MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1479 April, 18 1994
C.B.C.L. Paper No. 96

How are three-dimensional objects represented in the brain?

Heinrich H. Bülthoff, Shimon Y. Edelman & Michael J. Tarr

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1994/AIM-1479.ps.Z

Abstract

We discuss a variety of psychophysical experiments that explore different aspects of the problem of object recognition and representation in human vision. In all experiments, subjects were presented with realistically rendered images of computer-generated three-dimensional objects, with tight control over stimulus shape, surface properties, illumination, and viewpoint, as well as subjects' prior exposure to the stimulus objects. Contrary to the predictions of the paradigmatic theory of recognition, which holds that object representations are viewpoint invariant, performance in all experiments was consistently viewpoint dependent, was only partially aided by binocular stereo and other depth information, was specific to viewpoints that were familiar, and was systematically disrupted by rotation in depth more than by deforming the two-dimensional images of the stimuli. The emerging concept of multiple-views representation supported by these results is consistent with recently advanced computational theories of recognition based on view interpolation. Moreover, in several simulated experiments employing the same stimuli used in experiments with human subjects, models based on multiple-views representations replicated many of the psychophysical results concerning the observed pattern of human performance.

This report describes research done at the Center for Biological and Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. This research is sponsered by grants from the Office of Naval Research under contracts N00014-92-J-1879 and N00014-93-1-0385. Support for the Center is provided in part by a grant from the National Science Foundation under contract ASC–9217041 (funds provided by this award include funds from DARPA provided under the HPCC program) and by a grant from the National Institutes of Health under contract NIH 2-S07-RR07047. Support for the laboratory's artificial intelligence research is provided by ARPA contract N00014-91-J-4038. Heinrich H. Bülthoff is now at the Max-Planck-Institut für biologische Kybernetik, D-72076 Tübingen, Germany; Shimon Edelman is at the Dept. of Applied Mathematics and Computer Science, Weizmann Insitute of Science, Rehovot 76100, Israel and Michael J. Tarr at the Department of Psychology, Yale University, New Haven, CT 06520-8205. SE was supported by the Basic Research Foundation, administered by the Israel Academy of Arts and Sciences. MJT was supported by the Air Force Office of Scientific Research, contract number F49620-91-J-0169, and the Office of Naval Research, contract number N00014-93-1-0305.

1 Introduction

How does the human visual system represent three-dimensional objects for recognition? Object recognition is carried out by the human visual system with such expediency that to introspection it normally appears to be immediate and effortless (Fig. 1 – CANONICAL). Computationally, recognition of a three-dimensional object seen from an arbitrary viewpoint is complex because its image structure may vary considerably depending on its pose relative to the observer (Fig. 1 – NON-CANONICAL). Because of this variability across viewpoint, simple two-dimensional template matching is unlikely to account for human performance in recognizing three-dimensional objects, since it would require that a discrete template be stored for each of the infinite number of view-specific images that may arise for even a single object. Consequently, the most prominent computational theories of object recognition (see Ullman, 1989 for a survey) have rejected the notion of view-specific representations. Other approaches, rooted in pattern recognition theory, have postulated that objects are represented as lists of viewpoint-invariant properties or by points in abstract multidimensional feature spaces (Duda and Hart, 1973). Another, more commonly held, alternative is characterized by the postulate that objects are represented by three-dimensional viewpoint-invariant part-based descriptions (Marr and Nishihara, 1978; Biederman, 1987), similar to the solid geometrical models used in computer-aided design.

Surprisingly, theories that rely on viewpoint-invariant three-dimensional object representations fail to account for a number of important characteristics of human performance in recognition. In particular, across a wide range of tasks, recognition performance, as measured by response times and error rates, has been found to vary systematically with the viewpoint of the perceiver relative to the target object. Such results provide converging evidence in favor of an alternative theory of recognition, which is based on multiple viewpoint-specific, largely two-dimensional representations To support this interpretation of the psychophysical results, we review briefly several computational theories of object recognition, each of which generates specific behavioral predictions that the experiments were designed to test. Many 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 two-dimensional representations. 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, 1991; Poggio and Edelman, 1990; Bülthoff and Edelman, 1992; Edelman and Weinshall, 1991).

2 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. More than any other single issue, the question of whether object representations are viewpoint invariant or viewpoint dependent has been identified as the crucial distinction on which theories of recognition stand or fall.

One can use the viewpoint-invariant/viewpoint-dependent distinction to make specific psychophysical predictions as follows. Intuitively, if the representation is viewpoint invariant, and if an object-centered reference frame can be recovered independently of object pose, then neither recognition time nor accuracy should be related to the viewpoint of the observer with respect to the object. In contrast, if the representation is viewpoint dependent, and as long as the complexity of the normalization procedure scales with the magnitude of the transformation, then both recognition time and accuracy should be systematically related to the viewpoint of the observer with respect to the object. Subtler predictions may be derived from a closer examination of specific theories.

2.1 Theories that rely on three-dimensional object representations

Theories of the first kind we mention attempt to achieve a computer-vision equivalent of complete object constancy, the apparent ability of humans to perceive and recognize three-dimensional objects irrespective of factors such as viewpoint (Ellis et al., 1989). Two major approaches to object constancy can be discerned. The first approach uses fully three-dimensional viewpoint-invariant representations, and requires that a similar three-dimensional representation of the input be recovered from the image before it is matched to like-representations in visual memory. The second approach uses viewpoint-specific three-dimensional representations (e.g., selected views that include depth information), and requires that three-dimensional representations of the input be normalized (by an appropriate spatial transformation) from the viewpoint of the image to the viewpoint of a view-specific representation in visual memory.

2.1.1 Viewpoint-invariant three-dimensional representations

The notion that the processing of the visual input culminates in a full restoration of its three-dimensional structure which may then be matched to three-dimensional viewpoint-invariant representations in memory was popularized by Marr and Nishihara (1978). Representation by reconstruction, which became known in computer vision under the name of intrinsic images (Barrow and Tenenbaum, 1978; Tenenbaum et al., 1981), was never implemented, due to persistent difficulties in solving the problem of a general reconstruction of the three-dimensional representation from input images. Despite the failure of this approach in computer vision, in psychology it has become widely accepted as a plausible model of recognition, following the work of Biederman and his associates.

Biederman's theory, known as Recognition By Com-

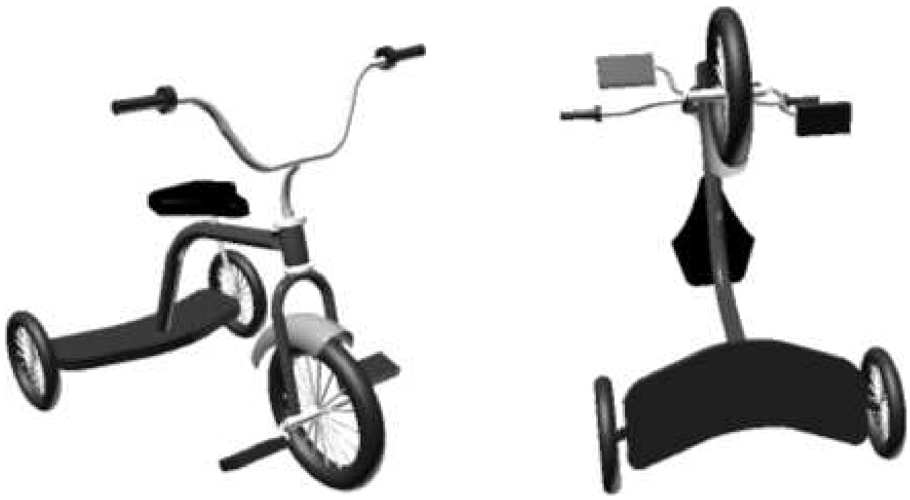

Figure 1: Canonical views: certain views of three-dimensional objects are consistently easier to recognize or process in a variety of visual tasks. Once this object is identified as a tricycle seen from the front, we find it difficult to believe its recognition was anything less than immediate. Nevertheless, recognition is at times prone to errors, and even familiar objects take longer to recognize if they are seen from unusual (non-canonical) viewpoints (Palmer et al., 1981). Exploring this and other related phenomena can help elucidate the nature of the representation of three-dimensional objects in the human visual system.

ponents (or more recently, Geon Structural Descriptions, or GSD (Hummel and Biederman, 1992)), postulates that the human visual system represents basic-level object categories by three-dimensional structural relationships between a restricted class of volumetric primitives known as "geons" (Biederman, 1987). The crucial property of the GSD approach is that the part descriptions upon which object representations are built are qualitative – the same object representation is derived, regardless of viewpoint, so long as the same configuration of perceptual features is present in the image. A consequence of this is that GSDs actually exhibit only *view-restricted* invariance in that a change in the visibility or occlusion of parts will alter the feature configurations present in the image (Hummel and Biederman, 1992; Biederman and Gerhardstein, 1993). Therefore, the representation of a single object will necessarily include several characteristic (Freeman and Chakravarty, 1980) or qualitative views, each composed of a distinct GSD and each viewpoint-invariant only for a limited range of viewpoints.

2.1.2 Viewpoint-specific three-dimensional 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 instance (Ullman, 1989). In the alignment approach the two-dimensional input image is compared with the projection of a stored three-dimensional 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 complete three-dimensional model. The aligning transformation is computed separately for each of the models stored in visual memory (but only one per object). The outcome of the recognition process is the model whose projection matches the input image most closely after the two are aligned. Related schemes (Lowe, 1986; Thompson and Mundy, 1987) select the most appropriate model in visual memory by using the "viewpoint consistency constraint" which projects each model to a hypothesized viewpoint and then relates the projected locations of the resultant image features to the input image, thereby deriving a mapping of the image to the three-dimensional structure of stored object representations.

Ullman (1989) distinguishes between a full alignment scheme that employs complete three-dimensional models and attempts to compensate for all possible three-dimensional transformations that objects may undergo, such as rotation in depth, and a partial alignment scheme that employs pictorial descriptions that decompose objects into (non-generic) parts and uses multiple views rather than a single viewpoint-invariant description to compensate for some three-dimensional transformations. Ullman notes (*ibid.*, p.228) that this latter multiple-views approach to alignment involves a 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 is "view-insensitive, since the differences between views are partially compensated by the alignment process." As such, this approach is similar to Biederman's (Hummel and

Biederman, 1992) most recent version of GSD theory in which multiple viewpoint-invariant GSDs are used to represent a single object (although because GSDs are considered to be qualitative descriptions, no alignment process is ever postulated to compensate for differences in viewpoint). Regardless of these subtle differences, both versions of alignment theory (hereafter referred to simply as alignment) may include the assumption that normalization procedures do not depend on the magnitude of the transformation – consequently, viewpoint-invariant performance in recognition tasks (e.g., response times and error rates) may be considered their central distinguishing feature. Alternatively, the complexity of normalization may scale with the magnitude of transformation, and as such, viewpoint-invariant performance is predicted only for error rates, with viewpoint-dependent patterns predicted for response times.

2.2 Theories that rely on viewpoint-dependent two-dimensional object representations

Theories of the second kind we mention here each attempt to achieve object constancy by storing multiple two-dimensional viewpoint-specific representations (e.g., image-based views) and including mechanisms for matching input images to stored views or to views derived computationally from stored views. While the specific mechanisms postulated for accomplishing this match vary among theories (and have consequences for the subtler predictions of each), they may all be considered as computational variants of the empirically-based multiple-views-plus-transformation (MVPT) theory of recognition (Tarr and Pinker, 1989). MVPT postulates that objects are represented as linked collections of viewpoint-specific images ("views"), and that recognition is achieved when the input image activates the view (or set of views) that corresponds to a familiar object transformed to the appropriate pose. There is evidence (Edelman and Weinshall, 1991; Tarr, 1989; Tarr and Pinker, 1989) indicating that this process can result in the same dependence of the response time on the pose of the stimulus object as obtained in the mental rotation experiments (Shepard and Cooper, 1982). We consider MVPT as a psychological model of human performance that predicts recognition behavior under specific conditions; the computational models reviewed below provide details on how this performance may be achieved.

2.2.1 Linear combination of views (LC)

Several recently proposed approaches to recognition dispense with the need to represent objects as three-dimensional models. The first of these, recognition by linear combination of views (Ullman and Basri, 1991), is built on the observation that, under orthographic projection, the two-dimensional coordinates of an object point can be represented as a linear combination of the coordinates of the corresponding points in a small number of fixed two-dimensional views of the same object. The required number of views depends on the allowed three-dimensional transformations of the objects and on the representation of an individual view. For a polyhedral object that can 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, 1991). A system that relies solely on the linear combination approach (LC) 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.

2.2.2 View interpolation by basis functions (HyperBF)

Another approach that represents objects by sets of two-dimensional views is view interpolation by regularization networks (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 stage 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 linear combination 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 (see (Poggio and Edelman, 1990; Edelman and Poggio, 1992) for details). 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.

2.2.3 Conjunction of localized features (CLF)

The third scheme we mention is also based on interpolation among two-dimensional 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; Tarr, 1989; Tarr and Pinker, 1989). 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 three-dimensional 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.

3 Recognition behavior as predicted by the different theories

3.1 Experimental issues

A wide range of psychophysical experiments have been reported that assess the impact of changes of viewpoint on the recognition of both familiar and novel stimuli. The core issue in all such studies is whether response times and/or error rates are equivalent for all changes in viewpoint or are systematically related to the magnitude of changes in viewpoint. Such behavioral patterns can help to decide which representations (viewpoint-invariant or viewpoint-dependent) are used in object recognition. However, one must be cautious in interpreting such patterns – there are instances of both viewpoint-invariant and viewpoint-dependent behavior that do not necessarily imply correspondingly viewpoint-invariant or viewpoint-dependent representations. In particular, there is an asymmetry in what may be concluded from viewpoint-invariant patterns of responses. For novel objects, because of the limited stimulus set sizes employed in many experiments, a viewpoint-invariant pattern may simply indicate that in the context of the experimentally defined recognition set, subjects were able to recognize objects via localized viewpoint-invariant features within each object (Eley, 1982). In contrast, in the context of all potentially recognizable objects in the world, such features would not be unique and consequently would not support viewpoint-invariant recognition. Thus, one of the many challenges that must be overcome in assessing recognition mechanisms in humans is the development of novel stimuli that do not facilitate the reliance on unique features (to the extent that such features are unlikely to be unique in the real world). A similar problem of interpretation exists for familiar objects: a viewpoint-invariant pattern may arise as a result of multiple familiar stored views (distributed across viewpoint so as to mask most effects of viewpoint). Thus, another challenge that must be overcome is how to assess the possible existence of multiple-views in cases where objects are very familiar, presumably leading to the instantiation of many views.

Examples of difficulties of interpretation may also be found in patterns of performance that are viewpoint-dependent. For instance, initial viewpoint-dependency for novel objects may occur because viewpoint-invariant representations may arise only over experience. Thus, learning processes must be considered in assessing recognition. Viewpoint-dependent patterns may arise because of reliance on perceptual information possibly irrelevant to recognition – for example, mirror-image discrimination requires left/right handedness information defined in only in our ego-centric frame of reference, therefore, mental rotation is apparently used to normalize objects to this frame (Shepard and Cooper, 1982). Thus, a final challenge is to ensure that extraneous factors, for instance, handedness, do not produce behavioral patterns that are not typical of recognition judgments. As discussed in Sections 4.3 and 5, these challenges are addressed in experiments conducted by Bülthoff and Edelman (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992a) and by Tarr (Tarr, 1989; Tarr and Pinker, 1989). Briefly, these experiments employed the following manipulations:

- Novel stimulus objects that shared similar parts in different spatial relationships (typical of subordinate-level recognition discriminations), thereby reducing the possibility of localized unique features mediating recognition (see Fig. 2).
- Measures assessing both the initial recognition of novel objects and recognition following extensive familiarization.
- Restricted sets of viewpoints during initial training or other controls (see below) to investigate the degree of viewpoint specificity encoded in object representations of familiar objects or novel objects following extensive familiarization.
- The introduction of unfamiliar "test" views to assess the underlying organization of views instantiated during learning.
- Recognition tasks that reduced the likelihood of extraneous influences on recognition performance. For instance, some studies controlled for handedness by using bilaterally symmetrical objects or treating both members of mirror-pairs as equivalent.

Additionally, to differentiate between the more subtle predictions of viewpoint-dependent theories of recognition, we have investigated the 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, test views re shown during training). In the second case, the test views are unfamiliar, but are related to the training views through a rigid three-dimensional 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. To better place the results of such experiments in a theoretical context, we first review the specific theoretical predictions generated by each theory of recognition.

3.2 Theoretical predictions

The theories discussed in Section 2 make different predictions about the effect of factors such as viewpoint on the accuracy and latency of recognition under the various conditions outlined above. As mentioned, at the most general level, theories that rely on viewpoint-invariant representations predict no systematic effect of viewpoint on either response times or error rates, both for familiar and for novel test views, provided that the representational primitives (i.e., invariant features or generic parts) can be readily extracted from the input image. In comparison, theories that rely on viewpoint-dependent representations naturally predict viewpoint-dependent performance. However, the details of such predictions vary according to the specifics of the approach postulated by each particular theory.

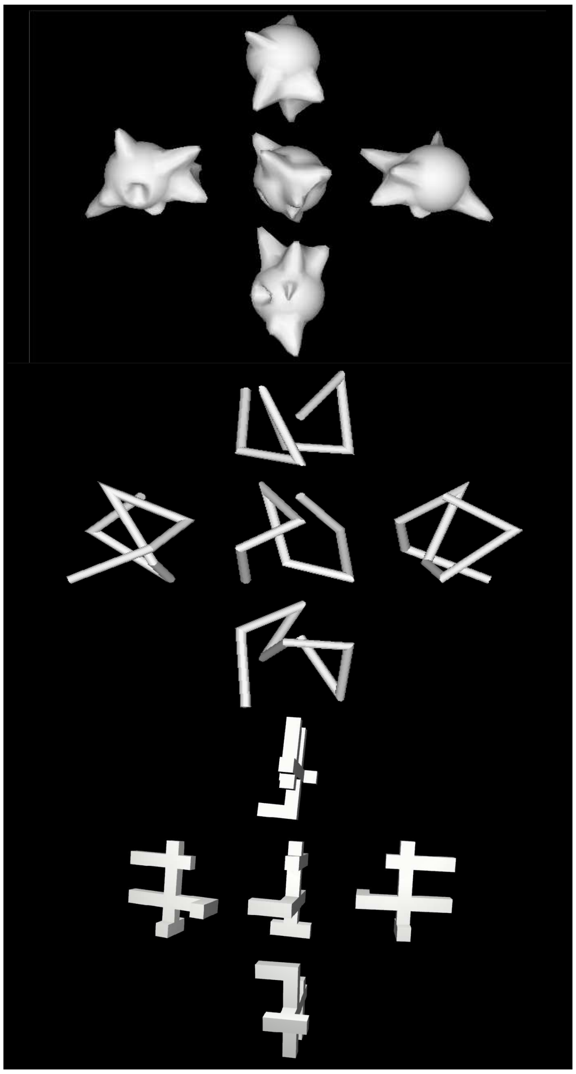

Figure 2: The appearance of a three-dimensional object can depend strongly on the viewpoint. The image in the center represents one view of a computer graphics object (wire-, amoeba-, or cube-like). The other images are derived from the same object by $\pm 75°$ rotation around the vertical or horizontal axis. The difference between the images illustrates the difficulties encountered by any straightforward template matching approach to three-dimensional 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.

3.2.1 Viewpoint-invariant three-dimensional representations

A recognition scheme based on viewpoint-invariant three-dimensional 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 viewpoint-invariant description is to be formed (Marr and Nishihara, 1978; 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 qualitative spatial relations.

Similar predictions are made by the most recent version of GSD theory (Biederman and Gerhardstein, 1993; Hummel and Biederman, 1992) to the extent that given GSD is considered to be viewpoint invariant up to changes in the visibility or occlusion of specific geons. Therefore, as long as the complete set of GSDs is familiar for a given object, recognition behavior will be completely viewpoint invariant. However, under conditions where some GSDs are unfamiliar or, more generally, under conditions where the GSD recovered from an image must be matched to a different GSD in memory, recognition behavior will degrade qualitatively, that is, without any systematic relationship to the magnitude of changes in viewpoint (Biederman and Gerhardstein, 1993). Thus, GSD theory predicts viewpoint invariance for the recognition of familiar objects and only step-like viewpoint-dependent patterns for the recognition of unfamiliar objects undergoing extreme changes in visible part structure.

3.2.2 Viewpoint-dependent three-dimensional 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. As mentioned, if the recognition system represents an object by multiple views and uses an incremental transformation process for viewpoint normalization, response times are expected to vary monotonically with the viewpoint 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 times on viewpoint, 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 (thereby leading to successful recognition). Alternatively, if normalizing or transforming the object uses a "one-shot" transformation process for viewpoint normalization, response times will likewise be viewpoint invariant. In either case, the predictions of this theory may be differentiated from theories that rely on two-dimensional representations and normalization procedures in that the latter predict effects of viewpoint for both response times and error rates (as discussed in the following sections). By comparison, theories based on three-dimensional representations predict that error rates will not vary with viewpoint (regardless of the pattern of response times).

3.2.3 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 three-dimensional viewpoint-invariant 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 two-dimensions and cannot accommodate 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, 1991). 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.

3.2.4 View interpolation

Finally, consider the predictions of the view interpolation theory. First, as with theories that rely on three-dimensional representations, effects of viewpoint on response times are expected to vary with specific implementation details. In one instance, there will be no systematic increase in response times with changes in viewpoint if the transformation (in this case, interpolation) mechanism is "one-shot" instead of incremental. In the other instance, response times will increase with increasing changes in viewpoint if the interpolation involves an incremental process, for example, a time-consuming spread of activation in a distributed implementation.

We note that while activation-spread models have been proposed as accounts of viewpoint-dependent response times in object recognition (Edelman and Weinshall, 1991), they may also offer a plausible mechanism for many so-called mental transformation phenomena. For instance, it is well documented that at the behavioral level, humans employ a transformation process commonly referred to as "mental rotation" during some perceptual judgments (Shepard and Cooper, 1982). The explanation offered by Shepard is that such transformations are mental analogs of actual physical transformations – a hypothesis which still stimulates a major debate in cognitive science, but does not seem to lead to a plausible neural or computational theory. In its stead, we propose that, to the extent that a given theory of view interpolation relies on an incremental process, it may provide a plausible account of mental transformation behavioral patterns across many tasks.[1]

Another prediction of the view interpolation theory is lower error rate for familiar test views than for novel test views, depending on the distance from the novel view to the nearest familiar stored view. 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.

4 Psychophysical background

4.1 Basic vs. subordinate-level recognition

Numerous studies in cognitive science (see Rosch et al., 1976 for a review) reveal that in the hierarchical structure of object categories there exists a level of category organization, referred to as the *basic level*, which is the most salient according to a variety of psychological 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". While basic-level categorical structure is unlikely to a product of either purely definitional or perceptual mechanisms (Armstrong et al., 1983), there is some evidence that basic-level categories are organized to some extent around perceptual properties of objects. For instance, Tversky and Hemenway (1984) have proposed that the presence of common parts in similar configurations is one of the essential properties in determining category membership. However, given this conjecture, it is clear that some apparent members of a particular basic-level category are inappropriate. For example, while robins, bluejays, and penguins all share membership in the category "bird," only the first two actually share many common parts. Both the shape and consequently the parts of penguins are dissimilar to prototypical birds. Likewise, in terms of naming performance, it is clear that the basic level fails to capture some aspects of categorization behavior; for example, the first label assigned to an image of a penguin is likely to be "penguin" rather than "bird" –

[1] Indeed, a view interpolation account of Tarr's data on object recognition supports this proposal. Tarr (1989; Tarr and Pinker, 1989) compared directly the response time patterns obtained in recognition tasks to those obtained using identical objects in identical viewpoints in perceptual judgments known to elicit the use of mental transformations. The comparison revealed that recognition and transformation tasks yield highly similar putative rates of "rotation" as well as deviations from monotonicity. While such evidence is necessarily only circumstantial, it provides some indications that well-specified computational theories of recognition may also inform us as to the mechanisms used in other aspects of visual cognition.

a behavior consistent with the dissociation at the perceptual level. Consequently, it has been suggested that for purposes of characterizing recognition performance, the basic level should be supplanted by the *entry level* – the first categorical label generally assigned to a given object (Jolicoeur et al., 1984). To the extent that theories of recognition attempt to account for classificatory behavior, they do so for entry-level performance (i.e., Biederman, 1987; Hummel and Biederman, 1992).

In contrast to the entry-level, objects whose recognition implies finer distinctions than those required for entry-level categorization are said to belong to a *subordinate level*. In terms of perceptual content, the subordinate level may be characterized by objects having similar overall shape as a consequence of sharing similar parts in similar spatial relationships. Typical examples of subordinate-level or within-category discriminations include recognizing individual faces or specific models of cars.

Crucially, 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 (Edelman, 1992). Specifically, if the subjects are required to *classify* the stimulus (that is, to determine its entry-level category), error rates and response times are often found to be viewpoint invariant (except in instances where the three-dimensional structure of the object is severely distorted, e.g., due to foreshortening; see Biederman 1987). In contrast, if the task is to *identify* a specific object (that is, to discriminate one individual from other, visually similar objects sharing parts and spatial relations), error rates and response times are normally viewpoint dependent. While this distinction is certainly true in its extreme form (for instance, objects having no parts in common will almost certainly be members of different entry-level categories and, likewise, may be discriminated by viewpoint-invariant unique features) it is less clear that "everyday" entry-level performance is mediated by viewpoint-invariant mechanisms. For example, as discussed in the following section, naming times (generally at the entry-level) for familiar common objects have been found to be viewpoint-dependent. More importantly, because entry-level categories are only acquired over extensive experience with many instances of each class, it is possible that multiple viewpoint-dependent representations are acquired as the category is learned. As discussed in Section 3.1, this leads to an asymmetry in the kind of conclusions that can be drawn from viewpoint-invariant performance: for familiar entry-level categories, the reliance on multiple views may mask the operation of any viewpoint-dependent mechanisms. Thus, it is difficult to assess the underlying structure of object representations through entry-level tasks employing familiar objects as stimuli. To address this problem, we are currently undertaking several psychophysical studies in which the acquisition of entry-level categories for novel objects is manipulated in conjunction with viewpoint. To the extent that entry-level categorization is normally viewpoint-invariant, such performance should be found regardless of which views have been displayed; alternatively, to the extent that entry-level categorization relies on multiple-views, performance should vary systematically in relation to the views that are familiar.

4.2 Canonical views

Most familiar common objects such as houses, animals, or vehicles are recognized faster or more slowly, depending on the viewpoint of the observer (as demonstrated in Figure 1). This phenomenon has been defined originally purely in descriptive and qualitative terms. For instance, Palmer, Rosch and Chase (1981) found that subjects consistently labeled one or two views, designated as *canonical views*, of such objects as subjectively "better" than all other views. Consistent with such ratings, a naming task revealed that subjects tended to respond fastest when the stimulus was shown in a canonical view (as determined independently in the aforementioned subjective judgment experiment), with response times increasing monotonically with changes in viewpoint relative to this view. This demonstration of viewpoint-dependent naming is consistent with the hypothesis that multiple-views mediate recognition even at the entry-level; in particular, theories of recognition that rely on viewpoint-specific representations may accommodate such results quite naturally, while theories that rely on viewpoint-invariant representations will require added complexity solely to account for this behavior. It should be noted however, that at the entry 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 as response times, where they constitute strong and stable evidence in favor of viewpoint-dependent nature of object representations (see Section 5.1). Thus, while entry-level and subordinate-level recognition may share some common representational structures, they may differ at some level of processing, for instance, in the threshold for what constitutes a correct match.

4.3 Mental transformation and its disappearance with practice

As discussed in Section 3.1, the body of evidence documenting the monotonic dependency of recognition time on viewpoint has been interpreted recently (Tarr, 1989; Tarr and Pinker, 1989; Tarr and Pinker, 1990) as an indication that objects are represented by a few specific views, and that recognition involves viewpoint normalization (via alignment, linear combinations, or HyperBF's) to the nearest stored view, by a process similar 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, 1989; Tarr and Pinker, 1989). Thus, experience with many viewpoints of an object leads to apparent viewpoint invariance. However, to reiterate the point made in Section 3.1, such performance is not diagnostic in that it may arise as a result of either multiple-views or as a single viewpoint-invariant representation.

To distinguish between these two possibilities, Tarr and Pinker (1989; also Tarr, 1989) investigated the effect of practice on the pattern of responses 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. Specifically, their experiments included three phases: training, practice, and surprise. Feedback about the correctness of their responses was provided to subjects in all phases. During training, subjects learned to identify three or four novel objects from a single viewpoint. Crucially, the stimulus objects shared similar parts in different spatial relationships, a perceptual discrimination characteristic of subordinate-level recognition. To assess the initial effects of changes of viewpoint on recognition, during practice, subjects named the objects in a small select set of viewpoints.[2] Consistent with the hypothesis that objects are recognized by a normalization to viewpoint-specific two-dimensional representations, initial naming times and accuracy were both monotonically related to the change in viewpoint (a finding also consistent with the results of Palmer, et. al., 1981, and Jolicoeur, 1985). In particular, the magnitude of this effect was comparable in terms of putative rate of rotation (as measured by the slope of the response time function) to the rates found in classic studies of mental rotation (Shepard and Cooper, 1982) and to control experiments in which the same novel stimuli were discriminated on the basis of left/right handedness in the identical viewpoints. However, as expected, this effect of viewpoint diminished to near equivalent performance at all familiar viewpoints with extensive practice. At this point, the surprise phase was introduced, during which subjects named the same now-familiar objects in new, never-before-seen viewpoints as well as in previously practiced familiar viewpoints (see Fig. 3).

The surprise phase manipulation is diagnostic for distinguishing between viewpoint-invariant and viewpoint-dependent theories in that the former class of theories predict that the mechanisms used to achieve invariance for the familiar viewpoints may be used to recognize stimuli independent of viewpoint in the unfamiliar viewpoints as well; in contrast, the latter class of theories predict that no such generalization will occur, rather, the viewpoint-dependent mechanisms used to match stimuli to stored familiar views will now necessitate that stimuli in unfamiliar views be normalized with stored views. Consistent with this latter prediction, numerous experiments have revealed patterns in both response times and error rates that vary monotonically with the distance between the unfamiliar viewpoint and the *nearest* familiar view (Fig. 3). Importantly, the magnitude of such effects was comparable to the viewpoint effects found in the initial practice phase of each experiment – indicating that the same viewpoint-dependent mechanism was employed both when the stimuli were relatively novel and when they were highly familiar (the crucial difference being the number of views encoded per object). Indeed, as before, further experience with a wide range of views (all of the viewpoints in the surprise phase) once again led to a dimunition in the effect of viewpoint on performance for those specific viewpoints, presumably because additional views were acquired with experience. Similar findings have been observed under numerous stimulus manipulations that controlled for the possibility that effects of viewpoint were the result of superfluous handedness checks, including experiments employing bilaterally symmetrical objects and cases where mirror-image pairs were treated as equivalent. Overall, these results provide strong evidence that, at least for purposes of subordinate-level recognition, objects are represented as viewpoint-specific multiple-views and recognized via viewpoint-dependent normalization processes.

4.4 Limited generalization

The pattern of error rates in 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 of the bent wire-forms used as stimuli), the error rate reaches chance level already at a misorientation of about 40° relative to a familiar attitude (Rock and DiVita, 1987), see also Figure 6. A similar limitation seems to hold for people's ability to imagine the appearance of such wire-forms from unfamiliar viewpoints (Rock, Wheeler and Tudor, 1989). However, such results may present an extreme case in terms of performance. Farah (Farah et al., 1994) observed that when Rock's wire-forms were interpolated with a smooth clay surface (creating "potato-chip" objects), subjects' recognition accuracy increased dramatically for changes in viewpoint equivalent to those tested by Rock. Thus, object shape and structure plays a significant role in the ability of humans to compensate for variations in viewpoint (for instance, see Koenderink and van Doorn, 1979). One possibility is that as the structure of objects becomes more regular (in terms of properties such as spatial relations and symmetries), the ability to compensate efficiently for changes in viewpoint is enhanced, in that the resultant image structure is predictable (Vetter et al., 1994). One consequence is that error rates may be reduced and performance will be enhanced, although it is possible that mixed strategies or verification procedures will yield response times that are still dependent on viewpoint (as seen in the naming of familiar common objects in non-canonical views, Palmer, et. al., 1981).

5 Psychophysics of subordinate-level recognition

Despite the availability of data indicating that multiple-views and normalization mechanisms play some role in subordinate-level recognition (Section 4.3), psychophysical research has left many of the questions vital to computational understanding of recognition unanswered.

[2] To ensure that subjects did not rely on unique features, several "distractor" objects were also included. Rather than naming such objects, subjects simply made a "none-of-the-above" response.

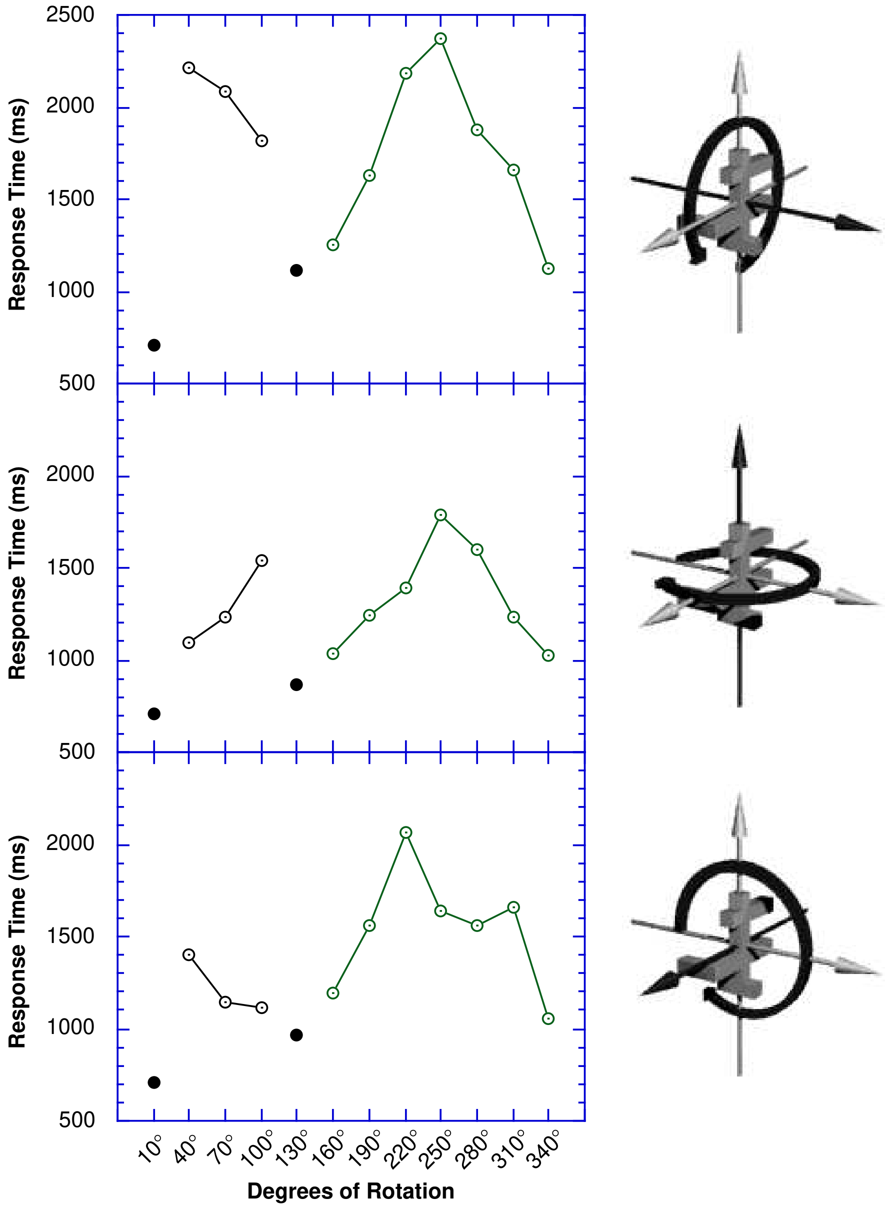

Figure 3: Mean response times for correctly naming familiar "cube" objects in familiar and unfamiliar viewpoints. Viewpoints were generated by rotations in depth (around the x or y axis) or in the picture-plane (around the z axis). Filled data points represent familiar viewpoints learned during training and extensive practice; open points represent unfamiliar viewpoints introduced in the "surprise" phase of the experiment. Prior to this phase, extensive practice resulted in the onset of equivalent naming performance at all familiar viewpoints – a pattern consistent both with the acquisition of multiple viewpoint-dependent "views" and with the acquisition of a single viewpoint-invariant description. Performance in the surprise phase distinguishes between these two possibilities: naming times (and error rates) increased systematically with angular distance from the nearest familiar viewpoint, indicating that subjects represented familiar objects as multiple-views and employed a time-consuming normalization process to match unfamiliar viewpoints to familiar views. One of the 7 "cube" objects is shown along with the axis of rotation to the right of each plot (data and stimuli adapted from Tarr, 1989).

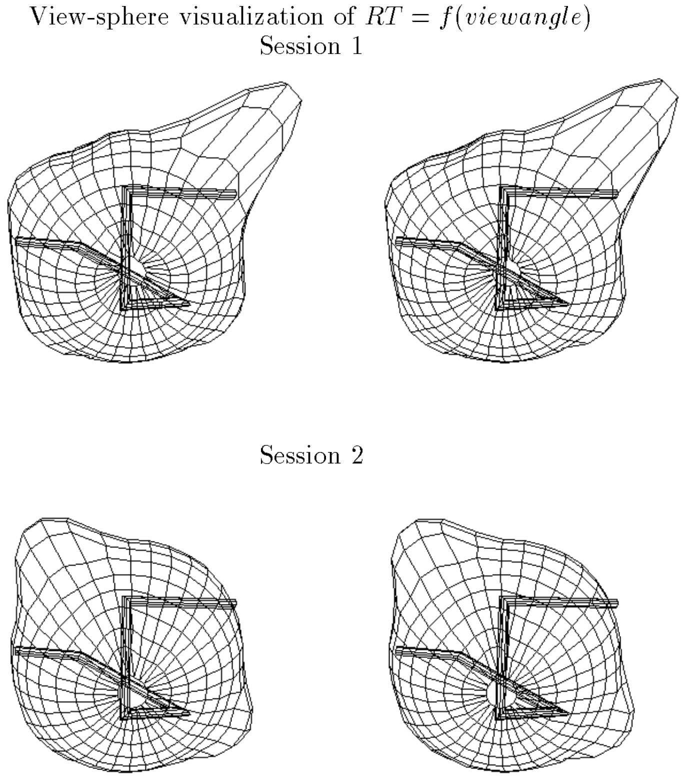

Figure 4: Canonical views and practice: the advantage of some views over others, as manifested in the pattern of response times (RTs) to different views of wire-like objects, is reduced with repeated exposure. The spheroid surrounding the target is a three-dimensional stereo-plot of response time vs. aspect (local deviations from a perfect sphere represent deviations of response time from the mean). The three-dimensional 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 Session 2. 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. Adapted from Edelman and Bülthoff, 1992.

For example, it is still unclear whether the canonical views phenomenon reflects basic viewpoint dependence of recognition, or is due to particular patterns of the subjects' exposure to the stimuli.[3] More importantly, existing data are insufficient for testing the subtler predictions of the many computational theories concerning generalization to novel views and across object deformations. Finally, the role of depth cues in recognition has been largely unexplored. The experiments described in this section were designed to address many such issues, concentrating on subordinate-level identification, which, unlike entry-level classification (Biederman, 1987), has been relatively unexplored.

All the experiments described below employed tasks in which subjects were asked to explicitly recall whether a currently displayed object had been previously presented.[4] Each experiment consisted of two phases: training and testing. In the training phase subjects were shown a novel object defined as the target, usually as a motion sequence of two-dimensional views that led to an impression of three-dimensional shape through structure from motion. 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

[3]Recent psychophysical and computational studies indicate that viewpoint dependence may be to a large extent an intrinsic characteristic of 3D shapes (Cutzu and Edelman, 1992; Weinshall et al., 1993).

[4]Such a judgment is commonly referred to as an "explicit" memory task. While some dissociations in performance have been found between similar explicit tasks and so-called "implicit" tasks such as priming or naming (Schacter, 1987), there is little evidence to indicate that this dissociation holds for changes across viewpoint (Cooper and Schacter, 1992). Moreover, Palmer, et. al.'s, (1981) and Tarr's (1989; Tarr and Pinker, 1989) studies employed implicit tasks, yet still revealed robust effects of viewpoint.

possible. No feedback was provided as to the correctness of the response.

5.1 Canonical views and their development with practice

To explore the first issue raised above, that of the determinants of canonical views, we tested the recognition of views all of which have been previously seen as a part of the training sequence (for further details see (Edelman and Bülthoff, 1992a), 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 Figure 4 as a three-dimensional 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. 4, top) decreases in the second session (Fig. 4, bottom). The pattern of error rates, in comparison, remained largely unaffected by repeated exposure.

5.2 Role of depth cues

5.2.1 Depth cues and the recognition of familiar views

A second set of experiments explored the role of three different cues to depth in the recognition of familiar views (for details, see (Edelman and Bülthoff, 1992a), Experiment 2). Whereas in the previous experiment test views were two-dimensional 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).

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 Corp.). 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%).

5.2.2 Depth cues and the generalization to novel views

A second manipulation probed the influence of binocular disparity (shown to be the strongest contributor of depth information to recognition) on the generalization of recognition to novel views (for details, see Edelman

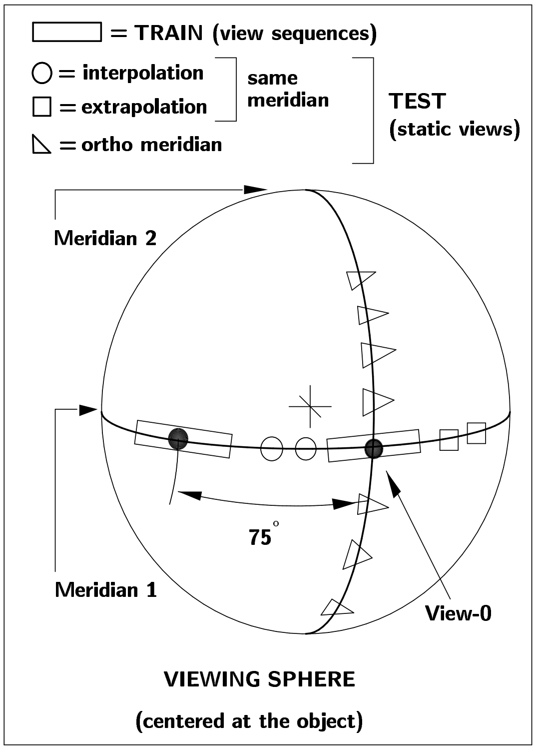

Figure 5: Generalization to novel views: An illustration of the INTER, EXTRA and ORTHO conditions. Computational theories of recognition outlined in Section 2 generate different predictions as to the relative degree of generalization in each of the three conditions. We have used this to distinguish experimentally between the different theories.

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 have 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 still significantly dependent on misorientation.

5.3 Generalization to novel views

A related experiment used an elaborate generalization task to distinguish among three classes of object recognition theories mentioned in Section 2: alignment, linear combination of views (LC), and view interpolation by 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

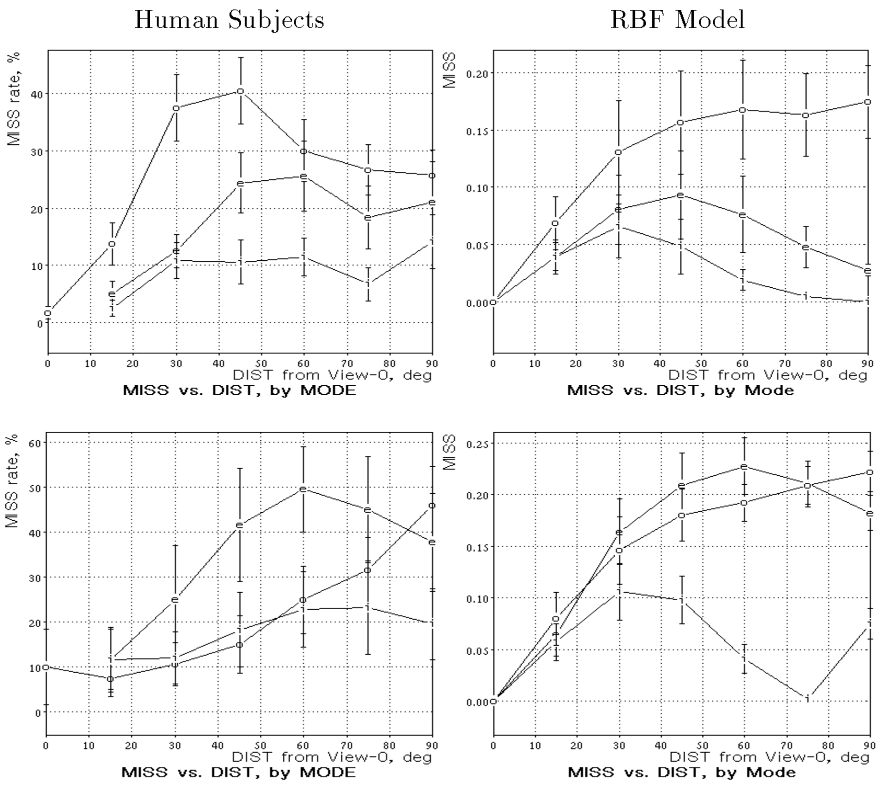

Figure 6: Generalization to novel views: *Top left:* Error rate vs. misorientation relative to the reference ("view-0" in Fig. 5) for the three types of test views – INTER, EXTRA and ORTHO, horizontal training plane. *Top right:* performance of the HyperBF model in a simulated replica of this experiment. *Bottom left and right:* same as above, except vertical training plane. Adapted from Bülthoff and Edelman, 1992.

training sequences was produced by letting the camera oscillate with an amplitude of $\pm 15^o$ around a fixed axis (Fig. 5). Target test views were situated either on the equator (on the 75^o or on the $360^o - 75^o = 285^o$ 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. 5).

The results of the generalization experiment, along with those of its replica involving the HyperBF model, appear in Figure 6. 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%. Repeated experiments involving the same subjects and stimuli, as well as control experiments under a variety of conditions yielded an identical pattern of error rates. The order of the mean error rates was changed, however, when the training views lay in the vertical instead of the horizontal plane. The means for the INTER, EXTRA and ORTHO conditions were in that case 17.9%, 35.1% and 21.7%.

The experimental results fit most closely the predictions of the HyperBF scheme and contradict theories that involve three-dimensional viewpoint-invariant models or viewpoint alignment models that do not allow for errors in recognition. In particular, the differences in generalization performance between the horizontal and the vertical arrangements of training views can be accommodated within the HyperBF framework by assigning different weights to the horizontal and the vertical dimensions (equivalent to using non-radial basis functions).

5.4 Generalization across deformations

In the last experiment reported in this section, we compared the generalization of recognition to novel views belonging to several different categories: those obtained from the original target object by rigid rotation, by three-dimensional affine transformation, and by non-uniform deformation (Edelman and Bülthoff, 1990; Sklar et al., 1993; Spectorov, 1993). The views in the rigid rotation category were obtained by rotation around the X axis (that is, in the sagittal plane), around the Y axis, and in the image-plane. In the deformation category, the methods were shear, stretch, quadratic stretch, and non-uniform stretch, all in depth. Altogether, views obtained through seven different transformation and deformation 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 (two-dimensional) deformation as measured in the image plane (cf. Cutzu and Edelman, 1992) than by the direction and the distance between the novel and the training views in the abstract space of all views of the tar-

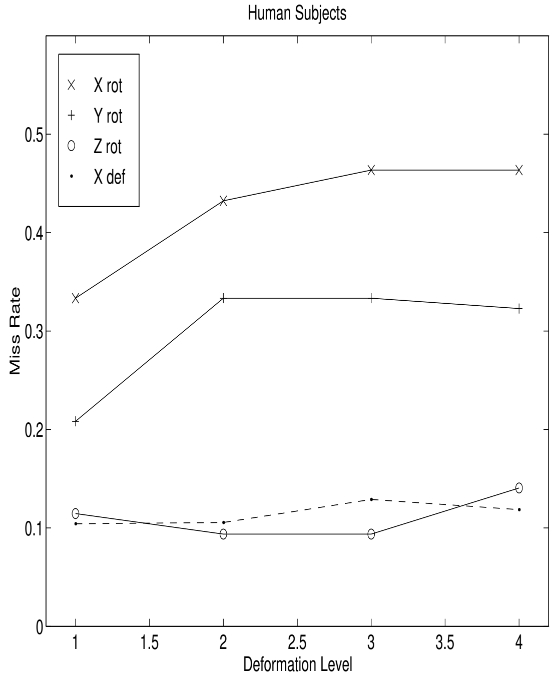

Figure 7: Human performance in the recognition of rotated and deformed objects. The subjects had to attribute briefly displayed static images of isolated objects to one of two classes (17 subjects participated; data are from 24 experimental sessions, which involved 5 different object pairs; for details, see Spectorov, 1993). The four curves show mean error (miss) rate for view related to the single training view by rotation around the X, Y, and Z axes (the latter is image-plane rotation), and by deformation along the X axis (data from four deformation methods, all of which produced similar results, are collapsed for clarity). Note that both image-plane rotation and deformation were easy, and elicited near-floor error rate.

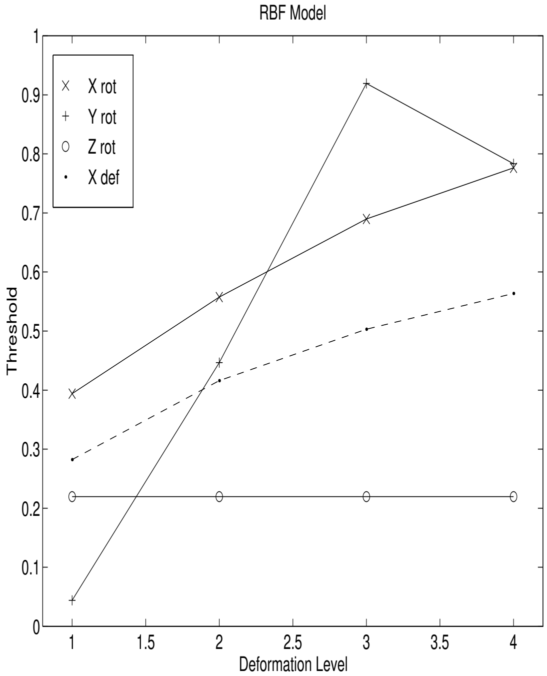

Figure 8: RBF model performance (measured by the classification threshold needed to achieve correct acceptance of all test views) in the recognition of rotated and deformed objects (for details, see Spectorov, 1993). The four curves are as in Figure 7. The wire-frame stimuli were encoded by vectors of angles formed by the various segments. Consequently, the image-plane rotation (which leaves these angles invariant) was as easy for the model as for the human subjects, but the deformations elicited somewhat worse performance (the rotations in depth were the most difficult, as they were for the humans). A choice of features other than angles may bring the performance of the model closer to that of humans.

get object. The HyperBF scheme was recently shown to produce a similar pattern of performance (Spectorov, 1993). More generally, such findings are consistent with the conception of multiple-views object representations as being exemplar-based, and consequently, recognition performance showing sensitivity to variations in two-dimensional image properties such as global shape, color, or illumination (Wurm et al., 1993).

5.5 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 viewpoint-invariant 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-invariant representation to allow privileged views and a built-in limit on generalization greatly weakens it, by breaking the symmetry that holds for truly viewpoint-invariant representations, in which all views, including novel ones, are equivalent.

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, 1989; Tarr and Pinker, 1989), can be accounted for in terms of viewpoint alignment (Ullman, 1989). According to Ullman's (1989) 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 (nor with Tarr's studies in which subjects received feedback about the correctness of their responses), which, on the one hand, show a marked and persistent dependency of error rate (also observed in Tarr's studies) on the distance to the training view for rigid rotations,[5] and, on the other hand, indicate that people are capable of generalization across object deformations. Moreover, the viewpoint dependency of the representations formed by subjects, manifested in the limitation on generalization to novel views, cannot be due exclusively to an absolute lack of three-dimensional 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 three-dimensional viewpoint-invariant 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 three-dimensional object is represented by a collection of specific views, each of which is essentially an image-based representation of the object as it is seen from a certain viewpoint, augmented by limited depth information.[6] 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 and share some structural information in common) are more closely associated with each other (Edelman and Weinshall, 1991; Perrett et al., 1989). 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 three-dimensional object represented by the collection of stored views (Ullman and Basri, 1991).

6 What are the features of recognition?

Most of the psychophysical findings reported above have been replicated by a computational model (Poggio and Edelman, 1990) based on interpolation of stored two-dimensional views (Bülthoff and Edelman, 1992). A natural question arising at this point is how those two-dimensional 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 Figure 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 one possible approach to the study of the features of recognition in human vision (see Edelman, 1991a for more details).

The central tenet of this approach, supported by the evidence presented in the preceding sections, is that recognition normally requires neither three-dimensional reconstruction of the stimulus, nor the maintenance of a library of three-dimensional models of objects (Edelman and Poggio, 1989). Instead, information sufficient for recognition can be found in the two-dimensional 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), but 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 ensuing generic recognition scheme is based on the idea of a hierarchy of image features, and is designed to address the major issue that remains at this point unresolved, namely, the capability of a recognition scheme based on interpolation among specific views for viewpoint-invariant performance exhibited by human subjects under certain circumstances (especially in tasks requiring basic-level classification, rather than the identification of individual objects; see (Biederman, 1987)). Evidence of viewpoint-invariant recognition has served in the past as an argument against multiple-view representation of objects. We propose that such evidence can

[5] 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.

[6] The basic limitation on the use of depth in recognition stems from its representation in a viewpoint-dependent coordinate frame (in Marr's terminology (Marr, 1982), such representation would be called a $2\frac{1}{2}D$-sketch). Another possible limitation is expected in view of the recent findings regarding the imperfections of the perception of three-dimensional shape, as mediated by different depth cues (Bülthoff and Mallot, 1988).

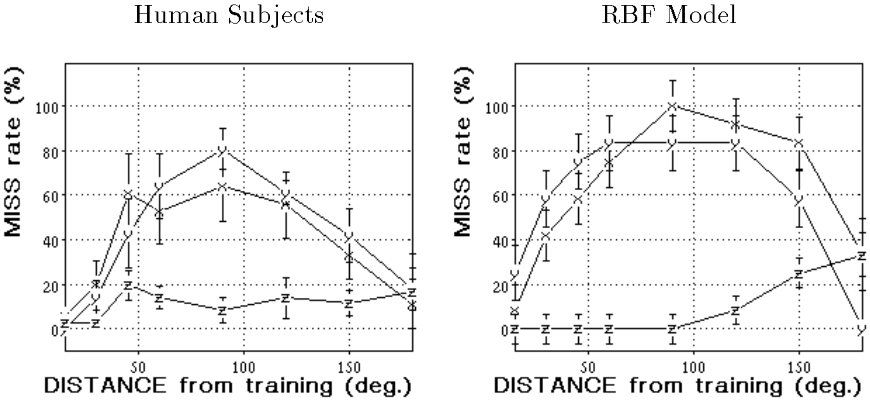

Figure 9: *Left:* human performance in the recognition of rotated wire-like 3D objects (Bülthoff and Edelman, 1992). Error rate of subjects trained on single view is plotted vs. distance between training and test views. Note poor generalization across rotations in depth (curves marked by x, y), compared to rotation in the image plane (curve marked by z; see text). *Right:* performance of the HyperBF network model (Poggio and Edelman, 1990; Edelman and Poggio, 1992; Bülthoff and Edelman, 1992) in the same task.

be accommodated within the framework of multiple-view representation by allowing for an appropriate encoding of the stored views. In other words, we propose to capture the varying degree of viewpoint invariance found in human recognition performance by endowing the model with an extensive repertoire of feature detectors, whose output (and not the raw input image) is fed into the classification stage (Edelman, 1991a).

Those of the detected features that are well-localized in the image (e.g., polyhedral vertices, as mentioned in the preceding section; see also Intrator et al., 1992) would allow fine distinctions among objects at the expense of relatively strong sensitivity to viewpoint (the location of a corner in the projected image is highly dependent on the object's attitude with respect to the observer). On the other hand, the so-called non-accidental features (Lowe, 1986; Biederman, 1987) offer relative insensitivity to viewpoint at the expense of reduced power of discrimination among objects. An example of such a feature is the presence of near-parallel lines in the image, which is highly unlikely to be caused by an accident of a viewpoint, but at the same time only allows to discriminate between objects that possess such parallel lines and those that do not. Finally, "diffuse" features such as surface color or texture may support recognition performance that is basically viewpoint-invariant and is exact to the extent that the surface markings are distinctive for each object under consideration. It is important to note that all three kinds of features — localized, non-accidental, and diffuse — can be detected by computational mechanisms resembling receptive fields, and can be considered, therefore, as a natural extension of a basis-function classification network (Poggio and Edelman, 1990).

A concrete example of the potential tradeoff between discrimination power and viewpoint invariance of a feature set is provided by recent experimental data (Edelman and Bülthoff, 1992a) shown in Figure 9. The plot on the left suggests that humans recognize 3D wire-like objects nearly independently of their image-plane orientation (but not of the orientation in depth; cf. Figure 7,8). A similar behavior is exhibited by a view-interpolation model which includes lengths of segments between connected vertices in the object representations (in addition to the coordinates of individual vertices). This relative insensitivity to rotation in the image plane is expected to cause the model to be more prone to confuse objects that have similar projected segment lengths, but different 3D structure. A complete invariance to image-plane rotation could be achieved by encoding vertex angles. For rotation in the image plane vertex angles stay constant but the projected angles are deformed by rotations in 3D.

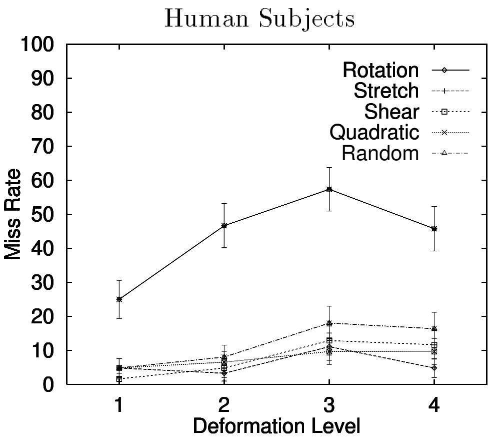

Figure 10: Subjects trained on two $\pm 15°$ motion sequences, centered at $\pm 37.5°$ to the reference view were tested on different deformation types based on the reference view. The deformation levels were normalized so the Level 4 is equivalent to the maximum amount of 2D vertex displacement possible with rotation in depth. Average miss rate of eight subjects for 6 objects.

Feature Model

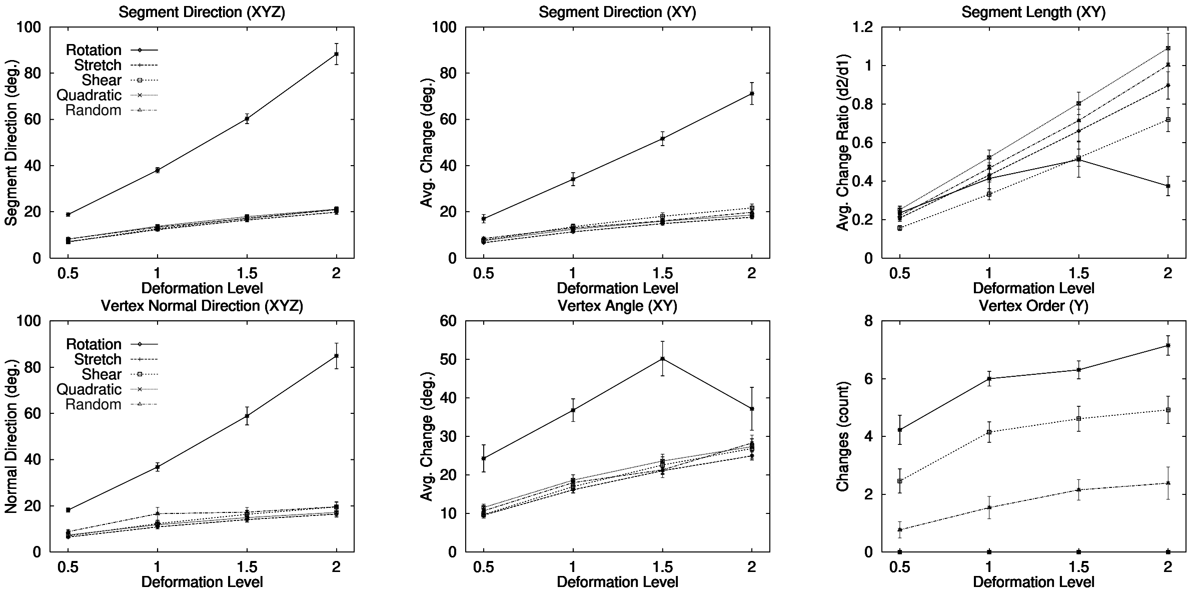

Figure 11: The various deformation methods effect measurements in 3D (XYZ) or in the image plane (XY) in different ways. The 2D vertex angle (XY) resembles best the psychophysical data presented in Figure 10.

A similar comparison between image-plane and rotation-in-depth may be found in the experiments reported by Tarr (1989; see, Fig. 3). However, in contrast to the results discussed above, subjects in these experiments exhibited large effects of viewpoint for both image-plane and in-depth rotations. One possible explanation for the discrepancy between these experiments may be the extent to which subjects relied on geometrically-defined versus familiarity-defined views. In terms of changes in image structure, all image-plane rotations are equivalent (e.g., constitute a single qualitative or characteristic view or aspect), and therefore may be undifferentiated with regard to multiple-views representations that encode views solely on the basis of qualitative changes in visible features.

However, Tarr's (1989) study intentionally manipulated the frequency with which selected views appeared to subjects (including views in the image-plane), thereby biasing them towards differentiating between featurally-equivalent image-plane rotations. Indeed, the fact that most canonical views of familiar objects seem to have a preferred orientation relative to gravitational upright, indicates that familiarity with specific viewpoints, as well as the presence of specific clusters of features, mediates what constitutes a view.

In order to test which feature (e.g., vertex position, vertex angle, segment direction, segment length, etc) is most likely the distinguishing feature used by the visual system in the recognition task, we compared recognition performance for a number of 2D and 3D deformation methods including rotation-in-depth, stretching, shearing, and random deformations (Sklar et al., 1993). For these experiments subjects first viewed a target object rotating $\pm 15°$ about a reference view. They were then asked to discriminate deformed versions of the target object from distractor objects which have undergone the same types and degrees of deformation. The results in Figure 10 show that the error rate for rotation-in-depth is clearly more pronounced than for the other deformation methods.

We then calculated how the different deformation methods and levels effect the following measurements: (1) segment direction (XYZ) is a 3D direction which can be only derived under stereoscopic display conditions; (2) segment direction (XY) is the projected 3D direction on the image plane; (3) segment length (XY) is the length of a wire segment measured in the image plane; (4) vertex normal direction (XYZ) is again a 3D measure which could only be derived under perfect 3D viewing conditions; (5) vertex angle (XY) is the projected angle measured in the image plane; (6) vertex order is a more topological measure which describes the change in top/bottom order of the vertices. A comparison with the psychophysical deformation data in Figure 10 shows that the vertex angle is the best 2D descriptor for human recognition performance of wire-like objects under varying image deformations (Fig. 11).

7 General conclusions

The psychophysical results reviewed in this paper present evidence that viewpoint-dependent representations and recognition processes play an important role in human object recognition. In particular, given that most studies have employed stimulus objects that share parts and have some spatial relations in common, viewpoint dependency is most strongly implicated in subordinate-level recognition. However, one must be cautious not to extend such conclusions to the more general assumption that viewpoint-dependent mechanisms are limited to the subordinate-level. Rather, the framework we have presented indicates that extreme viewpoint dependence and extreme viewpoint invariance lie at two ends of a continuum, with appropriate mechanisms and features recruited according to task demands, context, and the organization of visual memory. This conception of recognition in humans leaves less room for exclusively viewpoint-invariant theories of recognition, for instance, Geon-Structural-Descriptions (Biederman, 1987; Hummel and Biederman, 1992) in that a great deal of the extant psychophysical data on object recognition in humans is expressly inconsistent with such accounts (Bartram, 1974; Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992b; Cave and Kosslyn, 1993; Humphrey and Khan, 1992; Srinivas, 1993). Furthermore, the plausibility of such accounts is severely strained by their failure to accommodate the more flexible recognition mechanisms we have proposed. Indeed, even to the extent that such viewpoint-invariant theories are intended solely as explanations of entry-level performance, they are hampered by evidence for viewpoint-dependent patterns in naming familiar common objects (Palmer, et. al., 1981) and by their inability to provide both the stability and the sensitivity necessary to account for entry-level organization (cf., Marr and Nishihara, 1978).

A second important point to be drawn from the work surveyed here is that modeling psychophysically obtained response patterns permits us to "reverse-engineer" the human visual system – 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. In particular, the success of a HyperBF model that relied on simple receptive-field-like 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 has led to a number of insights:

- *Multiple-views.* Psychophysical evidence indicates that humans encode three-dimensional objects as multiple viewpoint-specific representations that are largely two-dimensional (but may include some depth information as well).
- *Normalization.* Psychophysical evidence indicates that subordinate-level recognition is achieved by employing a time-consuming normalization process to match objects seen in unfamiliar viewpoints to familiar stored viewpoints. The role of such mechanisms in entry-level recognition is less clear, but is more plausible than exclusively three-dimensional viewpoint-invariant accounts of recognition.
- *HyperBF Model and View Interpolation.* Psychophysical evidence in conjunction with computational simulations indicates that view interpolation offers a plausible explanation for viewpoint-dependent patterns of performance in terms of both response times and error rates. Moreover, this model offers an account of subtle aspects of generalization performance inconsistent with other viewpoint-dependent theories.

Our research program currently concentrates on the issue of feature extraction for recognition, on perceptual learning involved in the acquisition of object representations, and on the unification of theories of recognition spanning all levels of categorization. First, in modeling feature extraction in recognition, the identity and the relative importance of features discovered by computational learning models can be compared to a psychophysical characterization of the features of recognition relied upon by human subjects. Second, to the extent that both feature extraction and classification exhibit considerable flexibility, we are exploring the degree to which both priors and environmentally determined factors constrain learning and representation in human object recognition. Such factors include those relevant to general recognition, for instance, common feature sets, and those that differ for different classes of objects, for instance, subsets of non-generic features and restricted-class categorization methods. Finally, we believe that the concept of features of recognition, of varying complexities and degrees of spatial localization, may offer a unified approach spanning the continuum of subordinate-level to entry-level performance in human object recognition.

References

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). What some concepts might not be. *Cognition*, 12:263–308.

Barrow, H. G. and Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. In Hanson, A. R. and Riseman, E. M., editors, *Computer Vision Systems*, pages 3–26. Academic Press, New York, NY.

Bartram, D. J. (1974). The role of visual and semantic codes in object naming. *Cognitive Psychology*, 6:325–356.

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

Biederman, I. and Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 19(6):1162–1182.

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

Bülthoff, H. H. and Mallot, H. A. (1988). Interaction of depth modules: stereo and shading. *Journal of the Optical Society of America*, 5:1749–1758.

Cave, C. B. and Kosslyn, S. M. (1993). The role of parts and spatial relations in object identification. *Perception*, 22:229–248.

Cooper, L. A. and Schacter, D. L. (1992). Dissociations between structural and episodic representations of visual objects. *Current Directions in Psychological Science*, 1(5):141–146.

Cutzu, F. and Edelman, S. (1992). Viewpoint-dependence of response time in object recognition. CS-TR 10, Weizmann Institute of Science.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis*. Wiley, New York.

Edelman, S. (1991a). Features of recognition. CS-TR 91-10, Weizmann Institute of Science.

Edelman, S. (1991b). A network model of object recognition in human vision. In Wechsler, H., editor, *Neural networks for perception*, volume 1, pages 25–40. Academic Press, New York.

Edelman, S. (1992). Class similarity and viewpoint invariance in the recognition of 3D objects. CS-TR 92-17, Weizmann Institute of Science.

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

Edelman, S. and Bülthoff, H. H. (1992a). Modeling human visual object recognition. In *Proc. IJCNN-92*, volume IV, pages 37–42.

Edelman, S. and Bülthoff, H. H. (1992b). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400.

Edelman, S. and Poggio, T. (1992). Bringing the Grandmother back into the picture: a memory-based view of object recognition. *Int. J. Pattern Recog. Artif. Intell.*, 6:37–62.

Edelman, S. and Poggio, T. (May 1989). Representations in high-level vision: reassessing the inverse optics paradigm. In *Proc. DARPA Image Understanding Workshop*, pages 944–949, San Mateo, CA. Morgan Kaufman.

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

Eley, M. G. (1982). Identifying rotated letter-like symbols. *Memory & Cognition*, 10(1):25–32.

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

Farah, M. J., Rochlin, R., and Klein, K. L. (1994). Orientation invariance and geometric primitives. *Cognitive Science*, In Press.

Freeman, H. and Chakravarty, I. (1980). The use of characteristic views in the recognition of three-dimensional objects. In Gelsema, E. S. and Kanal, L. N., editors, *Pattern Recognition in Practice*, pages 277–288. North-Holland Publishing Company, New York.

Hummel, J. E. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99(3):480–517.

Humphrey, G. K. and Khan, S. C. (1992). Recognizing novel views of three-dimensional objects. *Canadian Journal of Psychology*, 46:170–190.

Intrator, N., Gold, J. I., Bülthoff, H. H., and Edelman, S. (1992). Three-dimensional object recognition using an unsupervised neural network: understanding the distinguishing features. In Moody, J., Hanson, S. J., and Lippman, R. L., editors, *Neural Information Processing Systems*, volume 4, pages 460–467. Morgan Kaufmann, San Mateo, CA.

Jolicoeur, P. (1985). The time to name disoriented objects. *Memory and Cognition*, 13:289–303.

Jolicoeur, P., Gluck, M., and Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, 16:243–275.

Koenderink, J. J. and van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216.

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

Lowe, D. G. (1986). *Perceptual organization and visual recognition*. Kluwer Academic Publishers, Boston, MA.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B*, 200:269–294.

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

Perrett, D. I., Mistlin, A. J., and Chitty, A. J. (1989). Visual neurones responsive to faces. *Trends in Neurosciences*, 10:358–364.

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

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

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

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

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

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13:501–518.

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

Sklar, E., Bülthoff, H. H., Edelman, S., and Basri, R. (1993). Generalization of object recognition across stimulus rotation and deformation. In *Investigative Opthalmology and Visual Science*, volume 34.

Spectorov, A. (1993). Generalization of object recognition across stimulus deformations. Master's thesis, Weizmann Institute of Science, Rehovot, Israel.

Srinivas, K. (1993). Perceptual specificity in nonverbal priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3):582–602.

Tarr, M. J. (1989). *Orientation dependence in three-dimensional object recognition*. PhD thesis, Massachusetts Institute of Technology.

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

Tarr, M. J. and Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, 1:253–256.

Tenenbaum, J. M., Fischler, M. A., and Barrow, H. G. (1981). Scene modeling: a structural basis for image description. In Rosenfeld, A., editor, *Image Modeling*, pages 371–389. Academic Press, New York.

Thompson, D. W. and Mundy, J. L. (1987). Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings of IEEE Conference on Robotics and Automation*, pages 208–220, Raleigh, NC.

Tversky, B. and Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, 113:169–193.

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

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

Vetter, T., Poggio, T., and Bülthoff, H. H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology*, 4:18–23.

Weinshall, D., Werman, M., and Tishby, N. (1993). Stability and likelihood of views of three dimensional objects. In Basri, R., Schild, U., and Stein, Y., editors, *Proc. 10th Israeli Symposium on Computer Vision and AI*, pages 445–454.

Wurm, L. H., Legge, G. E., Isenberg, L. M., and Luebker, A. (1993). Color improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human Perception and Performance*, 19(4):899–911.