (CLF)

The third scheme we mention is also based on interpolation among 2D views and, in addition, is particularly suitable for modeling the time course of recognition, including long-term learning effects (Edelman and Weinshall 1991; Edelman 1991b). The scheme is implemented as a two-layer network of thresholded summation units. The input layer of the network is a retinotopic feature map (thus the model's name). The distribution of the connections from the first layer to the second, or representation, layer is such that the activity in the second layer is a blurred version of the input. Unsupervised Hebbian learning augmented by a winner-take-all operation ensures that each sufficiently distinct input pattern (such as a particular view of a 3D object) is represented by a dedicated small clique of units in the second layer. Units that stand for individual views are linked together in an experience-driven fashion, again through Hebbian learning, to form a multiple-view representation of the object. When presented with a novel view, the CLF network can recognize it through a process that amounts to blurred template matching and is related to nonlinear basis function interpolation.

IMPLICATIONS OF THE DIFFERENT THEORIES

Experimental Issues

We have investigated the subjects' performance in three distinct cases, each corresponding to a different kind of test views. In the first and easiest case, the test views are familiar to the subject (that is, were shown during training). In the second case, the test views are unfamiliar, but are related to the training views through a rigid 3D transformation of the target. In this case the problem can be regarded as generalization of recognition to novel views. In the third case, which is especially relevant in the recognition of articulated or flexible objects, the test views are obtained through a combination of rigid transformation and nonrigid deformation of the target object.

An additional and important issue we have addressed is the role of depth cues in recognition. Providing the subject with ample depth cues should increase the plausibility of attributing any subsequent manifestation of viewpoint dependency in recognition to viewpoint-dependent representation, rather than to general scarcity of 3D information in the stimulus during training. Second, adding depth to the stimulus eliminates the possibility that the subjects do form a 3D object-centered representation during training; they fail, however, to make full use of it because test images normally used in psychophysical investigations of recognition, being inherently two dimensional, do not activate the "real" 3D pathway to recognition.

Theoretical Predictions

The theories discussed above make different predictions about the effect of factors such as object orientation and the presence of depth information on the accuracy of recognition and on the amount of time it takes under the various conditions mentioned above. Roughly speaking, theories that rely on viewpoint-invariant representations predict no systematic effect of orientation either on the response time or on the error rate, both for familiar and for novel test views, provided that the representation primitives (i.e., invariant features or generic parts) can be readily extracted from the input image. In comparison, theories that involve viewpoint-dependent representations naturally predict viewpoint-dependent performance. The details of the predictions vary according to the recognition method postulated by each particular theory and are discussed below.

Viewpoint-independent 3D Representations

A recognition scheme based on object-centered 3D representations may be expected to perform poorly only for those views which by an accident of perspective lack the information necessary for the recovery of the reference frame in which the object-centered description is to be formed (Biederman 1987). In a standard example of this situation, an elongated object is seen end-on, causing a foreshortening of its major axis

and an increased error rate, due presumably to a failure to achieve a stable description of the object in terms of its parts (Marr and Nishihara 1978; Biederman 1987). In all other cases this theory predicts independence of response time on orientation and a uniformly low error rate across different views. Furthermore, the error rate should remain low even for deformed objects as long as the deformation does not alter the make-up of the object in terms of its parts and their interelationships.

Viewpoint-dependent 3D Representations

Consider next the predictions of those theories that explicitly compensate for viewpoint-related variability of apparent shape of objects, by normalizing or transforming the object to a standard viewpoint. A system that represents an object by one or more of its views and uses an incremental transformation process for viewpoint normalization is expected to exhibit a response time that will depend monotonically on the misorientation of the test view relative to one of stored views. This pattern of response times will hold for many of the familiar, as well as for novel test views, since the system may store selectively only some of the views it encounters for each object and may rely on normalization for the recognition of other views, either familiar or novel. In contrast to the expected dependence of response time on orientation, the error rate under the viewpoint normalization approach will be uniformly low for any test view, either familiar or novel, in which the information necessary for pose estimation is not lost.

Linear Combination of Views

The predictions of the LC scheme vary according to the particular version used. The basic LC scheme predicts uniformly successful generalization to those views that belong to the space spanned by the stored set of model views. It is expected to perform poorly on views that belong to an orthogonal space. In contrast, the mixed-basis LC (MLC) is expected to generalize perfectly, just as the 3D object-centered schemes do. Furthermore, the varieties of the LC scheme should not benefit significantly from the availability of depth cues because they require that the views be encoded as lists of coordinates of object features in 2D and cannot accomodate depth information. Regarding the recognition of deformed objects, the LC method will generalize to any view that belongs to a hyperplane spanned by the training views (Ullman and Basri 1990). For the LC+ scheme (that is, LC augmented by quadratic constraints verifying that the transformation in question is rigid), the generalization will be correctly restricted to the space of the rigid transformations of the object, which is a nonlinear subspace of the hyperplane that is the space of all linear transformations of the object.

View Interpolation

Finally, consider the predictions of the view interpolation theory. First, no effect of orientation on response time is expected, except as a byproduct of a specific implemen-

tation.[2] Second, a lower error rate is predicted for familiar than for novel test views, depending on the distance to the nearest stored view. Similarly, some variation in the error rate among the familiar views is also possible, if the stored prototypical views form a proper subset of the previously seen ones (in which case views that are the closest to the stored ones will be recognized more reliably than views that have been previously seen, but were not included in the representation). For deformed objects, generalization is expected to be as significant as for novel views produced by rigid transformations. Furthermore, better generalization should be obtained for test views produced by the same deformation method used in training.

COMPUTER GRAPHICS PSYCHOPHYSICS

The possibility of directly testing the intricate implications of the different computational theories surveyed above depends critically on the availability of a sufficient variety of stimuli that are, on one hand, easily controlled, and, on the other, complex enough to yield results relevant to real-world situations. Until recently, this dilemma placed severe limitations on experimental work in visual object recognition. Clearly, its resolution depended on the emergence of new psychophysical paradigms. We have addressed this need for a new approach by developing experimental techniques that use state-of-the-art computer graphics to create and display 3D objects (see Fig. 11.2) with the following attributes under full software control:

- *Novelty*: An unlimited number of novel objects can be produced to a given set of specifications. In addition to interactive manipulation of objects in a graphics environment, we use a powerful 3D graphics language (Symbolics S-Geometry based on object-oriented programing in Lisp) to generate arbitrary large object sets under complete program control. This makes the study of perceptual learning possible, which otherwise is difficult to do with real-world objects because of the effects of the subject's previous exposure.
- *Shape*: The set of possible shapes is only limited by the imagination of the experimenter; the Symbolics S-Geometry system is a flexible and versatile tool for the generation of arbitrary 3D shapes.
- *Features*: Both local features and global properties can be precisely controlled. Texture mapping or albedo manipulation facilitates the inclusion of nongeometric features.

[2] Note that both the prediction of a monotonic increase in response time with misorientation relative to a stored view made by the normalization theories, and the prediction of constant response time made by the view interpolation theory, are weak because of their potential dependence on implementation details. In the first case, the monotonic increase in response time will disappear if the transformation mechanism is "one-shot" instead of incremental. In the other case, response time will depend on orientation if the interpolation involves a time-consuming spread of activation in a distributed implementation.

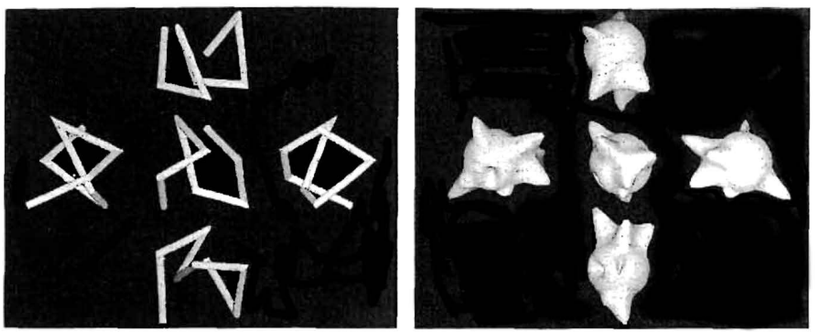

Figure 11.2 Wires and Amoebae. The appearance of a 3D object can depend strongly on the viewpoint. The image in the center represents one view of a computer graphics object (wire- or amoeba-like). The other images are derived from the same object by ± 75° rotation around the vertical or horizontal axis. The difference between the images illustrates the difficulties encountered by any straightforward template matching approach to 3D object recognition. Thin wire-like objects have the nice property that the negligible amount of occlusion provides any recognition system with equal amount of information for any view. A realistic recognition system has to deal with the more difficult situation of self-occlusion as demonstrated with the amoeba-like objects.

- *Depth cues*: The simulated objects can be given surface texture and realistic shading. In addition, they can be displayed with a variable amount of binocular disparity using a StereoGraphics 3Display system.
- *Motion*: The stimuli can be shown to the subject in simulated motion (to engage the structure from motion mechanisms) or statically. Even complicated objects with many polygons can be rendered in real-time on a graphics mini-supercomputer (GS1000, Stardent Inc.).
- *Deformation*: Controlled distortions can be introduced into the shape of the simulated objects. This facilitates the study of phenomena hitherto excluded from the psychophysical repertoire (e.g., the dependence of the recognition of deformable objects such as human faces on the amount and the nature of the deformation).

In addition to allowing us extensive control over stimulus presentation, the new paradigm presents some unique opportunities for analyzing the experimental data. In particular, we compare the subjects' performance with the performance of detailed computational models of recognition in simulated experiments involving the *same* stimuli and exposure conditions, made possible by the computer graphics environment.

We now turn to survey the experimental findings concerning the recognition of 3D objects, starting with a review of the four main characteristics of object recognition that emerge from previous studies: the distinction between classification and identifi-

cation, the existence of canonical views, the phenomenon resembling mental rotation, and the limited ability of subjects to identify objects from novel viewpoints.

PSYCHOPHYSICAL BACKGROUND

Basic vs. Subordinate-level Recognition

Numerous studies in cognitive science (for a review, see Rosch et al. 1976) reveal that in the hierarchical structure of object categories there exists a certain level, called basic level, that is the most salient according to a variety of criteria (such as the ease and preference of access). Taking as an example the hierarchy "quadruped, mammal, cat, Siamese," the basic level is that of "cat." Objects whose recognition implies more detailed distinctions than those required for basic-level categorization are said to belong to a subordinate level. The pattern of response times and error rates in recognition experiments appears to be influenced to a large extent by the category level at which the distinction between the different stimuli is to be made. Specifically, if the subjects are required to *classify* the stimulus (that is, to determine its basic-level category), they normally exhibit near-zero error rate, independently of the stimulus orientation (except when the 3D structure of the object is severely distorted, e.g., due to foreshortening; see Biederman 1987). In contrast, if the task is to *identify* a given stimulus (that is, to distinguish it from other, similar stimuli belonging to the same basic category), the subjects' performance is strongly dependent on the viewpoint.

Canonical Views

Commonplace objects such as houses or cars are recognized faster or more slowly, depending on the attitude of the object with respect to the observer. Palmer, Rosch, and Chase (1981) found that human subjects consistently labeled certain views of such objects as subjectively "better" than other, random views. In a naming task, subjects tended to respond quicker when the stimulus was shown from a good or canonical perspective, with the response time increasing monotonically with misorientation relative to a canonical view (determined independently in a subjective judgement experiment). At the basic level, canonical views are largely a response time phenomenon (the error rate for basic-level naming, as found by Palmer et al., was very low, with the errors being slightly more frequent for the worst views than for others). In comparison, at the subordinate levels canonical views are apparent in the distribution of error rates as well, where they constitute strong and stable evidence in favor of viewpoint-dependent nature of object representations (see section below on **Canonical Views and Their Development with Practice**).

Mental Rotation and Its Disappearance with Practice

The body of evidence documenting the monotonic dependency of recognition time on the object's attitude relative to a canonical view has been interpreted recently (Tarr

and Pinker 1989, 1990) as an indication that objects are represented by a few specific views, and that recognition involves viewpoint normalization or alignment (Ullman 1989) to the nearest stored view, by a process related to mental rotation (Shepard and Cooper 1982). A number of researchers have shown the differences in response time among familiar views to be transient, with much of the variability disappearing with practice (see, e.g., Jolicoeur 1985; Koriat and Norman 1985; Tarr and Pinker 1989). Tarr and Pinker (1989) investigated the effect of practice on the pattern of response times in the recognition of novel objects, which are particularly suitable for this purpose because they offer the possibility of complete control over the subjects' prior exposure to the stimuli. They have found that the monotonic dependency of response times on the stimulus attitude, which disappeared after repeated exposure to the same set of test views, reappeared for "surprise" test views, only to fade away again as these novel views became familiar to the subjects.

Limited Generalization

The pattern of error rates in recent experiments by Rock and his collaborators (Rock and DiVita 1987) indicates that when the recognition task can only be solved through relatively precise shape matching (such as required for subordinate-level recognition), the error rate reaches chance level already at a misorientation of about 40° relative to a familiar attitude (Rock and DiVita 1987; see also Fig. 11.5). A similar limitation seems to hold for people's ability to imagine what an object looks like from an unfamiliar viewpoint (Rock et al. 1989).

PSYCHOPHYSICS OF SUBORDINATE-LEVEL RECOGNITION

Previously published psychophysical works left many of the questions vital to computational understanding of recognition unanswered. First, it was unclear whether the canonical views phenomenon reflected basic viewpoint dependence of recognition or was due to particular patterns of the subjects' exposure to the stimuli. Second, the existing data were insufficient to test the predictions of the different theories concerning generalization to novel views and across object deformations. Third, the role of depth cues in recognition remained largely unknown. The experiments described in this section were designed to address those issues, concentrating on subordinate-level identification, which, unlike basic-level classification (Biederman 1987), has been relatively unexplored.

The experiments described below consisted of two phases: training and testing. In the training phase subjects were shown an object defined as the target, usually as a motion sequence of 2D views that led to an impression of 3D shape through the kinetic depth effect. In the testing phase the subjects were presented with single static views of either the target or a distractor (one of a relatively large set of similar objects). The subject's task was to press a "yes"-button if the displayed object was the current target

and a "no"-button otherwise, and to do it as quickly and as accurately as possible. No feedback was provided as to the correctness of the response.

Canonical Views and Their Development with Practice

To explore the first issue raised above, that of the determinants of canonical views, we tested subjects' recognition of views all of which have been previously seen as a part of the training sequence (for further details see Edelman and Bülthoff 1992, experiment 1). Our stimuli proved to possess canonical views, despite the fact that in training all views appeared with equal frequency. We also found that the response times for the different views became more uniform with practice. The development of canonical views with practice is shown in Fig. 11.3 as a 3D stereo-plot of response time vs. orientation, in which local deviations from a perfect sphere represent deviations of response time from the mean. For example, the difference in response time between a "good" and a "bad" view in the first session (the dip at the pole of the sphere and the large protrusion in Fig. 11.3, top) decreases in the second session (Fig. 11.3, bottom). The pattern of error rates, in comparison, remained largely unaffected by repeated exposure.

Role of Depth Cues

Depth Cues and the Recognition of Familiar Views

We next explored the role of three different cues to depth in the recognition of familiar views (for details, see Edelman and Bülthoff 1992, experiment 2). Whereas in the previous experiment, test views were 2D and the only depth available cues were shading of the objects and interposition of their parts, we now added texture and binocular stereo to some of the test views and manipulated the position of the simulated light source to modulate the strength of the shape from shading cue (cf. Bülthoff and Mallot 1988; Pentland 1989).

The stimuli were rendered under eight different combinations of values of three parameters: surface texture (present or absent), simulated light position (at the simulated camera or to the left of it), and binocular disparity (present or absent). Training was done with maximal depth information (oblique light, texture and stereo present). Stimuli were presented using a noninterlaced stereo viewing system (StereoGraphics 3Display). A fixed set of views of each object was used, both in training and in testing. We found that both binocular disparity and, to a smaller extent, light position affected performance. The error rate was lower in the STEREO compared to MONO trials (11.5% as opposed to 18.0%) and lower under oblique lighting than under head-on lighting (13.7% compared to 15.8%).

Depth Cues and the Generalization to Novel Views

We then proceeded to probe the influence of binocular disparity (shown to be the strongest contributor of depth information to recognition) on the generalization of

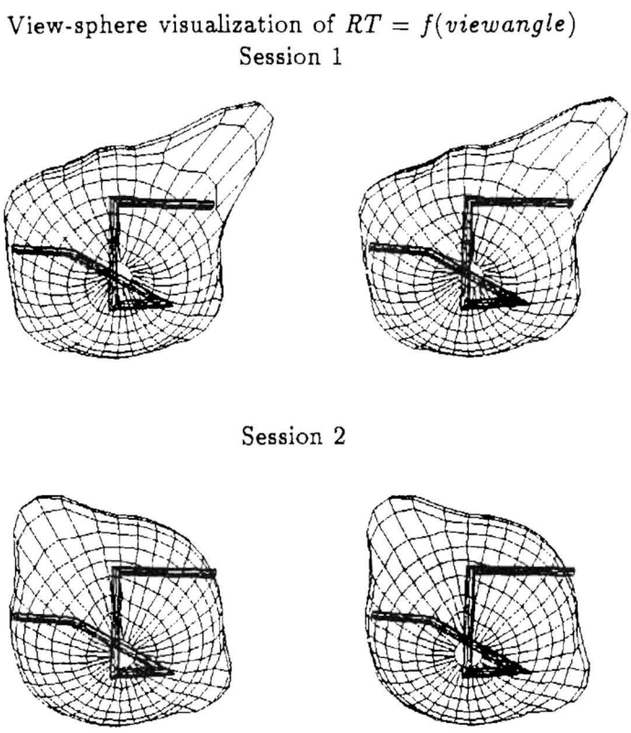

Figure 11.3 Canonical views and practice: the advantage of some views over others, as manifested in the pattern of response times (RT) to different views, is reduced with repeated exposure. The spheroid surrounding the target is a 3D stereo-plot of response time vs. aspect (local deviations from a perfect sphere represent deviations of response time from the mean). The 3D plot may be viewed by free-fusing the two images in each row, or by using a stereoscope. *Top*: Target object and response time distribution for session 1. Canonical aspects (e.g., the broadside view, corresponding to the visible pole of the spheroid) can be easily visualized using this display method. *Bottom*: The response time difference between views are much smaller in the second session. Note that not only did the protrusion in the spheroid in session 1 disappear but also the dip in the polar view is much smaller in session 2.

recognition to novel views (for details, see Edelman and Bülthoff 1992, experiment 4). The subjects were first trained on a sequence of closely spaced views of the stimuli, then tested repeatedly on a different set of views, spaced at 10 intervals (0° to 120° from a reference view at the center of the training sequence).

The mean error rate in this experiment was 14.0% under MONO and 8.1% under STEREO. In the last session of the experiment, by the time the transient learning effects had disappeared, the error rate under MONO approached the error rate under STEREO,

except for the range of misorientation between 50° and 80°, where MONO was much worse than STEREO. Notably, error rate in each of the two conditions in the last session was significantly dependent on misorientation.

Generalization to Novel Views

Our next experiment used an elaborate generalization task to distinguish among three classes of object recognition theories mentioned earlier: alignment, LC of views, and view interpolation by radial basis functions (HyperBF). Specifically, we explored the dependence of generalization on the relative position of training and test views on the viewing sphere (for details, see Bülthoff and Edelman 1992).

We presented the subjects with the target from two viewpoints on the equator of the viewing sphere, 75° apart. Each of the two training sequences was produced by letting the camera oscillate with an amplitude of ±15° around a fixed axis (Fig. 11.4).

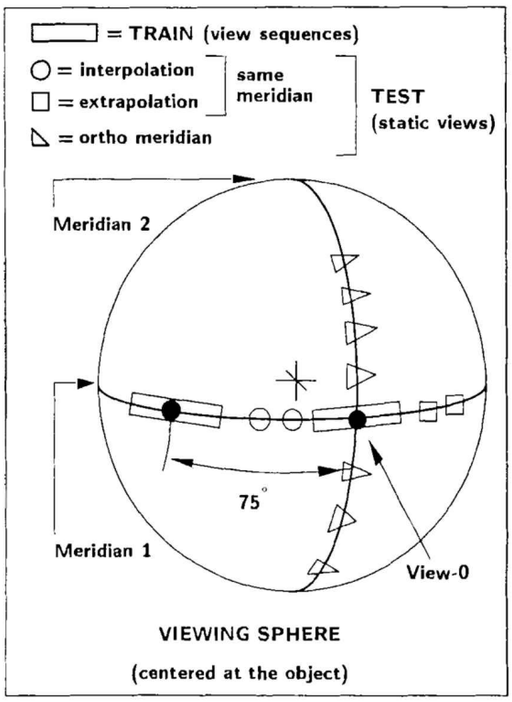

Fig. 11.4 Generalization to novel views: An illustration of the INTER, EXTRA, and ORTHO conditions. Computational theories of recognition outlined in the first section (COMPUTATIONAL THEORIES OF OBJECT RECOGNITION) generate different predictions as to the relative degree of generalization in each of the three conditions. We have used this to distinguish experimentally between the different theories.

Target test views were situated either on the equator (on the 75° or on the 360° – 75° = 285° portion of the great circle, called INTER and EXTRA conditions), or on the meridian passing through one of the training views (ORTHO condition; see Fig. 11. 4).

The results of the generalization experiment, along with those of its replica involving the HyperBF model, appear in Fig. 11.5 (see also the summary in Table 11.1). As expected, the subjects' generalization ability was far from perfect. The mean error rates for the INTER, EXTRA, and ORTHO view types were 9.4%, 17.8%, and 26.9%, respectively. Repeated experiments involving the same subjects and stimuli, as well as control experiments under a variety of conditions yielded an identical pattern of error

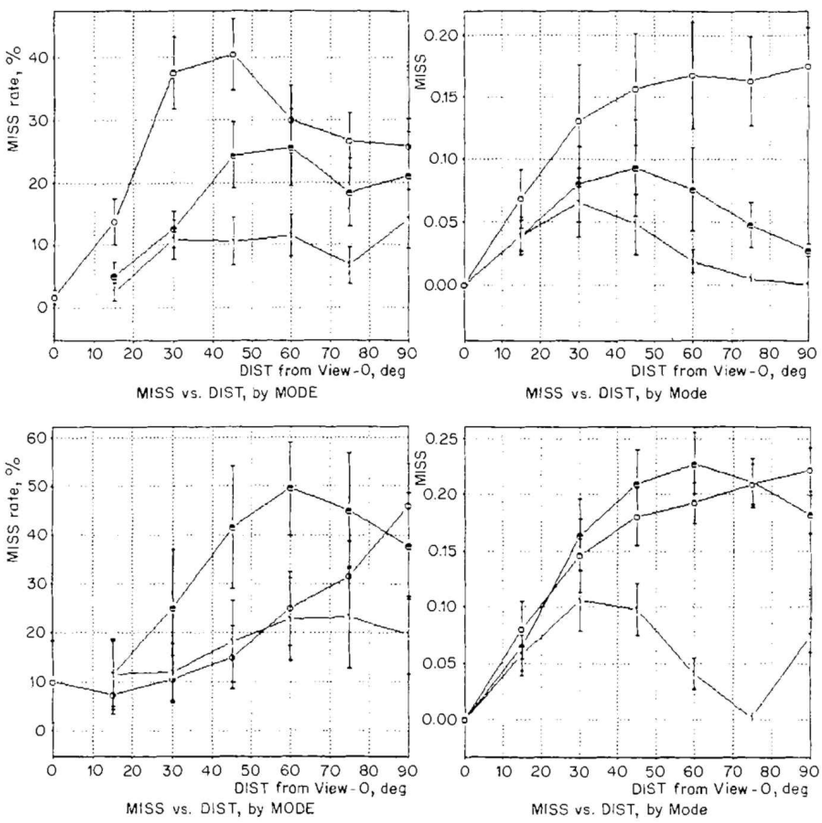

Figure 11.5 Generalization to novel views. Top left: Error rate vs. misorientation relative to the reference ("view-0" in Fig. 11.4) for the three types of test views—INTER, EXTRA, and ORTHO, horizontal training plane. Top right: performance of a HyperBF model in a simulated replica of this experiment. Bottom left and right: same as above, except vertical training.

Table 11.1. Generalization to novel views: error rate for novel views, by condition, as predicted by the different theories of object recognition outlined in the first section (COMPUTATIONAL THEORIES OF OBJECT RECOGNITION). The reciprocal of the error rate is an indicator of generalization. The last line describes human performance (error rate in percent).

Condition → / Theory ↓	Train in Horizontal Plane			Train in Vertical Plane		
	INTER	*EXTRA*	*ORTHO*	*INTER*	*EXTRA*	*ORTHO*
Alignment	Low	Low	Low	Low	Low	Low
LC	Low	Low	High	Low	Low	High
CLC	Low	Medium	High	Low	Medium	High
MLC	Low	Low	Low	Low	Low	Low
HyperBF	Low	Medium	High	Medium	High	Low
Humans	13.3	22.0	48.3	17.9	35.1	21.7

rates. The order of the mean error rates was changed, however, when the training views lay in the vertical instead of the horizontal plane. In this case, the means for the INTER, EXTRA, and ORTHO conditions were 17.9%, 35.1%, and 21.7%, respectively.

The experimental results fit most closely the predictions of the HyperBF scheme and contradict theories that involve 3D object-centered models or viewpoint normalization. The differences of generalization performance between the horizontal and the vertical arrangements of training views (compare the figures in the last line of Table 11.1 for the two training conditions) can be accommodated within the HyperBF framework by assigning different weights to the horizontal and the vertical dimensions (equivalent to using nonradial basis functions).

Generalization across Deformations

In the last experiment reported here, we compared the generalization of recognition to novel views belonging to three different categories: those obtained from the original target object by rigid rotation in depth, by 3D affine transformation, and by nonuniform (hence, nonlinear) deformation (for details, see Edelman and Bülthoff 1990a). Half of the views in the rigid rotation category were obtained by rotation around the X axis (that is, in the sagittal plane) and half around the Y axis. In the affine category, the transformation methods were shear in X (specifically, $x = ay + bz$ for each object point), shear in Y ($y = ax + bz$) and general linear (represented by an arbitrary 3×3 matrix). Thus, altogether views obtained through six different transformation classes were tested.

From the experimental results it appears that the degree of generalization exhibited by the human visual system is determined more by the amount of (2D) deformation as measured in the image plane than by the direction and the distance between the novel and the training views in the abstract space of all views of the target object. The

HyperBF scheme (incorporating the horizontal/vertical asymmetry) can be made to produce a similar pattern of performance.

Interpretation of the Experimental Data: Support for a View Interpolation Theory of Recognition

The experimental findings reported above are incompatible with theories of recognition that postulate object-centered representations. Such theories predict no differences in recognition performance across different views of objects and therefore cannot account either for the canonical views phenomenon or for the limited generalization to novel views, without assuming that, for some reason, certain views are assigned a special status. Modifying the thesis of viewpoint-independent representation to allow privileged views and a built-in limit on generalization greatly weakens it by breaking the symmetry that holds for truly object-centered representations, in which all views, including novel ones, are equally easily accessed.

Part of the findings on viewpoint-dependent recognition, including mental rotation and its disappearance with practice, and the lack of transfer of the practice effects to novel orientations or to novel objects (Tarr and Pinker 1989), can be accounted for in terms of viewpoint normalization or alignment (Ullman 1989). According to the alignment explanation, the visual system represents objects by small sets of canonical views and employs a variant of mental rotation to recognize objects at attitudes other than the canonical ones. Furthermore, practice causes more views to be stored, making response times shorter and more uniform. At the same time, the pattern of error rates across views, determined largely by the second stage of the recognition process in which the aligned model is compared to the input, remains stable due to the absence of feedback to the subject.

This explanation, however, is not compatible with the results of the generalization experiments, which, on one hand, show a marked and persistent dependency of error rate on the distance to the training view for rigid rotations,[3] and, on the other, indicate that people are capable of generalization across object deformations. Moreover, the viewpoint dependency of the representations formed by our subjects, manifested in the limitation on generalization to novel views, cannot be due exclusively to the lack of 3D information in the stimuli, since the same dependency of error rate on viewpoint was obtained (in the depth-cues experiment) both in MONO and STEREO trials.

In view of the experimental results discussed above, theories that rely on fully 3D object-centered representations appear to be poor models of human performance, at least in tasks that require subordinate-level recognition. A plausible alternative account of the experimental data assumes that object representations involved in such tasks are inherently viewpoint-dependent. According to this account, a 3D object is repre-

[3] These findings also rule out the possibility that the increase in the uniformity of response time over different views, caused by practice, is due to the formation of a viewpoint-invariant representation of the target object.

sented by a collection of specific views, each of which is essentially a snapshot of the object as it is seen from a certain viewpoint, augmented by limited depth information.[4] The collection of stored views is structured, in the sense that views that "belong" together (e.g., because they appeared in close succession during previous exposure) are more closely associated with each other (Edelman and Weinshall 1991). To precipitate recognition, an input stimulus must bring the entire structure to a certain minimal level of activity. This process of activation may be mediated by a correlation-like operation that compares the stimulus (possibly in parallel) with each of the stored views, and activates the representation of that view in proportion to its similarity to the input (Edelman 1991b). Computationally, this method of recognition is equivalent to an attempt to express the input as an interpolation of the stored views (Poggio and Edelman 1990; Edelman and Weinshall 1991), which is much more likely to succeed if the input image is indeed a legal view of the 3D object represented by the collection of stored views (Ullman and Basri 1990).

WHAT ARE THE FEATURES OF RECOGNITION?

All of our psychophysical findings reported above have been replicated by a computational model based on interpolation of stored 2D views[5] (Edelman et al. 1989; Edelman and Bülthoff 1990a; Bülthoff and Edelman 1992); more details about the model are given by Poggio and Girosi (see chapter 8). A natural question arising at this point is how those 2D views are represented in the human visual system. It is instructive to compare the different possibilities that suggest themselves to the method of representation used by the HyperBF network model. The input to the model is a vector of measurements of certain image parameters. In the simplest case, these parameters are the image coordinates of primitive features such as edge terminators or corners. While these features are suitable for the class of thin tube-like objects used in most of our experiments to date, they are clearly inadequate for the description of objects in which intensity edges and, in particular, edges due to the occluding contour are of secondary importance. An example of an object class that dictates a reconsideration of the feature issue appears in Fig. 11.2. It should be noted that amoeba-like stimuli yield the same pattern of results as do the wire-like objects used throughout the experiments reported above. These results, however, cannot be replicated computationally without an in-depth study of the feature extraction stage of recognition in human vision. In this section we outline our approach to the study of the features of recognition in human vision (see Edelman 1991a for more details).

---

[4] The basic limitation on the use of depth in recognition stems from its representation in a viewer-centered coordinate frame (in Marr's terminology [Marr 1982] such representation would be called a 2 1/2 D-sketch). Another possible limitation is expected in view of the recent findings regarding the imperfections of the perception of 3D shape, as mediated by different cues (Bülthoff and Mallot 1988).

[5] An exception is the effect of binocular stereo, which has not yet been incorporated into the model.

The central tenet of our approach, supported by the evidence presented in the preceding sections, is that recognition normally requires neither 3D reconstruction of the stimulus nor the maintenance of a library of 3D models of objects (Edelman and Poggio 1989). Instead, information sufficient for recognition can be found in the 2D image locations of object *features*. The choice of features and their complexity may vary between objects. For example, a pineapple can be recognized by its characteristic pattern of spiny scales. The main feature in this case is textural and is distributed over the object's surface. In comparison, the relevant features of a peanut are both its texture and a characteristic outline (in a line drawing, a round peanut can be confused with a golf ball). Finally, a road vehicle can be recognized as such by the presence of wheels (each of which may be considered a complex feature); however, for the drawing of a vehicle to be classified, e.g., as a car, simple additional features such as contour elements and corners must be appropriately situated in the image (presumably, in the vicinity of the locations of corresponding features in the image of a prototypical car). The rest of this section describes a generic recognition scheme based on the idea of a hierarchy of image features.

Feature-based Recognition

The scheme whose implications we are currently exploring learns to recognize new objects by constructing and storing their representations in terms of the current feature set. The first stage of processing both in learning and in subsequent recognition is feature extraction. A small sample of features extracted at this stage includes contour elements and events (corners, inflections, terminations), color and texture labels, as well as *compound* features — recurring combinations of simple features proven useful by previous experience (e.g., consider the elliptical blob or "wheel" feature in vehicle recognition).

Only a few of the potentially many features detected in the first stage are involved in the recognition of any given object. The relevance of a feature for the recognition of an object is determined by its diagnosticity. Thus, in basic-level recognition, a good feature is one that distinguishes the object's category from other basic categories, while on a subordinate level a distinction is sought between similar members of the same category (LaBerge 1976).

Once the relevant features of an object are identified, the system synthesizes a recognizer for the object, essentially by storing the pattern of activities of the chosen feature detectors (Edelman and Weinshall 1991; for a background, see, e.g., Aha et al. 1991). Generalization to other views of the same object or to other members of the same basic category is achieved by interpolating among patterns corresponding to several different stored instances (Poggio and Edelman 1990). This interpolation can be carried out, e.g., by HyperBF networks (Poggio and Girosi 1990). The advantages of using this method include a natural distributed implementation by a network of "receptive fields," and the possibility of integrating a variety of feature types within the same interpolation mechanism (Edelman and Poggio 1990).

PSYCHOPHYSICS OF RECOGNITION: A UNIFIED VIEW

The plausibility of a unified account of basic-level and subordinate-level performance centered on the notion of features of recognition is supported by the existence of several points of similarity between performance characteristics at the two levels. First, objects possess canonical views, whether they are considered at the basic (Biederman 1985, p.66) or at a subordinate level (Edelman and Bülthoff 1990b; Edelman et al. 1991a). Second, the same processes may underlie the emergence of canonical views at both levels. For instance, an examination of the data of the canonical views experiment (see section on **Canonical Views and Their Development with Practice**) reveals that the canonical views can be frequently characterized as displaying one or more of the nonaccidental features mentioned by Biederman (1985, p. 36). Third, at the basic level, viewpoint-invariant performance breaks down when key information (the nonaccidental features in Biederman [1985], or the anchor features in Ullman [1989]) cannot be reliably extracted from the image. These similarities support the unified feature-based account outlined in the section above. In particular, the dichotomy between viewpoint invariance of basic-level recognition and viewpoint sensitivity of subordinate-level recognition does not stand up to scrutiny.

Computational Issues

The psychophysical inquiry into feature-based recognition can greatly benefit from an interaction with a computational study of the features issue, briefly described below.

A Taxonomy of Compound Features

The possibility of extending view interpolation models to include basic-level recognition appears to depend to a large extent on the availability of a wide variety of image-based features of different diagnosticity and complexity. A basic observation regarding computational aspects of the feature issue is that any recurring image pattern whose identification, in turn, does not presuppose prior recognition of the entire object it is included in, should be considered a potential feature. This suggests that the features of recognition may form a loose hierarchy, in which higher-order or compound features are built on top of simpler ones. Prime candidates for lower-level features are simple functions of image intensity (e.g., edges, extrema, etc.; see Watt and Morgan 1985). As to the higher levels of the hierarchy, some of the possibilities for their formation are outlined below.

Subsumption by Symbolic Labeling

Marr's notion of the primal sketch (Marr 1976), in which primitive place tokens are grouped according to rules that resemble the Gestalt laws of perceptual organization,

is an early example of what may be considered a hierarchy of spatially localized features. In basic-level recognition, alignment groups of place tokens carrying relatively complex shape information can be more useful if they are assigned symbolic labels. In Ullman's example (Ullman 1989, p.233; cf. Edelman and Poggio 1989), in the representation of a rooster, the label "wiggly" could subsume the complex shape of the crown in most cases (except, possibly, when a particular rooster is to be recognized and the precise shape details matter).

Local Elaboration

When fine details of local image structure are important for the task at hand (as in the case of individual exemplar recognition), such details can be provided by computing higher-order derivatives of image intensity than the first derivative, which is strictly necessary for the detection of intensity edges. A specific computational hypothesis regarding such higher-order local features has been advanced by Koenderink and van Doorn (1990), who suggested that the various receptive fields found in the primary visual area of mammalian brain can be described as derivative operators.

Spatial Agglomeration

An alternative to local elaboration of image description through the computation of higher-order derivatives is joining together of spatially distinct features on a larger scale. A well-known example of a method that yields spatially distributed features is the Fourier transform. If one considers space instead of spatial frequency, compound features may be obtained by encoding conjunctions of the occurrences of simple features that satisfy certain constraints (e.g., constraints on type and location) or groups of place tokens (Bravo and Blake 1990). Biederman's geons (Biederman 1987) fall into this category of compound features.

Hashing

Under certain conditions (in particular, for the purpose of initial fast indexing into a large database of object representations), one may wish to define features whose only required property is that they allow efficient object access (cf. the diffuse features such as color used in (Wixson and Ballard 1990). In this case, object recognition may be likened to texture discrimination, with the individual objects regarded as textons (which must be classified statistically when present in quantity, but need not be described in detail). Linear receptive fields (Amari 1978), considered as implementations of hashing functions (cf. Breuel 1989), are a possible candidate for a biologically motivated computational approach to the extraction of indexing features. Recently, such features have been used with success in a computer scheme for the recognition of human faces (Edelman et al. 1992).

The Role of Learning

It seems plausible to assume that the visual system acquires many of the possible features of recognition through experience (psychophysical evidence for perceptual learning at the feature level can be found in Steinman [1987], Fahle et al. [1991], Karni and Sagi [1991]). Computational theory showing how this may happen could be based on recent studies of unsupervised feature extraction. This issue has been approached using tools of statistics (principal component analysis of image intensities [Sanger 1989]), exploratory projection pursuit (Intrator and Gold 1992) and information theory (maximizing mutual information between the input and the output of a layer of feature detectors (Linsker 1990), or between two separate groups of detectors (Zemel and Hinton 1991).

The scheme we are developing relies on learning also for the synthesis of object-specific recognition modules (cf. Hurlbert and Poggio 1988; Poggio et al. 1991; Poggio 1990). Learning recognition from examples has been demonstrated in (Poggio and Edelman 1990; Edelman and Weinshall 1991) for synthetic 3D objects, represented by image locations of simple features such as corners and terminators. Its extension to real-world objects may depend on the use of more appropriate features, a conclusion not unexpected in view of the preceding discussion.

Features of Recognition: A Summary

In this section we have outlined a unified approach to the understanding and modeling of human performance in object recognition. The highlights of this approach are as follows:

1. *Versatility*: Recognition starts with the extraction of a large variety of image-based features.
2. *Plasticity*: A recognition procedure for a given object at a given category level is synthesized at need and is optimized with practice.
3. *Hierarchy*: One of the ways of optimizing recognition performance involves formation of compound features out of simpler ones, and subsequent reliance on such features.
4. *Invariance/diagnosticity trade-off*: Some of the features are well-localized within the 2D image-based reference frame. Exclusive reliance on such features under certain circumstances makes recognition viewpoint-dependent. Other features, such as color, are "diffuse," which makes them relatively viewpoint-invariant. Yet another kind, the nonaccidental features, are viewpoint-invariant for geometrical reasons. There is a trade-off between viewpoint invariance of a feature and its diagnosticity, or the degree of discrimination among object instances that it affords.

GENERAL CONCLUSIONS

Modeling the subjects' response patterns in an attempt to reverse-engineer the human visual system is an integral part of our research effort. Insight gained through modeling proves to be useful both for understanding experimental results and for the planning of experiments that explore further theoretical issues.

The success of a HyperBF-based model that relied on simple arbitrary features in replicating nontrivial aspects of human performance in recognition experiments (Bülthoff and Edelman 1992) indicates that even better results can be obtained with more sophisticated feature-extraction and learning techniques. The integrated psychophysical and computational study of these issues is expected to lead to progress in three main directions:

1. *Modeling of feature extraction in human vision.* The identity and the relative importance of features discovered by the computational learning model can be compared to a psychophysical characterization of the features of recognition relied upon by human subjects.
2. *Perceptual learning in feature-based recognition.* Both the feature extraction and the classification stages of the model of recognition that emerges from our study exhibit a considerable degree of plasticity or learning. A central issue regarding the nature and role of learning in human object recognition is the distinction among the contributions of genetically determined factors and those acquired through learning. The latter can be further classified into factors, such as a common set of features, relevant to the recognition of a variety of objects, and factors that differ from object to object, such as a particular subset of the universal feature set and a particular classification method employed for a given category of objects.
3. *Unification of theoretical accounts of recognition.* It is possible that the visual system synthesizes a feature-based recognition module for a given object when it is first encountered. The details of the synthesized module, such as the types of features it relies upon and the computations applied to the features, depend on the kind of the object and on the required level of recognition. Specifically, reliance on local features such as edges or corners should allow relatively precise discrimination among similar shapes, as required for subordinate-level recognition and, at the same time, should result in viewpoint-dependent performance. On the other hand, reliance on compound features, such as geons or nonaccidental properties, should lead to viewpoint-insensitive recognition in conjunction with a limited ability for fine shape discrimination. Thus, the notion of features of recognition, of varying complexities and possessing different degrees of spatial localization, may offer a unified approach to the understanding of both basic-level and subordinate-level object recognition in human vision.

ACKNOWLEDGEMENT

This work was done at the Center for Biological Information Processing at MIT with support from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division, at the Department of Applied Mathematics and Computer Science, Weizmann Institute and at Brown University. HHB would like to thank his graduate students at Brown University—Zili Liu, Philippe Schyns, Dave Sheinberg, Erik Sklar, and Greg Zelinsky—for many discussions. Continued intellectual support by Tommy Poggio and Shimon Ullman is very much appreciated.

REFERENCES

Aha, D.W., D. Kibler, and M.A. Albert. 1991. Instance-based learning algorithms. *Machine Learning* **6**:37–66.

Amari, S. 1978. Feature spaces which admit and detect invariant signal transformations. In: Proc. 4th Intl. Conf. Pattern Recognition, pp. 452–456. Tokyo.

Biederman, I. 1985. Human image understanding: Recent research and a theory. *Comp. Vis., Graphics, Image Proc.* **32**:29–73.

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

Bravo, M., and R. Blake. 1990. Preattentive vision and perceptual groups. *Perception* **19**:515–522.

Breuel, T.M. 1989. Adaptive model base indexing. Cambridge, MA: Artificial Intelligence Lab, MIT, A.I. Memo No. 1008.

Broomhead, D.S., and D. Lowe. 1988. Multivariable functional interpolation and adaptive networks. *Complex Systems* **2**:321–355.

Bülthoff, H.H., and S. Edelman. 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci.* **89**:60–64.

Bülthoff, H.H., and H.A. Mallot. 1988. Interaction of depth modules: Stereo and shading. *J. Optical Soc. Am.* **5**:1749–1758.

Duda, R.O., and P.E. Hart. 1973. Pattern Classification and Scene Analysis. New York: Wiley.

Edelman, S. 1991a. Features of recognition. CS-TR10, Weizmann Inst. of Science.

Edelman, S. 1991b. A network model of object recognition in human vision. In: Networks for Vision, ed. H. Wechsler. New York: Academic.

Edelman, S., and H.H. Bülthoff. 1990a. Generalization of object recognition in human vision across stimulus transformations and deformations. Proc. 7th Israeli AICV Conference, ed. Y. Feldman and A. Bruckstein, pp. 479–487. Amsterdam: Elsevier.

Edelman, S., and H.H. Bülthoff. 1990b. Viewpoint-specific representations in 3D object recognition. Cambridge, MA: Artifical Intelligence Lab, MIT A.I. Memo No. 1239.

Edelman, S., and H.H. Bülthoff. 1992. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Res.*, in press.

Edelman, S., H.H. Bülthoff, and E. Sklar. 1991a. Task and object learning in visual recognition. Cambridge, MA: Artificial Intelligence Lab, MIT A.I. Memo No.1285.

Edelman, S., H.H. Bülthoff, and D. Weinshall. 1989. Stimulus familiarity determines recognition strategy for novel 3D objects. MIT A.I. Memo No. 1138.

Edelman, S., and Poggio, T. 1989. Representations in high-level vision: Reassessing the inverse optics paradigm. Proc. DARPA Image Understanding Workshop, pp. 944–949. San Mateo, CA: Morgan Kaufman.

Edelman, S., and Poggio, T. 1990. Bringing the grandmother back into the picture: A memory-based view of object recognition. MIT A.I. Memo No. 1181.

Edelman, S., D. Reisfeld, and Y. Yeshurun. 1992. Learning to recognize faces from examples. Proc. 2nd European Conf. on Computer Vision, ed. G. Sandini, vol. 588, pp. 787–791. Berlin: Springer.

Edelman, S., and D. Weinshall. 1991. A self-organizing multiple-view representation of 3D objects. *Biolog. Cybern.* **64**:209–219.

Edelman, S., D. Weinshall, H. Bülthoff, and T. Poggio. 1990. A model of the acquisition of object representations in human 3D visual recognition. Proc. NATO Advanced Research Workshop on Robots and Biological Systems, ed. P. Dario, G. Sandini, and P. Aebischer. New York: Springer.

Ellis, R., D.A. Allport, G.W. Humphreys, and J. Collis. 1989. Varieties of object constancy. *Q. J. Exp. Psychol.* **41A**:775–796.

Hurlbert, A., and T. Poggio. 1988. Synthesizing a color algorithm from examples. *Science* **239**:482–485.

Intrator, N., and J.I. Gold. 1992. Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. *Neural Comp.*, in press.

Jolicoeur, P. 1985. The time to name disoriented objects. *Memory Cogn.* **13**:289–303.

Karni, A., and D. Sagi. 1991. Where practice makes perfect in texture discrimination. *Proc. Natl. Acad. Sci.* **88**:4966–4970.

Koenderink, J.J., and A.J. van Doorn. 1990. Receptive field families. *Biolog. Cybern.* **63**:291–297.

Koriat, A., and J. Norman. 1985. Mental rotation and visual familiarity. *Perception and Psychophysics* **37**:429–439.

LaBerge, D. 1976. Perceptual learning and attention. In: Handbook of Learning and Cognitive Processes, ed. W.K. Estes, vol. 4, pp. 237–273. Hillsdale, NJ.: Erlbaum

Linsker, R. 1990. Perceptual neural organization: Some approaches based on network models and information theory. *Ann. Rev. Neurosci.* **13**:257–281.

Lowe, D.G. 1986. Perceptual Organization and Visual Recognition. Boston: Kluwer Academic.

Marr, D. 1976. Early processing of visual information. *Phil. Trans. R. Soc. Lond. B* **275**:483–524.

Marr, D. 1982. Vision. San Francisco: W.H. Freeman.

Marr, D., and H.K. Nishihara. 1978. Representation and recognition of the spatial organization of 3D structure. *Proc. Roy. Soc. Lond. B* **200**:269–294.

Moody, J., and C. Darken. 1989. Fast learning in networks of locally tuned processing units. *Neural Comp.* **1**:281–289.

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

Pentland, A. 1988. Shape information from shading: A theory about human perception. Proc. 2nd Intl. Conf. on Computer Vision, pp. 404–413, Tarpon Springs, FL. Washington, DC: IEEE.

Poggio, T. 1990. A theory of how the brain might work. *Cold Spring Harbor Symp. Quant. Biol.* **LV**:899–910.

Poggio, T., and S. Edelman. 1990. A network that learns to recognize 3D objects. *Nature* **343**:263–266.

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

Poggio, T., M. Fahle, and S. Edelman. 1991. Synthesis of visual modules from examples: Learning hyperacuity. Cambridge, MA: Artificial Intelligence Lab, MIT A.I. Memo No. 1271.

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

Rock, I., and J. DiVita. 1987. A case of viewer-centered object perception. *Cog. Psych.* **19**:280–293.

Rock, I., D. Wheeler, and L. Tudor. 1989. Can we imagine how objects look from other viewpoints? *Cog. Psych.* **21**:185–210.

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

Sanger, T. 1989. Optimal unsupervised learning in feedforward neural networks. Cambridge, MA: Artificial Intelligence Lab, MIT A.I. Memo No. 1086.

Shepard, R.N., and L.A. Cooper. 1982. Mental images and their transformations. Cambridge, MA: MIT Press.

Steinman, S.B. 1987. Serial and parallel search in pattern vision? *Perception* **16**:389–398.

Tarr, M., and S. Pinker. 1989. Mental rotation and orientation-dependence in shape recognition. *Cog. Psych.* **21**:233–282.

Tarr, M., and S. Pinker. 1990. When does human object recognition use a viewer-centered reference frame? *Psych. Sci.* **1**:253–256.

Thompson, D.W., and J.L. Mundy. 1987. 3D model matching from an unconstrained viewpoint. Proc. IEEE Conf. on Robotics and Automation, pp. 208–220, Raleigh, NC.

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

Ullman, S., and R. Basri. 1990. Recognition by linear combinations of models. Cambridge, MA: Artificial Intelligence Lab, MIT A.I. Memo No. 1152.

Watt, R.J., and M.J. Morgan. 1985. A theory of primitive spatial code in human vision. *Vision Res.* **25**:1661–1674.

Wixson, L.E., and D.H. Ballard. 1990. Real-time qualitative detection of multi-colored objects for object search. Proc. AAAI–90 Workshop on Qualitative Vision, pp. 46–50. San Mateo, CA: Morgan Kaufmann.

Zemel, R.S., and G.E. Hinton. 1991. Discovering viewpoint-invariant relationships that characterize objects. In: Neural Information Processing Systems, ed. D. Touretzky, vol. 3. San Mateo, CA: Morgan Kaufmann.