COGNITION

Cognition 67 (1998) 1–20

Image-based object recognition in man, monkey and machine

Michael J. Tarr[a,]*, Heinrich H. Bülthoff[b]

[a]Brown University, Department of Cognitive and Linguistic Sciences,
P.O. Box 1978, Providence, RI 02912, USA
[b]Max-Planck-Institut für Biologische Kybernetik, Tübingen, Germany

Abstract

Theories of visual object recognition must solve the problem of recognizing 3D objects given that perceivers only receive 2D patterns of light on their retinae. Recent findings from human psychophysics, neurophysiology and machine vision provide converging evidence for 'image-based' models in which objects are represented as collections of viewpoint-specific local features. This approach is contrasted with 'structural-description' models in which objects are represented as configurations of 3D volumes or parts. We then review recent behavioral results that address the biological plausibility of both approaches, as well as some of their computational advantages and limitations. We conclude that, although the image-based approach holds great promise, it has potential pitfalls that may be best overcome by including structural information. Thus, the most viable model of object recognition may be one that incorporates the most appealing aspects of both image-based and structural-description theories. © 1998 Elsevier Science B.V. All rights reserved

Keywords: Object recognition; Image-based model; Structural description

1. Introduction

It has been over a decade since *Cognition* published its special issue on 'Visual Cognition' (Pinker, 1984a). That volume addressed topics such as mental imagery, visual attention and object recognition. Since that time there has been tremendous progress in each of these domains, but no where more so than in visual object recognition. In 1984 relatively little was known about the nature of the mental

* Corresponding author. Tel.: +1 401 8631148; fax: +1 401 8632255; e-mail: Michael_Tarr@brown.edu

PII S0010-0277(98)00026-2

representations used in human object recognition. Cognitive neuroscientific methods were still in their infancy and computational models of recognition were based primarily on Marr's (1982) work. In 1998 we know much more about object recognition through research in each of these domains. First, psychophysical studies have revealed many facets of the amazing human capacity to recognize objects (Jolicoeur, 1985; Biederman, 1987; Tarr and Pinker, 1989; Bülthoff and Edelman, 1992; Humphrey and Khan, 1992). Second, a wide range of neuroscientific methods have been used to investigate the neural basis of object recognition in non-human primates and brain-damaged humans (Perrett et al., 1987; Farah, 1990; Goodale and Milner, 1992; Logothetis et al., 1995; Tanaka, 1996). Third, there have been significant advances in the sophistication, robustness and ecological validity of computational models (Poggio and Edelman, 1990; Ullman and Basri, 1991; Hummel and Stankiewicz, 1996b).

In this special issue we present recent work by some of the most creative scientists studying the problem of visual recognition. Moore and Cavanagh take a classic demonstration, the perception of 'two-tone' images, and turn it into a method for understanding the nature of object representations in terms of surfaces and the interaction between bottom-up and top-down processes. Tarr and Gauthier use computer graphics to explore whether viewpoint-dependent recognition mechanisms can generalize between exemplars of perceptually-defined classes. Goodale and Humphrey use innovative psychophysical techniques to investigate dissociable aspects of visual and spatial processing in brain-injured subjects. Perrett, Oram and Wachsmuth combine neurophysiological single-cell data from monkeys with computational analyses to provide a new way of thinking about the mechanisms that mediate viewpoint-dependent object recognition and mental rotation. Ullman's work also addresses possible mechanisms that may account for viewpoint-dependent behavior, but from the perspective of machine vision. Finally, Schyns synthesizes work from many areas, providing a coherent account of how stimulus class and recognition task interact. What is notable is that this group of contributors brings together a wide range of methodologies to a common problem. Moreover, much of the work presented in this volume provides converging evidence for a common approach – what we refer to as 'image-based' or 'view-based' recognition. The key idea of the image-based approach is that object representations encode visual information as it appears to the observer from a specific vantage point. Note that, although such a claim is actually neutral with regard to particular types of features, including pixel regions, shape contours, texture, etc., it does imply that features, regardless of their content, are viewpoint-dependent. Consequently, the usefulness of a given feature for recognition will diminish as that feature changes its appearance with changes in viewpoint and overall recognition performance will be viewpoint-dependent.

2. Models of recognition

The study of visual object recognition is often motivated by the problem of recognizing 3D objects given that we only receive 2D patterns of light on our retinae. A commonly-held solution, popularized by Marr (1982), is that the goal of vision is to reconstruct the 3D scene. Reconstruction assumes that visual perception is a hierarchical process which begins with local features that are combined into progressively more complex descriptions (also see Pinker, 1984b). Note that the types of features used and how they are combined is completely deterministic. That is, particular types of features and the relations between them are pre-defined and used for reconstruction across all images. Moreover, the presence or absence of a given feature is absolute-there is no ‘middle ground’ in which there is partial or probabilistic evidence for a feature. Thus, lines are grouped into contours, contours into surfaces, and surfaces into objects. At the endpoint of the reconstruction process Marr and Nishihara (1978) assumed that viewer-centered descriptions (what Marr termed ‘sketches’) are remapped into 3D object-centered representations. This final step was motivated by Marr and Nishihara’s (1978) suggestion that object representations should be relatively stable, that is, they should generalize or be invariant over changes in the retinal image. Otherwise, Marr and Nishihara argued, new, distinct representations would be required for each small variation in the image of a given object, e.g. for each change in 3D position, each change in illumination, etc. More concretely, this meant that object representations should be object-centered rather than viewer-centered-hence their conjecture that objects are represented as configurations of 3D parts or volumes.

Although there is a theoretical elegance to this approach, it has never been obvious that recovering descriptions of 3D parts from 2D images is generally possible. Indeed, during the 1980s numerous machine vision researchers attempted to implement reconstruction algorithms with only marginal success (Nalwa, 1993). Thus, one argument that favors the image-based approach is that it does not require reconstruction. Indeed, given that our visual systems are given viewer-centered images as input, it would not be altogether surprising if visual recognition was based on similar mental representations.

Notwithstanding these potential problems, Marr’s work has had tremendous impact on the study of vision, and, in particular, helped to shift the focus of high-level vision research from visual imagery (e.g. Kosslyn, 1980) to visual object recognition during the 1980s. One of the most prominent theories to come out of this era was the ‘Recognition-By-Components’ model (RBC) by Biederman (1987). The RBC model built on Marr and Nishihara’s earlier work on object recognition, proposing that objects are represented as collections of volumes or parts. What RBC added, however, were additional syntactic constraints that specified the allowable types of volumes, how such volumes might be recovered from 2D images, and the types of qualitative spatial relations that connect such volumes. RBC also followed the stricture that object representations should be stable and, consequently, proposed that the configurations of parts that are used to describe objects are invariant across changes in viewpoint (up to significant changes in the visible part structure (Biederman and Gerhardstein, 1993)), illumination, and color (Biederman and Ju, 1988). Thus, the RBC approach, often referred to as a ‘structural-description’ model, provides a computationally-elegant, but completely deterministic (i.e. the elements of the representation are pre-defined and such elements are either present or absent), answer to the question of how human perceivers recognize objects across changes in viewpoint.

Although RBC has been very influential, it is still not clear that any approach that relies on the recovery of 3D volumes is robust enough to subserve general object recognition. Moreover, the actual evidence for viewpoint-invariance in human visual recognition (as predicted by RBC) is somewhat thin – the most notable experiments that obtain viewpoint invariance for rotations in depth[1] (Biederman and Gerhardstein, 1993) having only limited generalizabilty to other recognition tasks and stimulus sets (Hayward and Tarr, 1995; Tarr and Bülthoff, 1995; Tarr et al., 1997). In contrast, psychophysical and neurophysiological studies from the late 1980s and early 1990s offer a somewhat different conclusion – under a wide variety of experimental conditions, human object recognition performance is strongly viewpoint-dependent across rotations in depth (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Humphrey and Khan, 1992; Tarr, 1995). Converging evidence for this result has come from single-cell recording studies in the inferior temporal cortex of monkeys (Logothetis et al., 1995). From a computational perspective this result is rather surprising because view-dependent object representations are necessarily less stable than view-invariant representations – yet the data seem to imply that humans rely on image-based representations that are viewpoint-dependent.

On the face of it, the image-based approach to recognition appears to be subject to Marr and Nishihara’s criticism of viewer-centered models – that each distinct viewpoint of an object necessitates a separate representation. What Marr and Nishihara omitted was that the stability constraint only holds if there is no means for generalizing from one image to another. For instance, if observers can compensate for changes in viewpoint by a normalization process, they may be able to use a small number of viewer-centered representations to recognize objects in any orientation in space (i.e., a ‘multiple-views’ representation). Indeed, proponents of the image-based approach have offered a variety of different mechanisms for generalizing from unfamiliar to familiar views, including mental rotation (Tarr and Pinker, 1989), view interpolation (Poggio and Edelman, 1990) and linear combinations of views (Ullman and Basri, 1991). Even more sophisticated (Ullman, 1998) and neurally-plausible (Perrett et al., 1998) generalization mechanisms are presented in this volume.

While generalizing over viewpoints has been accepted as one way of providing stability within image-based models, generalizing over different instances of a perceptually-defined class has been seen as a far more difficult problem (Biederman and Gerhardstein, 1995). Consider that almost every behavioral study that has reported viewpoint-dependent recognition has also used tasks in which subjects must discri-

[1] There are several studies that have obtained orientation invariance for rotations in the picture plane (Corballis et al., 1978; Tarr and Pinker, 1990). This result, however, is not considered diagnostic for theories of recognition in that there are both image-based and structural-description models that predict recognition costs over changes in picture-plane orientation (Tarr and Pinker, 1989; Hummel and Biederman, 1992).

minate between visually-similar objects, not object classes. For example, subjects might be asked to distinguish robins from sparrows, but not birds from cars. Thus, there is little data that directly addresses the question of how basic-level (Rosch et al., 1976) or entry-level (Jolicoeur et al., 1984) recognition is accomplished. Although more specific recognition discrim-inations are no doubt important, it is relatively uncontroversial that visual recognition most frequently functions at the basic level. Moreover, with the exception of RBC theory, most models of human visual recognition have failed to provide well-specified mechanisms for class-level recognition. Image-based models seem particularly problematic in this regard-because objects are represented as viewpoint-specific images it is assumed that the representations are also specific to particular exemplars, not object classes. The common claim is that image-based or view-based representations are templates based on inflexible linear coordinates, and, as such, cannot accommodate the variations in image geometry that characterize different exemplars of a single object class (Hummel, 1998). Indeed, as reviewed below, this is only one of several oft-cited critiques of image-based models.

3. Evidence for the image-based approach

Criticisms that portray image-based theories as overly simplistic are no longer tenable as arguments against such theories. This claim is supported by recent extensions of the image-based approach (Edelman, 1995b; Beymer and Poggio, 1996; Moses et al., 1996) and by the new work presented in this volume. While it is true that earlier image-based models suffered because of simplifying assumptions (e.g. locating features at fixed x, y coordinates in the image-plane), theorists were well-aware of the problem. For instance, in one of the seminal papers on the image-based approach, Poggio and Edelman (1990; p. 264) state that ‘The key issue of how to detect and identify image features that are stable for different illuminations and viewpoints is outside the scope of this paper...[the model] does not require the x, y coordinates of image features as inputs: other parameters of appropriate features could also be used...Recognition of noisy or occluded objects, using realistic feature identification schemes, requires an extension of the scheme...’. What we have witnessed over the past several years are serious attempts to extend the image-based approach. For instance, Bricolo et al. (1997) employ features characterized by small brightness regions that can be located at any position within the image, thereby facilitating flexible object representations. Amit and Geman (1997) use similar features, but relate the spatial positions of such features in a manner that allows for a highly robust matching scheme. There have also been attempts to develop models that use less local representations of shape. For example, Hayward and Tarr (1997) found that observers were sensitive to both the metric and qualitative structure of image contours in 3D objects.

Concurrent with efforts to extend the image-based approach, there has been a great deal of scrutiny regarding the biological validity of the structural-description approach. In particular, a variety of labs has tested the specific predictions of RBC and related theories that posit the recovery of view-invariant parts. Behavioral results suggest that such models offer only limited explanatory power. For example, several studies (Bülthoff and Edelman, 1992; Humphrey and Khan, 1992; Tarr, 1995) have demonstrated that, when subjects are trained to recognize novel objects in a small set of viewpoints, not only are the generalization patterns viewpoint-dependent, but, critically, they are related to the distance between an unfamiliar test view and the nearest familiar view. Such results provide strong evidence for object representations based on multiple image-based views matched to input shapes through normalization processes. Supporting this claim, Logothetis et al. (1995) trained monkeys to recognize novel objects similar to those used in Bülthoff and Edelman (1992). Recordings in the inferior temporal cortex of these monkeys reveal 'view-tuned' neurons, that is, cells that are preferentially active for specific instances of these trained objects in specific views. Moreover, for a given object, Logothetis et al. (1995) found that different neurons coded for different views, thereby providing a multiple-views representation similar to that inferred from behavioral data. It should be noted, however, that Logothetis et al. (1995) also found some evidence for view-independent neurons for some objects. The question is whether such neurons arise as a result of the derivation of truly viewpoint-invariant object representations or because multiple view-tuned neurons simply feed into a single neuron.

As mentioned, one criticism of this body of results is that the stimuli used in these experiments were typically drawn from a single visually-similar class, e.g. 'paper-clip objects' (Bülthoff and Edelman, 1992) or 'cube objects' (Tarr, 1995; Fig. 1a). In part due to this limitation, it is popularly held that both structural-description and image-based models explain elements of human visual recognition. For exam-

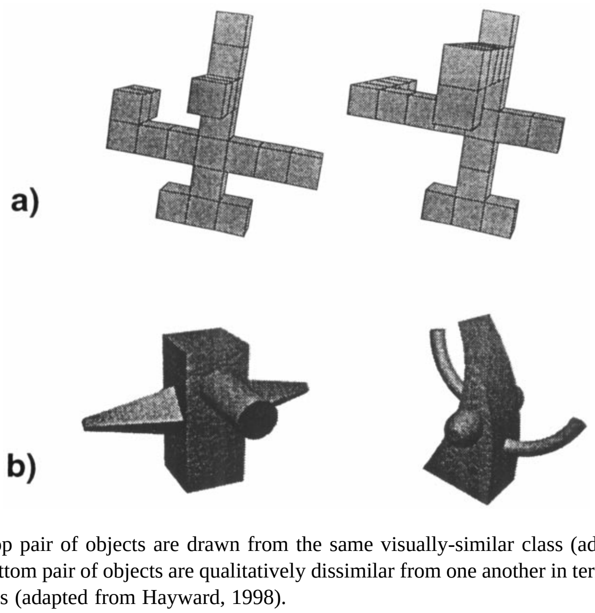

Fig. 1. (a) The top pair of objects are drawn from the same visually-similar class (adapted from Tarr, 1995). (b) The bottom pair of objects are qualitatively dissimilar from one another in terms of both image structure and parts (adapted from Hayward, 1998).

ple, structural-descriptions providing categorical-level access and image-based mechanisms providing within-class or exemplar-specific level access (Jolicoeur, 1990).

Why then can we claim that current empirical results strongly support image-based models and provide little evidence for view-invariant models? The answer lies in a series of recent behavioral studies based on critiques of image-based studies (Biederman and Gerhardstein, 1993, 1995). Specifically, Biederman and Gerhardstein (1993) proposed three ‘conditions for invariance’ claimed to be ‘typical’ of human object recognition. Briefly, the conditions are that: (i) objects must be decomposable into parts; (ii) each object in the recognition set must be composed of a distinct configuration of parts; (iii) different viewpoints of the same object must show the same configuration of parts. These conditions attempt to exclude almost all earlier studies, e.g. Bülthoff and Edelman (1992) and Tarr (1995), from consideration as diagnostic of visual recognition. In response to this critique, researchers began to test recognition performance using experimental designs that satisfied Biederman and Gerhardstein’s conditions. In particular, each target object is qualitatively different from the other objects in the recognition set (Fig. 1b). Results in these studies strongly support image-based models. In almost every case, even given highly dissimilar objects, recognition performance has been found to be viewpoint-dependent (Liter, 1995; Hayward and Tarr, 1997; Suzuki et al., 1997; Hayward, 1998; Tarr et al., 1997, 1998); but see also Biederman and Gerhardstein (1993).

4. Reconciling image-based and structural-description models

Recent empirical results seem to pose problems for a particular family of structural-description models, and, most notably, RBC. However, they do not indicate that all approaches to structural-descriptions are invalid, only that we need to rethink what kind of structural knowledge is encoded. Indeed, a major goal of vision scientists should be to develop models that provide robust accounts of human performance within a combined image-based/structural-description framework (insofar as the preponderance of behavioral data supports such a framework and that there are computational advantages to both approaches).

What are the challenges in developing such an approach? First of all, we must consider the fact that the spectrum of results measuring viewpoint dependency ranges from almost complete viewpoint invariance (Biederman and Gerhardstein, 1993; Tarr et al., 1998) to extreme viewpoint dependence (Bülthoff and Edelman, 1992; Humphrey and Khan, 1992; Tarr, 1995). What is not the case is that we see a pattern across experiments in which there is simply either invariance or dependence. Rather, depending on the homogeneity of the stimulus class and the particular recognition task, we obtain relatively more or less of an effect (Edelman, 1995a; Schyns, 1998). This is exemplified by the results of nine experiments reported by Tarr et al. (1998). They found that under the specific conditions used by Biederman and Gerhardstein (1993), i.e. match-to-sample recognition of qualitatively-distinct

3D objects and response time feedback on each trial, recognition performance was close to viewpoint-invariant. However, given a different recognition task, e.g. sequential matching or naming, or no feedback, recognition of the same 3D objects was viewpoint-dependent. A viable model of recognition must account for this continuum and the conditions under which different values along it are obtained. Constraining any such account, several recent studies have tested some of the conditions that appear to determine the degree of viewpoint invariance or dependence. For example, Tarr and Pinker (1990) found that performance was viewpoint invariant when subjects recognized 2D shapes that could be discriminated by a unique one-dimensional ordering of features, but was viewpoint-dependent when the shapes could only be discriminated by using 2D relations between features. Similarly, Tarr et al. (1997) found that the recognition of 3D objects containing single unique parts was much less view-dependent as compared to the recognition of objects containing multiple parts that had to be related to one another.

Second, we must consider the fact that human perceivers are capable of recognizing objects at multiple categorical levels, ranging from basic-level (Bartram, 1976; Jolicoeur, 1985; Biederman and Gerhardstein, 1993) to subordinate-level (Gauthier et al., 1997) to item-specific (Bülthoff and Edelman, 1992; Humphrey and Khan, 1992; Tarr, 1995) recognition. Models of recognition must account for how we represent object information that supports multiple levels of access – either through multiple systems that interact (e.g. with structural-descriptions supporting the category level and image-based mechanisms supporting the more specific levels (Jolicoeur, 1990; Tarr and Pinker, 1990; Marsolek and Burgund, 1997) or through a single system that is highly adaptable to varying recognition conditions (Biederman et al., 1997; Edelman, 1995b; Gauthier and Tarr, 1997b).

Third, we must consider the fact that human perceivers vary in the level of expertise they have for a given stimulus class. The degree of experience an individual has had with a class may help to determine the default level of access for recognition, how sensitive recognition is to image transformations, e.g. brightness reversal, and to changes in configural information (Gauthier and Tarr, 1997a; Tanaka and Sengco, 1997; Gauthier et al., 1998). Models of recognition must be sufficiently plastic to adapt as experience with an object class accumulates. Moreover, it is not enough for a model to simply allow recognition at different levels of expertise. There must be an account for why performance across various behavioral measures changes with changes in expertise.

Finally, it is crucial to realize that performance in a given recognition task is actually the product of a complex interaction between all of these factors: homogeneity of the stimulus class; categorical level; and level of expertise (Gauthier, 1998; Schyns, 1998). As a rule, extant models of recognition have tended to focus on only one or two of these factors, for example, comparing face recognition to non-face object recognition (Farah, 1992), or contrasting basic-level with subordinate-level recognition (Biederman, 1987). Recent models have certainly begun to move away from such simple dichotomies (e.g. Edelman, 1995a), but there is clearly still a great deal of work to be done.

5. Current problems with image-based models

It is our contention that new approaches to image-based recognition can account for the complete range of human recognition performance. However, meeting this challenge may necessitate abandoning old notions of what is meant by an image-based or view-based model. In particular, because there is some behavioral evidence to support both image-based representations and structural-descriptions, as well as computational strengths for each, it seems likely that a viable model will encompass elements of both. In order to see why this is the case, let us examine some of the most oft-cited problems with 'traditional' image-based models.

5.1. Class generalization and categorical representation

Image-based models typically represent the appearance of a specific object from a specific viewpoint. As such, they are exemplar-based and seemingly poor candidates for class-level recognition. Moreover, even if it is possible to generalize from familiar exemplars of a class to unfamiliar exemplars, mechanisms for specifying category membership and representing perceptually-defined categories as categories are less than obvious.

5.2. Hyper-sensitivity, inflexibility and combinatorial explosions

Even for the recognition of a single object, an image-based approach may have difficulties in generalizing across slight variations in appearance. Marr and Nishihara (1978) suggested that object representations should be sensitive in order to discriminate between visually-similar objects. However, sensitivity should not be so great that each specific change in the image necessitates a distinct representation. Therefore, if image-based information does not generalize across viewing conditions, an excessive number of representations may be required to capture the appearance of only a single object. Indeed, image-based models often appear prone to this problem in that some approaches have posited inflexible or 'holistic' representations that are ill-suited for generalizing from known to unknown viewing conditions. Although it has been argued that trading 'memory for computation' in this manner is acceptable, it is unclear that there can ever be sufficient memory to compensate for a system that allows for only minimal generalization.

5.3. Matching algorithms and normalization mechanisms

In order for image-based representations to generalize between exemplars or between views, robust matching algorithms must be specified. That is, there must be some mechanism for measuring the perceptual similarity (within some domain) between an input image and known objects. One possibility is that we simply measure local pixel or brightness similarity across images, but it is doubtful that such representations will exhibit the necessary robustness because they are likely to be highly unstable over image transformations. An alternative might be to measure similarity across the output of receptive fields, although it is unclear that what are still relatively local descriptions of the image will suffice. More plausibly, relational information between local features is needed. A second issue is how to match an unfamiliar view of an object to a familiar view of the same object. Image-based models have often appealed to mental transformations or alignment models that seem to beg the question. At issue is that such processes must establish the direction of rotation before executing a rotation or alignment. Determining this information seems to imply that recognition, at least at a coarse level, has already occurred.

6. Extending the image-based approach

6.1. Interpolation across views

How do we extend image-based models to address such problems? As alluded to earlier, there has been increasing interest in developing image-based models that can generalize between familiar and unfamiliar views for a given object and between familiar and unfamiliar exemplars for a given class. Indeed, some of the earliest computational approaches to image-based recognition relied on mechanisms that effectively measured the visual similarity between different views rather than executing a transformation, for example, the view interpolation model of Poggio and Edelman (1990). In this approach, specific object views are described as sets of viewpoint-dependent features (e.g. the output of receptive fields). Each view can then be considered a point in a high dimensional space that captures the appearance of all possible views. Generalization from unknown to known views (those in memory) is accomplished by establishing the location of the unknown view within this space and measuring the similarity of its features relative to the features of the nearest known views, that is, 'interpolating' across the view space. Such models are appealing in that they do not require the precomputation of 'alignment keys' (Ullman, 1989) or other information about the shape prior to recognition (see also Ullman, 1998). Moreover, there is some psychophysical evidence to support the view interpolation approach (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992). Critically, more recent computational instantiations of view interpolation have adopted more flexible representations of image features (Bricolo et al., 1997; Riesenhuber and Poggio, 1998) based on neurophysiological results that provide evidence for view-tuned neurons (Logothetis et al., 1995). Reinforcing the biological plausibility of this approach, Perrett et al. (1998) offer specific neurophysiological evidence for 'evidence accumulation' across collections of local features, a mechanism similar to that proposed in some of the recent computational models.

6.2. Interpolation across exemplars

One insight that has helped extend the image-based approach is that interpolation need not only occur between views of an object. It is equally plausible that interpolation can occur between different exemplars of a perceptually-defined object class. Thus, just as an unknown view can be recognized through interpolation to a visually-similar nearby view, an unknown exemplar can be recognized through interpolation to a visually-similar exemplar (Lando and Edelman, 1995; Beymer and Poggio, 1996). Psychophysical evidence for exactly this sort of class generalization has begun to accrue (Moses et al., 1996; Gauthier and Tarr, 1997b) and is discussed in detail by Tarr and Gauthier (1998). Indeed, the neural mechanisms proposed by Perrett et al. (1998) could readily extend to measuring the similarity of local features across class instances[2].

One caveat about image-based generalization processes is that view interpolation is possible because the view space for a given object or class tends to vary smoothly. That is, across a wide range of adjacent viewpoints, there are only small qualitative changes in the projected image of the object. When dramatic changes in the image do occur, such as when a major part comes in or out of view (Biederman and Gerhardstein, 1993; Hayward and Tarr, 1997; Hayward, 1998), it is probable that interpolation mechanisms may fail across this boundary (referred to as a ‘visual event’ by Koenderink, 1987). Under such conditions it may be that explicit[3] view-invariant structural information is required to map a view of an object onto a qualitatively different view of that same object – a possibility we discuss below.

Similarly, interpolation between different exemplars is only likely when the two are visually similar, that is, when the space of exemplars defining the object class varies smoothly. How likely is the assumption of smoothness? For many basic-level classes, the answer may be quite likely. Several recent computational studies have assessed how easily familiar objects can be categorized into stable visually-defined classes. Using only silhouette or boundary contour information readily extracted from images, it has been found that large numbers of exemplars can be separated into perceptual categories. Critically, these categories correspond quite closely to those that might be delineated by human perceivers (Ullman, 1996; Cutzu and Tarr, 1997). A similar conclusion regarding the perceptual stability of basic-level classes has been reached by developmental psychologists studying the acquisition of visual category information in infants. For example, Quinn et al. (1993) found that 3–4-month-old infants were capable of discriminating images of birds from dogs and cats, as well as images of dogs and cats from one another. Presumably, given the limited experience such young infants have had with these object classes, their performance must be based on visual information available in the images, not on conceptual knowledge acquired through the names assigned to the objects. Indeed, in the original formulation of basic-level categories, Rosch et al. (1976) posited that

[2] While the view-tuned neurons reported in Logothetis et al. (1995) appeared to have their highest activation when presented with a specific exemplar, e.g. a particular ‘paperclip’ object, the same neurons sometimes showed above-resting-level activation for the presentation of visually-similar objects, suggesting that within-class generalization may have occurred.

[3] We use the term explicit here because we mean a distinct, explicitly represented description of an object’s structure. As discussed below, we propose that image-based representations also encode implicit structural information in terms of the relative positions of local features.

most classes have a perceptual basis and, in particular, that silhouettes might provide this basis. Overall, these results suggest that generalization within perceptually-defined classes is a plausible extension to image-based models. On the other hand, much as multiple-views are necessary to represent the complete 3D structure of an object (Tarr, 1995), multiple exemplars will be necessary to represent the complete range of object classes.

Multiple-views or multiple-exemplar representations alone do not provide a basis for representing 3D objects. What are needed are organizing principles that provide a ‘glue’ between qualitatively dissimilar views or exemplars (qualitatively similar views or exemplars may be related on the basis of visual similarity and interpolation mechanisms). For multiple-views representations, two types of information may be available as evidence that distinct and geometrically dissimilar views arise from the same 3D object.

6.3. Temporal associations

Consider that from one moment to the next, the most likely image to follow an image of given object is another view of that same object. Using simple occurrence-based association mechanisms (i.e. Hebbian learning) the visual system could come to associate distinct views. Specifically, the more often that two images temporally co-occur, regardless of image similarity (Miyashita, 1993), the more strongly they will be associated. If we couple this with some measure of perceptual similarity, we have a powerful mechanism for building multiple-views representations. Although the existence this type of temporal association is somewhat speculative, there are recent psychophysical, neurophysiological, and computational results that provide some evidence in this direction (Miyashita, 1993; Wallis, 1996a,b).

6.4. Explicit structural information

Associations between views may also be formed by explicitly represented structural information. We have already alluded to the fact that there are instances for which structural information about an object may be critical. Insofar as a structural-description of a given object is stable over changes in viewpoint, it may provide a mechanism for linking two distinct views. However, given the problems we have raised for structural-descriptions based on 3D parts, what kind of structural information might offer sufficient stability, yet not predict complete invariance? One candidate is a ‘medial-axis’ representation derived from an object’s silhouette, that is, a skeletal description of the object. The idea of using medial-axis representations is quite old, being first proposed in Blum’s ‘Grassfire’ model (Blum, 1967). Blum’s idea was that if the edges of an object’s silhouette are simultaneously ‘ignited’, the flames will burn inward until they collide or interfere with one another, thereby leaving a skeleton describing the shape of the object. More recent instantiations of this idea have provided computationally robust methods for recovering skeletal descriptions (Kimia et al., 1995; Zhu and Yuille, 1996). Additionally, there is recent behavioral evidence suggesting that medial-axis representations are computed early in visual processing (Kovacs and Julesz, 1994).

Medial-axis representations are appealing for several reasons. First, they are readily computed from an object's silhouette or bounding contour, a type of information that is recoverable from 2D images (in contrast with 3D part descriptions). Second, they provide a topological description of object shape that allows the representation to remain relatively stable over changes in viewpoint, illumination, color and object configuration (Marr and Nishihara, 1978; Zhu and Yuille, 1996). Third, the topological nature of the representation facilitates fast and efficient matching between object descriptions (Kimia et al., 1995).

Given these positives, why do we claim that explicit structural information only supplements image-based recognition? The answer is that medial-axis representations provide only limited information about an object, that is, a coarse description of its shape. Recognition based on such information would not be entirely reliable and, at best, might provide a 'ballpark' estimate of the category (see the examples provided in Zhu and Yuille, 1996). Moreover, we have already made it clear that recognition may occur at many different categorical levels. Thus, skeletal descriptions may help constrain the search space during recognition, but in and of themselves they are not sufficient for recognition. As an example, consider an observer that has learned to recognize several views of a 3D object; unfamiliar views that are similar to familiar views may be recognized through normalization processes such as interpolation. However, because of qualitative differences between new views and stored views, it may not always be possible to recognize unfamiliar views through interpolation. Therefore, new views should be learned as distinct nodes ('aspects' in Koenderink, 1987) in a multiple-views representation. The question is, a node in *which* multiple-views representation? The answer may be provided by relatively view-invariant structural information: the skeletal description for a given view may be similar enough to the skeletal descriptions derived from other views of the same object or class to help constrain which particular multiple-views object representation is selected.

It is worth noting that other theorists have taken different directions in combining image-based information and explicit structural-descriptions. For example, Hummel and Stankiewicz (1996b) have sought to extend a neural-net implementation of RBC (Hummel and Biederman, 1992) to include both structural-descriptions based on 3D parts and image-based surface information. Their model is motivated by the computational problem of 'binding' together the different component parts that form a structural-description of an object. Interestingly, the representation of surfaces in their model helps to defray the costs of the binding process. Thus, image-based and structural information may be both functionally and computationally complementary. Indeed, their model is much more successful than its predecessor–some specific predictions regarding the need for attention in the binding process and sensitivity to left-right reflection, translation, and scale changes have been born out in behavioral experiments (Stankiewicz et al., 1998).

6.5. Implicit structural information

In contrast to explicit medial-axis descriptions, image-based models may also incorporate what we refer to as implicit structural information. We use the term implicit here to denote the fact that this type of structural information does not provide a global description of object shape, but rather simply codes relations between local features. Consider that images may be described as collections of local measures of the image at different locations (e.g. the output of oriented receptive fields, Edelman, 1993; small pixel regions, Bricolo et al., 1997; Amit and Geman, 1997; qualitative and quantitative measures of shape, color, texture, etc.). At one extreme, it may be possible to represent these features in a completely unordered fashion, thereby losing all information regarding the spatial relation between one local feature and the next. Such a representation would retain only information about the presence of each feature anywhere within the image. It is clear that such a model has severe limitations: for example, randomly scrambling the positions of the features will produce an image that cannot be distinguished from the original. On the other hand, there is computational evidence that even such simplistic representations have a surprising degree of explanatory power in terms of recognizing novel views of 3D objects (Bricolo et al., 1997). At the other extreme, the relations between local features may be completely deterministic, as in a literal image where the point-to-point positions between features are rigidly fixed relative to one another, e.g. described in linear coordinates (Poggio and Edelman, 1990). This is the kind of shape representation often associated with templates and image-based models (Hummel and Stankiewicz, 1996a). Obviously, such rigid and completely deterministic templates where features match absolutely or not at all also have severe limitations: for example, hyper-sensitivity to trivial metric changes in the image (Hummel, 1998).

What we propose is a representation of image features somewhere between completely unordered vectors and rigid templates, that is, a model in which there is implicit structural information regarding the spatial relations between local features. The form of the structural information is not a global description of an object in terms of parts or skeletons, rather, it is a relatively local description that captures the positional certainty between image measurements. In contrast to this type of statistical relation between features, structural-description models such as RBC relate far less local features, i.e. 3D parts, in a purely qualitative fashion (e.g. a part is simply ‘above’ a second part rather than more or less above). One way of coding these implicit relations is as set of weights within a neural network. For example, both Edelman and Weinshall (1991) and Lades et al. (1993) have proposed taking the output of receptive-field-like image filters and mapping them directly onto a recognition layer (Williams (1997) has developed a similar model that uses individual pixels as input). The weights between the input and output layers effectively code the likelihood of co-occurrence between local features. Thus, the relative positions of features are probabilistic, thereby providing ‘flexible’ or ‘deformable’ templates for recognition. Critically, metric variation between a known image of an object and a new image will not be catastrophic–recognition performance will degrade smoothly as the relative positions of features deviate further and further from their associated relations. The fact that the degree of match varies smoothly with changes in the image suggests that models incorporating local features in conjunction with implicit structural information may be compatible with view and class interpolation mechanisms, for instance, by computing the likelihood of each local feature at particular locations within the image (Edelman, 1995b).

The majority of models incorporating local image measurements in neural-network architectures have included only simple mappings between features (Edelman and Weinshall, 1991; Lades et al., 1993; Williams, 1997), e.g. one set of weights between input and output. As such, these implementations are still template models, albeit deformable, in that there is only a one layer description for each shape. One method for increasing the power of this approach is to add compositional structure to the representation (Bienenstock and Geman, 1995). In this framework local image-based features would be organized hierarchically into multiple levels of increasing complexity. For example, highly associated first order relations between image measurements could themselves be associated at the next level. However, in contrast to the reconstruction approach, these assemblies would be based on the statistics of the images shown to the system rather than fixed syntactic constraints.

One appealing aspect of compositionality is that it allows input shapes to be matched to stored representations through randomized tree searches (Amit and Geman, 1997). That is, rather than attempting to match all of the features of the representation during recognition, a series of binary ‘queries’ are performed. Each of these queries relates the position of one additional feature to the positions of features already queried. Critically, no particular set of features is required for successful identification–queries can begin with almost any feature (more informative features are selected during learning) and can follow many different search paths. Therefore, recognition should be robust over occlusion and other image variations. Equally important is that only a small number queries are likely to be necessary to recognize the input shape (as in the children’s ‘20-questions’ game). Thus, recognition should be computationally efficient.

A second appealing aspect of the compositional approach is that it allows for emergent structures at many scales within the image. Thus, more global representational elements, for instance, surfaces or parts, may arise at some level of the hierarchy depending on the co-occurrence of image measurements (Fukushima, 1980; Bienenstock et al., 1997). Indeed, because many image measurements are likely to co-occur repeatedly when one encounters the surfaces of a specific part, it may be possible to capture the part structure of most objects without the need for recovery or deterministic processes. However, such representations are still image-based in that the fundamental units of the representation are measurements of the image from a particular viewing direction–as such they are individually unlikely to remain stable over large changes in viewpoint or other viewing parameters. On the other hand, the inclusion of implicit structural information in the form of compositionality allows for more invariance than would otherwise be possible. To some extent this is precisely the goal of structural-description models (Marr and Nishihara, 1978), and indeed, differences between such models and our extended image-based approach come down to differences in the choice of features and the relations between such features. It is important to note, however, that these choices are critical for how each type of theory accounts for behavioral data.

6.6. Perceptual expertise

Our conjectures to this point have not addressed how the visual system achieves expertise with an object class. Consider the above framework in which object representations are comprised of features for which the spatial positions are more or less strongly related to one another. Two characteristics of this approach indicate that it may help provide an explanation for the phenomenon of perceptual expertise. First, features that co-occur more frequently will become more strongly associated. Second, extensive experience with the same features in a consistent configuration will give rise to more complex features. These simple statistical learning mechanisms offer an explanation for the configural sensitivity found in cases of perceptual expertise, including face recognition (Gauthier and Tarr, 1997a; Tanaka and Sengco, 1997). Consider that the acquisition of expertise is marked by extensive practice differentiating similar instances from within a class. Many class-level features will co-occur in the same configuration with great frequency, for example, the eyes, nose and mouth of human faces. Such oft-seen features will become tightly interdependent as the system is fine tuned by experience. Thus, relocating the position of one such feature will impact the recognition of the other features much as has been found for parts of human faces (Tanaka and Sengco, 1997) and for parts of non-face objects when recognized by experts (Gauthier and Tarr, 1997a). Moreover, because of compositionality, new, more stable configurations of features that have greater discriminatory power may emerge as an observer gains experience discriminating exemplars within a class (Gauthier et al., 1998).

7. Conclusion

Tremendous progress in understanding visual object recognition has been made over the past decade. Models of recognition have become far more computationally sophisticated. New and exciting findings from cognitive neuroscience and neurophysiology have offered insights into the brain mechanisms used during recognition. There has also been an impressive body of behavioral data collected on human recognition performance. Insights from all of these domains suggest that new theories hold great promise for explaining biological object recognition. At the same time, recent work has also illuminated some of the potential pitfalls of these theories. We have identified some of the most notable problems and offer possible solutions. What you will find in this special issue are the ideas of researchers that are working towards this goal, that of understanding visual recognition using a wide range of new methodologies.

Acknowledgements

The authors wish to thank Steven Pinker, Jacques Mehler and Susana Franck for their help and guidance throughout the development of this special issue and Talia Ben-Zeev, Guy Wallis, Chris Christou, John Hummel, and Bosco Tjan for their comments on this paper. This work was supported by a TRANSCOOP grant to MJT and HHB

References

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588.

Bartram, D.J., 1976. Levels of coding in picture-picture comparison tasks. Memory and Cognition 4, 593–602.

Beymer, D., Poggio, T., 1996. Image representations for visual learning. Science 272, 1905–1909.

Biederman, I., 1987. Recognition-by-components: a theory of human image understanding. Psychological Review 94, 115–147.

Biederman, I., 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.

Biederman, I., Gerhardstein, P.C., 1995. Viewpoint-dependent mechanisms in visual object recognition. Journal of Experimental Psychology: Human Perception and Performance 21 (6), 1506–1514.

Biederman, I., Ju, G., 1988. Surface versus edge-based determinants of visual recognition. Cognitive Psychology 20, 38–64.

Biederman, I., Subramaniam, S., Kalocsai, P., Bar, M., in press. Viewpoint-invariant information in subordinate-level object classification. In: Gopher, D., Koriat, A. (Eds.), Attention and Performance XVII. MIT Press, Cambridge, MA.

Bienenstock, E., Geman, S., 1995. Compositionality in neural systems. In: Arbib, M.A. (Ed.), The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA, pp. 223–226.

Bienenstock, E., Geman, S., Potter, D., 1997. Compositionality, MDL priors, and object recognition. In: Mozer, M.C., Jordan, M.I. Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA.

Blum, H., 1967. A transformation for extracting new descriptors of shape. In: Wathen-Dunn, W. (Ed.), Models for the Perception of Speech and Visual Form. MIT Press, Cambridge, MA, pp. 362–380.

Bricolo, E., Poggio, T., Logothetis, N.K., 1997. 3D object recognition: A model of view-tuned neurons. In: Mozer, M.C., Jordan, M.I., Petsche, T. (Eds.), Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA, pp. 41–47.

Bülthoff, H.H., Edelman, S., 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proceedings of the National Academy of Sciences of the United States of America 89, 60–64.

Corballis, M.C., Zbrodoff, N.J., Shetzer, L.I., Butler, P.B., 1978. Decisions about identity and orientation of rotated letters and digits. Memory and Cognition 6, 98–107.

Cutzu, F., Tarr, M.J., 1997. The representation of three-dimensional object similarity in human vision. In: SPIE Proceedings from Electronic Imaging: Human Vision and Electronic Imaging II, Vol. 3016. SPIE, San Jose, CA, pp. 460–471.

Edelman, S., 1993. Representing three-dimensional objects by sets of activities of receptive fields. Biological Cybernetics 70, 37–45.

Edelman, S., 1995a. Class similarity and viewpoint invariance in the recognition of 3D objects. Biological Cybernetics 72, 207–220.

Edelman, S., 1995b. Representation, similarity, and the chorus of prototypes. Minds and Machines 5 (1), 45–68.

Edelman, S., Bülthoff, H.H., 1992. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Research 32 (12), 2385–2400.

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

Farah, M.J., 1990. Visual agnosia: disorders of object recognition and what they tell us about normal vision. MIT Press, Cambridge, MA.

Farah, M.J., 1992. Is an object an object an object? Cognitive and neuropsychological investigations of domain-specificity in visual object recognition. Current Directions in Psychological Science 1 (5), 164–169.

Fukushima, K., 1980. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36, 193–202.

Gauthier, I., 1998. Dissecting face recognition: the role of categorization level and expertise in visual ob.ject recognition. Unpublished doctoral dissertation. Yale University.

Gauthier, I., Anderson, A.W., Tarr, M.J., Skudlarski, P., Gore, J.C., 1997. Levels of categorization in visual objects studied with functional MRI. Current Biology 7, 645–651.

Gauthier, I. and Tarr, M.J., 1997a. Becoming a ‘Greeble’ expert: exploring the face recognition mechanism. Vision Research 37 (12), 1673–1682.

Gauthier, I., Tarr, M.J., 1997b. Orientation priming of novel shapes in the context of viewpoint-dependent recognition. Perception 26, 51–73.

Gauthier, I., Williams, P., Tarr, M.J., and Tanaka, J., 1998. Training ‘Greeble’ experts: a framework for studying expert object recognition processes. Vision Research, in press.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends in Neuroscience 15 (1), 20–25.

Hayward, W.G., 1998. Effects of outline shape in object recognition. Journal of Experimental Psychology: Human Perception and Performance, in press.

Hayward, W.G., Tarr, M.J., 1995. Spatial language and spatial representation. Cognition 55, 39–84.

Hayward, W.G., Tarr, M.J., 1997. Testing conditions for viewpoint invariance in object recognition. Journal of Experimental Psychology: Human Perception and Performance 23 (5), 1511–1521.

Hummel, J.E., 1998. Where view-based theories break down: The role of structure in shape perception and object recognition. In: Dietrich, E., Markman, A. (Eds.), Cognitive Dynamics: Conceptual Change in Humans and Machines. MIT Press, Cambridge, MA.

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

Hummel, J.E., Stankiewicz, B.J., 1996a. Categorical relations in shape perception. Spatial Vision 10, 201–236.

Hummel, J.E., Stankiewicz, B.J., 1996b. An architecture for rapid, hierarchical structural description. In: Inui, T., McClelland, J. (Eds.), Attention and Performance XVI. MIT Press, Cambridge, MA. pp. 93–121.

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

Jolicoeur, P., 1985. The time to name disoriented natural objects. Memory and Cognition 13, 289–303.

Jolicoeur, P., 1990. Identification of disoriented objects: A dual-systems theory. Mind and Language 5 (4), 387–410.

Jolicoeur, P., Gluck, M., Kosslyn, S.M., 1984. Pictures and names: making the connection. Cognitive Psychology 243–275.

Kimia, B.B., Tannenbaum, A.R., Zucker, S.W., 1995. Shapes, shocks, and deformations, I: The components of shape and the reaction-diffusion space. International Journal of Computer Vision 15, 189–224.

Koenderink, J.J., 1987. An internal representation for solid shape based on the topological properties of the apparent contour. In: Richards, W., Ullman, S. (Eds.), Image Understanding 1985–86. Ablex, Norwood, NJ, pp. 257–285.

Kosslyn, S.M., 1980. Image and mind. Harvard University Press, Cambridge, MA.

Kovacs, I., Julesz, B., 1994. Perceptual sensitivity maps within globally defined visual shapes. Nature 370, 644–646.

Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.P., Konen, W., 1993. Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers 42, 300–311.

Lando, M., Edelman, S., 1995. Receptive field spaces and class-based generalization from a single view in face recognition. Network 6, 551–576.

Liter, J.C., 1995. Features affecting orientation-invariant recognition of novel objects. Unpublished doctoral dissertation. University of California, Irvine, CA.

Logothetis, N.K., Pauls, J., Poggio, T., 1995. Shape representation in the inferior temporal cortex of monkeys. Current Biology 5 (5), 552–563.

Marr, D., (1982). Vision: a computational investigation into the human representation and processing of visual information. Freeman, San Francisco, CA.

Marr, D., Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London B 200, 269–294.

Marsolek, C.J., Burgund, E.D., 1997. Computational analyses and hemispheric asymmetries in visual-form recognition. In: Christman, S. (Ed.), Cerebral Asymmetries in Sensory and Perceptual Processing. Elsevier, Amsterdam, pp. 125–158.

Miyashita, Y., 1993. Inferior temporal cortex: where visual perception meets memory. Annual Review of Neuroscience 16, 245–263.

Moses, Y., Ullman, S., Edelman, S., 1996. Generalization to novel images in upright and inverted faces. Perception 25, 443–462.

Nalwa, V.S., 1993. A guided tour of computer vision. Addison-Wesley, Reading, MA.

Perrett, D.I., Mistlin, A.J., Chitty, A.J., 1987. Visual neurones responsive to faces. Trends in Neuroscience 10 (96), 358–364.

Perrett, D.I., Oram, M.W., Ashbridge, E., 1998. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. Cognition 67, 111–145.

Pinker, S., 1984a. Visual cognition. Cognition 18.

Pinker, S., 1984b. Visual cognition: An introduction. Cognition 18, 1–63.

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

Quinn, P.C., Eimas, P.D., Rosenkrantz, S.L., 1993. Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. Perception 22, 463–475.

Riesenhuber, M., Poggio, T., 1998. Just one view: invariances in inferotemporal cell tuning. In: Advances in Neural Information Processing Systems 10. MIT Press, Cambridge, MA, in press.

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

Schyns, P.G., 1998. Diagnostic recognition: task constraints, object formation and their interactions. Cognition 67, 147–179.

Stankiewicz, B.J., Hummel, J.E., Cooper, E.E., 1998. The role of attention in priming for left-right reflections of object images. Journal of Experimental Psychology: Human Perception and Performance, in press.

Suzuki, S., Peterson, M.A., Moscovitch, M., Behrmann, M., 1997. Viewpoint specificity in the identification of simple volumetric objects (geons) is evident in control subjects and very exaggerated in visual object agnosia. Cognitive Neuroscience Society, Boston, MA.

Tanaka, J.W., Sengco, J.A., 1997. Features and their configuration in face recognition. Memory and Cognition 25 (5), 583–592.

Tanaka, K., 1996. Inferotemporal cortex and object vision. Annual Review of Neuroscience 19, 109–139.

Tarr, M.J., 1995. Rotating objects to recognize them: a case study of the role of viewpoint dependency in the recognition of three-dimensional objects. Psychonomic Bulletin and Review 2 (1), 55–82.

Tarr, M.J., Bülthoff, H.H., 1995. Is human object recognition better described by geon-structural-descriptions or by multiple-views?. Journal of Experimental Psychology: Human Perception and Performance 21 (6), 1494–1505.

Tarr, M.J., Gauthier, I., 1998. Do viewpoint–dependent mechanisms generalize across members of a class? Cognition 67, 71–109.

Tarr, M.J., Bülthoff, H.H., Zabinski, M., Blanz, V., 1997. To what extent do unique parts influence recognition across changes in viewpoint?. Psychological Science 8 (4), 282–289.

Tarr, M.J., Pinker, S., 1989. Mental rotation and orientation-dependence in shape recognition. Cognitive Psychology 21 (28), 233–282.

Tarr, M.J., Pinker, S., 1990. When does human object recognition use a viewer-centered reference frame?. Psychological Science 1 (42), 253–256.

Tarr, M.J., Williams, P., Hayward, W.G., Gauthier, I., 1998. Three-dimensional object recognition is viewpoint-dependent. Nature Neuroscience 1.

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

Ullman, S., 1996. High-level vision. The MIT Press, Cambridge, MA.

Ullman, S., 1998. Three-dimensional object recognition based on the combination of views. Cognition 67, 21–44.

Ullman, S., Basri, R., 1991. Recognition by linear combinations of models. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (10), 992–1006.

Wallis, G., 1996a. How neurons learn to associate 2D-views in invariant object recognition (Tech. Rep. No. 37). Max-Planck Institut für Biologische Kybernetik, Tübingen, Germany.

Wallis, G., 1996b. Presentation order affects human object recognition learning (Tech. Rep. No. 36). Max-Planck Institut für Biologische Kybernetik, Tübingen, Germany.

Williams, P., 1997. Prototypes, exemplars, and object recognition. Unpublished doctoral dissertation, Yale University.

Zhu, S.C., Yuille, A.L., 1996. FORMS: a flexible object recognition and modeling system. International Journal of Computer Vision 20 (3), 187–212.