RECONSIDERING THE ROLE OF STRUCTURE IN VISION

Elan Barenholtz and Michael J. Tarr

I. Introduction

How do we recognize complex objects? The problem of recognition can be summarized as determining how perceptual mechanisms map infinitely varying two-dimensional images into representations of a finite set of three-dimensional objects. Both learning and memory are intrinsic to this process; learning because inferences must be made about the regularities in the way objects vary across examples, and memory because theories of object recognition are in essence theories of how objects are instantiated as long-term visual representations. Thus, any account of the visual recognition of complex objects is an account of how observers learn about and remember the visual world.

In this chapter, we consider some fundamental arguments in favor of *structure* in mental representations and apply them to the particular challenges of visual object recognition. Structural accounts can be distinguished from what we will term pure "holistic" accounts such as simple template and some view-based models. These latter approaches achieve recognition by *globally* matching an incoming input whole cloth to some stored representation that does not distinguish individual features or their relations (e.g., Lades et al., 1993)—perhaps after performing some global transformation on the input or model (e.g., Ullman, 1989). According to these theories no individual features are differentiated; a match between an incoming image

THE PSYCHOLOGY OF LEARNING
AND MOTIVATION VOL. 47
DOI: 10.1016/S0079-7421(06)47005-5

0079-7421/06 $35.00

and a stored image is registered based on global similarity across a homogeneous input space—that is, each individual unit can play the same role as any other. Structural accounts can also be contrasted with ~~"pure" featural~~ accounts, which encode and recognize objects on the basis of some set of localized features but without *explicitly* encoding the position or relations among these features, that is, a "bag of features" (Mel, 1997).

It is important to point out that the term structure as we are using it here refers only to the question of whether relations and features are encoded separately—an issue that is orthogonal to the specific nature of the features or relations themselves. However, the term has also sometimes been recruited (perhaps inappropriately) in the empirical debate as to whether human object recognition is viewpoint dependent/viewpoint invariant (Biederman & Bar, 1999; Biederman & Gerhardstein, 1995; Hayward & Tarr, 2000; Tarr & Bülthoff, 1995). In these contexts, "structural" has been synonymous with viewpoint invariance, presumably because theories supporting invariance directly consider the properties of three-dimensional volumetric primitives. Indeed, the position of the second author of this chapter has often been mistakenly interpreted as antistructuralist because of his support for the claim that recognition shows viewpoint dependency. However, for the second author at least, the debate about viewpoint dependency was always about the nature of the *features* encoded in object representations (e.g., Geons or view-dependent local features) not about structural versus image-based accounts.[1] That is, image-based object representations were (and are) construed as collections of local, view-dependent image features that *may or may not be* related to one another in a structural manner (e.g., Edelman & Intrator, 2003; Tarr & Bülthoff, 1998). In other words, viewpoint-dependent features and volumetric features are *both* consistent with a structural architecture as we are defining it here. Consequently, the second author has taken issue with accounts of his view that paint it as "template-like" or "holistic" in the most simplistic sense (Hummel, 2000).

Intuitions about the critical role of relational information have played a part in a number of influential theories of visual object recognition (Biederman, 1987; Marr & Nishihara, 1978). According to both of these proposals, objects are represented and recognized on the basis of a set of generic volumetric (three-dimensional) primitives and their (explicit) spatial relationships. However, the focus of these theories—along with most other theories of recognition, including our own—has been on determining the nature of the molar features used by the visual system not on the nature of

AU:1

the spatial relations between such features. For example, Biederman's (1987) "Recognition-By-Components" model is quite specific about its vocabulary of three-dimensional primitives—Geons. In contrast, it makes only a brief mention of the way in which these features are related to one another and it eschews any strong theoretical justification for which relations are included, instead relying more on correspondence with verbalizable predicates, such as "on top of." The empirical research supporting these theories follows a similar pattern: the evidence for (and against) particular feature types is extensive, while the evidence for any explicit representation of relational information is somewhat sparse.

To some extent, this lack of data may be due to an inherent difficulty in establishing evidence for structure: by definition, structural representations require that features and their relations can vary independently; however, without *a priori* knowledge about what constitutes a visual feature there is no way to determine whether a particular experimental manipulation affects only one property and not the other. This "chicken-and-egg" problem raises serious challenges in examining structure directly and has not been successfully dealt with in the experimental literature. Instead, the primary strategy has been to use "likely" features that are chosen by the experimenter based on their plausibility. For example, the primary source of evidence for structural recognition concerns facial recognition in which the set of basic features are nameable parts of the face (e.g., eyes, nose) whose spacing is manipulated by translating the individual features in the image (for an example of attempts to leverage linguistic intuitions to study visual relations, see Hayward & Tarr, 1995). However, there are a number of models of recognition that include higher-order features that include more than one nominal part in a single feature (Ullman, Vidal-Naquet, & Sali, 2002; Zhang & Cottrell, 2005) in which case altering the spacing between the basic parts actually changes the features themselves.

One of the best pieces of evidence for structural models in visual recognition is text reading. Recent empirical evidence suggests that word recognition depends on identifying letters as opposed to recognizing the word holistically as a single pattern. Pelli, Farell, and Moore (2003) showed that words are unreadable unless their individual *letters* can be recognized independently. This means that words are recognized on the basis of constituent features, in this case letters; in addition, since we can quickly and efficiently distinguish between different words consisting of the same letters (e.g., "bat" $\neq$ "tab"), word identification clearly depends on representing the relations between these features. However, while this is an evidence for a form of structural recognition, it is unclear to what extent reading is representative of more general recognition; unlike other visual features, letters are initially learned as individual, separable symbols that contribute to a whole on the basis of their phonetic referents and

there appear to be regions of visual cortex specialized for letter processing (Cohen et al., 2000). In other words, they are explicitly structural in a way that other visual objects are not.

In recent years, the pendulum seems to have swung away from structural accounts to more image-driven theories of recognition. In particular, a number of theories in the computational and neuroscience literature have been proposed (Edelman, 1993; Riesenhuber & Poggio, 1999; Wallis & Rolls, 1997) that rely on a features-only strategy in which the features consist of image patches. Importantly, according to these theories the relations between individual features is not explicitly encoded, only the presence of a feature (in the location it happens to be) is considered as evidence. In this chapter, we reconsider a number of theoretical reasons why encoding features as *well* as some set of relations (i.e., structure) can be computationally advantageous. First, we attempt to define structure in its most general form as it has been defined within other domains of cognition and consider the application to visual recognition. Then we consider a number of varieties of structural theories of recognition and their potential application to specific problems in recognition. In addition, we consider some evidence—both experimental and also based on simple demonstrations—for structural representations in visual recognition.

II. Defining Structure

Structural accounts of recognition are those in which the identification of a visual object depends on both a set of features and the relations between those features within some representational space. Clearly, according to this definition, there are a large variety of potential theories that may be described as structural. "Features" in the visual domain can mean just about anything: image patches, object parts, or particular image properties (e.g., luminance), while "relations" between these features are any comparison between a set of individuated features. A meaningful feature refers to a property of a spatially localized region in an image, that is, it cannot be a global property of the entire image since relations can only be defined between two or more features. For the purposes of the discussion here, we will focus primarily on *spatial* relations between features, sometimes referred to as *configural* information. However, the general arguments we put forward can be applied to other forms of relational encoding as well.

A. Structure as Compositionality

The idea that the identification of a complex whole is dependent on some set of constituent features as well as their relations are related to the more

general concept of ~~*compositionally*~~ (Fodor, 1975; Fodor & Pylyshyn, 1988). In general, we may say that some complex object *X* is *composed* of some units *a,b,c*... *if and only if* the identity of X as *X* is contingent on the identities *a,b,c* and their relations within a space. For example, a grammatical sentence (in English) is composed of words since its grammaticality depends on the property of the words (i.e., their lexical category) and the relations between them (word order). A critical point to compositionality is that the properties of the units are defined *independently* of the composition. For example, the lexical category of the words does not change as a function of its place within the sequence (although some words belong to multiple categories). This latter property is central to compositionality since it allows for new compositions to be generated without new "emergent" properties that must be learned independently.

Compositionality is defined with regard to the assignment of some *property* to a complex entity such as a label, a meaning, or a referent. For visual recognition, a compositional approach is one in which the identification of some object—that is, corresponding to a specific name or category—depends on the object's features *and* their relations. Thus, it is not enough for the visual system to *extract* structure from visual input in a bottom-up manner (Markman, 1999; Marr & Nishihara, 1978); the assignment of a given property must be conditional on that structure. This is a critical distinction to keep in mind when considering evidence for structure in recognition. The primate visual system is clearly able to extract and represent structural information, for example, spatial relations, at a wide variety of scales. This is obviously true for images that contain multiple objects (we can describe, based on visual input, whether the cat is on the mat or the mat is on the cat). In addition, there is a great deal of evidence that the visual system naturally parses unitary objects into smaller units or "parts." Recent research on figure-ground assignment (Barenholtz & Feldman, in press), symmetry detection (Baylis & Driver, 1994), category learning (Goldstone, 2000; Schyns & Rodet, 1998), the movement of attention (Barenholtz & Feldman, 2003), and the perception of transparency (Singh & Hoffman, 1998) have all suggested a division of shapes into distinct parts.

AU:2

However, the ability and/or propensity to parse scenes into objects and parts and extract structural information is not the same thing as recognizing objects *on the basis* of their structure. There is a significant body of evidence that supports the use of featural information in mid- and high-level vision (Markman, 1999). Hinton (1979) demonstrated that mental imagery does include structural information without explicitly addressing how features come to be related to one another. But imagery is not recognition. For example, there is no doubt that we extract the eyes, nose, and mouth from a viewed face (and can potentially use such information in imagining a face);

however, this does not rule out the possibility that identifying an image as a face or recognizing a particular individual proceeds holistically—based, for example, on a template model of an entire face in which individual features are not differentiated (e.g., see Zhang & Cottrell, 2005).

Consider the following example: there is a natural tendency to see the large letter ''M'' in Fig. 1 as being constructed from smaller units—''H''s (such ''hierarchical'' letters have been used extensively in research concerning local versus global processing; e.g., Navon, 1977). However, we would not say that the ''M'' is *composed* of ''H''s—even though we can clearly discern these nominal ''features''—because the identity of the larger pattern as an ''M'' does not depend on the smaller units being identified as ''H''s. For example, if the ''H''s were interchanged with some other letter or blurred so as to be unrecognizable, this would not affect the identity of the ''M''. Alternatively, even if there is a structural description of an object inherently, this does not mean that the visual system employs it for recognition. For example, the square in Fig. 2 can be parsed into an arbitrary set of possible ''features'' with varying degrees of psychological plausibility. And, unlike the previous example, the identity of the square as such depends on the properties of these features as well as their configuration. However, it is unlikely that the visual system actually uses most of these structural descriptions (and might use none) in identifying squares.

This distinction between ''available'' structure and what we are calling compositional structure is important to keep in mind when considering evidence for structure in recognition. For example, Hummel (2000) argues against view-based theories of recognition based on the observation that we are able to compare images, such as the right and left images in Fig. 3, in a way that suggests structural encoding (e.g., we would say that both are formed by the same units, a circle and a triangle in different relations).

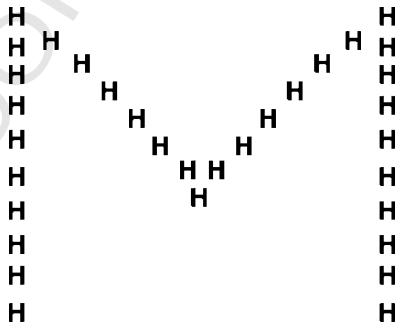

Fig. 1. An ''M'' built out of ''H''s but not *composed* of ''H''s.

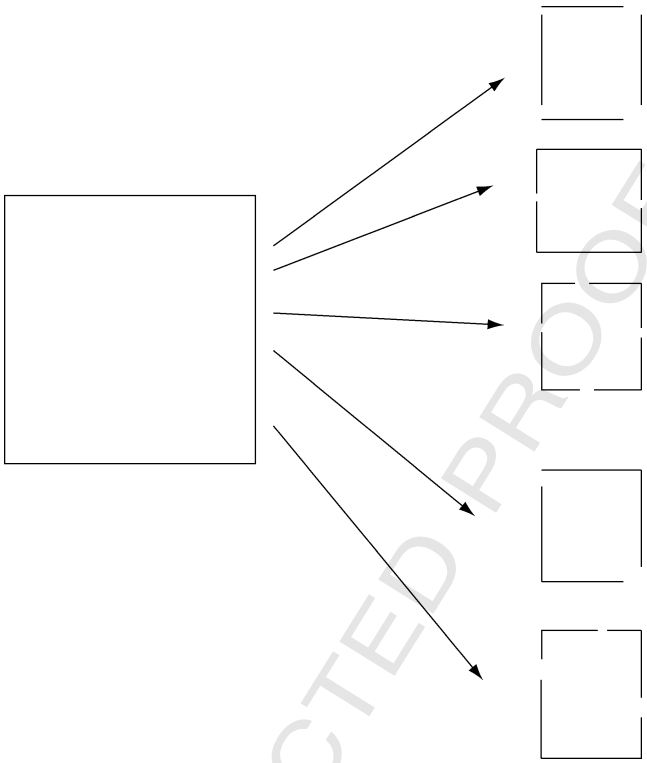

Fig. 2. Parsing a square into arbitrary "features."

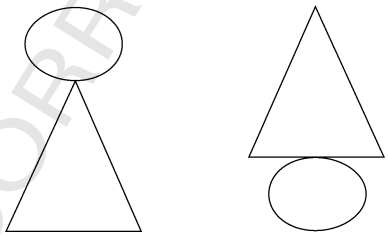

Fig. 3. Adapted from Hummel (2000). The same two features in different configurations.

However, this ability to extract and describe structural information in one context does not necessarily mean that this structure is used to assign a label or category—that is, the process of object recognition. Similarly, in another recent paper, Arguin and Saumier (2004) provide evidence that structural

representations are used in a visual search task. Specifically, search is observed to be more difficult when a target shares either parts *or* spatial configurations with distracters. However, visual search—finding a predefined target amongst similar distracters—makes very different processing demands as compared to *de novo* recognition. For example, object recognition is highly underconstrained in that you are unlikely to know *in advance* what you are looking at; even given some scene schema, the potential number of possible objects in enormous. In contrast, visual search typically predefines a single, unique, target and all an observer need to do is extract features from the image that disambiguate it from everything else in the scene. Thus, evidence for structure in one case is not necessarily evidence for structure in the other.

B. The Importance of Features in Structural Representations

An obvious issue arises with regard to the "M" in Fig. 1: even if the particular shape of the smaller units (i.e., "H"s) is not critical to the identification of the "M", successful recognition is ultimately dependent on the relations among *some* set of smaller units. Thus, the "M" might still be viewed as being composed of smaller features and their spatial relations. Along the same lines, all of visual recognition might be considered to be trivially compositional because it ultimately depends on the activation of localized units (i.e., ganglion cell receptive fields) whose relative position is clearly critical in encoding and recognizing an object (Edelman, 1993). In particular, the retina and downstream neural layers are organized spatiotopically such that proximity of activation on the retina reflects two-dimensional proximity in the real world. Moreover, it is clear that this structure is utilized by the visual system (to argue otherwise would be silly). A similar argument could be made for the presence of structure with regard to a template model that depends on the activity of individual receptors. However, it is our contention that these examples do not constitute genuine compositionality because these features, independent of their configuration, do not constrain the possible space of objects; *any* object can give rise to a response in almost any specific ganglion receptive field or photoreceptor. The locations that these units encode are only useful in that they contribute to larger patterns of activation based on the information across multiple units.

In other words, meaningful structural accounts, like pure feature-based accounts, require that the identity of the features be critical to the identity of the complex object. This can be captured by stating that in a structural representation both the features and the relations must contain diagnostic *information.* Information (in Shannon's sense) is a measure of the amount of uncertainty that is removed on the basis of some particular datum.

The information (or more precisely "self information," also called the "surprisal") of an individual datum is defined as the negative log of the probability of that datum's occurrence. For example, the result of a fair coin toss carries a certain amount of information because there is uncertainty between two possible outcomes ($\log_2(1/0.5) = \log_2(2) = 1$ bit), while the outcome of a dice roll carries more information because there are 6 possible outcomes ($\log_2(1/(1/6)) = \log_2(6) = 2.585$ bit).

With regard to recognition or categorization of an image, this construal of information can be used to measure the extent to which the presence of specific features and configurations constrain the space of possible objects in the world that may give rise to that image. For example, given a homogenous distribution of objects,[2] a high information feature would be one that only occurs in the presence of a small number of the objects—and thus whose presence eliminates a large number of potential objects—while a low information feature would be one that occurs in a larger number of objects—and thus whose presence eliminates a smaller number of potential objects. Thus, a visual feature that is common to *all* objects contains *no* information. The activation of any individual photoreceptor, while encoding some property of the distal stimulus, contains almost no information vis-à-vis recognition since nearly *any* object can presumably generate that activation under the right circumstances. A similar argument may apply to the oriented blobs that are characteristic of the receptive field responses in V1 neurons. Although these could potentially serve as candidate features in compositional system (Bienenstock, Geman, & Potter, 1997, treat oriented linelets as the basis of a compositional system) according to the requirement that individual features contain information, this seems unlikely in that virtually any object in the world that is free to move in space can produce a particular oriented edge. Thus, V1 may not be engaged in genuine "feature" extraction but instead may be functioning as a form of filtering that reduces redundancy in the image by only permitting certain patterns of activation to be further processed (Barlow, 1959) or encoding the statistics of natural images using Markov random fields (Black & Roth, 2005). This might simply serve to push the homogeneous input space—that is, in which any unit can play the same role as any other—higher up in the system rather than serving as a genuine compositional feature set.

Of course, just as the features in a structural representation must contain information, so must the relations among features. This means that for some

property to be based on compositional structure, it cannot be derived solely on the basis of the identity of the features alone. For example, if the presence of nose eyes and mouth in an image are sufficient evidence to decide that an image contains a face, than the identity ''face'' is not structural. In other words, in a compositional encoding, information is ''shared'' among the separable dimensions of feature-type and feature configuration. As we will discuss later, an important outcome of this is that the set of features needed to define a particular object will be more *generic* when configural information is included.

III. Why be Structural?

The computational utility of compositional structure has been considered in depth with regard to a number of cognitive domains, and in particular, with respect to language. A compositional architecture allows you to generate arbitrary constructions based on a limited set of starting units, a property referred to as *productivity* (Fodor & Pylyshyn, 1988). For example, natural language speakers are able to produce a theoretically infinite set of different grammatical sentences. Since it is not possible to learn every sentence in a language individually, speakers must have access to a compositional system consisting of a finite set of units (words) and rules for ordering them (syntax). By applying composition rules recursively, an infinite set of possible grammatical structures can be produced. Inversely, any grammatical sentence can be parsed into its constituent units and grammatical structure. A rigorous treatment of compositional linguistics was first proposed by Chomsky who showed that a formal language can be described (i.e., every sentence in the language can theoretically be generated) by a set of production rules and a finite set of symbols. While Chomsky's program relates exclusively to grammaticality, the idea of compositionality has been extended to the semantics of language and to our general cognitive capacities. For example, Fodor (1975) argues that thought itself is compositional, based in part on the observation that our mental capacity is open-ended (productive) in much the same way that language is, since we can understand and maintain beliefs about an infinite number of propositions (Fodor & Pylyshyn, 1988).

Is visual recognition productive in the same manner as thought and language? Despite some authors' suggestions to this effect (Edelman & Intrator, 2003), we believe that, strictly speaking, visual recognition is inherently nonproductive. Productivity refers to the capacity to generate or understand *new* compositions in which the set of known units are composed in a way that has not been encountered before (e.g., a novel sentence) but which can be understood based on our knowledge of the constituent units and the

composition rules. This capacity to understand new compositions, even those that are superficially radically different from those that have been encountered before is at the heart of Fodor and Pylyshyn's (1988) challenge to noncompositional architectures such as connectionism.[3] Object recognition, on the other hand, consists of determining that some viewed object contains the *same* (or nearly the same) compositional structure (i.e., a specific set of units and their relations) as one that has been viewed before, in order to assign that object to some predetermined category or label. There is no sense in which we are able to recognize some *new* composition of features except insofar as they are similar to some learned composition, an inherently nonproductive form of processing.

Instead, the utility of structural encoding is that it vastly increases the representational capacity of some restricted set of features—an argument cited by Biederman (1987) for using a limited collection of Geons plus spatial relations in his model. A structureless, feature-based approach can be described as representing each object in terms of an unordered set of features (e.g. $\{A,B,D\} \neq \{D,B,A\} \neq \{B,A,D\}$). Given a set of N unique features (i.e., the same feature cannot appear more than once in a given object) the number of possible representations is the number of subsets minus one or $2^n - 1$. Assuming that the *position* of the features is informational however (i.e., the sets are ordered so that {A,B} is distinguished from {B,A}), the number of possible objects is $\sum(k = 1{:}n)n!/(n-k)!$, where n is the number of features. For example, with 6 features you can represent 63 distinct objects based on a feature-only approach. If the sets are ordered, you can represent 1956 objects *based on the same set of features*. In other words, the category defined by the visual features can be further delineated on the basis of configural information. Of course, this number only reflects serial order which is analogous to allowing the encoding of left-of and right-of relations; it would grow more quickly in a spatial coordinate system that encodes above or below and *much* more quickly in an absolute coordinate system that measures distances between features. The potential utility of the increased capacity is twofold. First, it allows you to represent a much larger number of objects on the basis of the same set of features. This allows for a more economical representational system when the encoding cost of the features is higher than the encoding cost of the relations. Second, it allows you to capitalize on structural information that is already present in the image—information that pure feature-based strategies discard—in order to perform recognition when the features themselves might not be diagnostic.

To see this, we identify two broad classes of structural encoding that carry different potential utility for recognition: *category-specific* and *category-generic* structure. Category-specific structure is the application of structural information to a set of features that are bound to a particular category. For example, a given class of objects (e.g., elephants) may contain a unique set of visual features (e.g., big ears, trunk) that is common across multiple exemplars of that category. Moreover, these features may be specific to the particular category and be shared across all the members of that category. These features may be extracted *after* the initial identification has taken pace (e.g., based on a holistic strategy) or they may be the basis of the original identification (i.e., cases where the features, independent of their relations are diagnostic for the category).

Category-generic structure refers to structural information applied to a set of features that are common across multiple categories. For example, Biederman's (1987) theory relies on a small number of primitives to represent virtually all complex objects. As discussed earlier, features in a compositional representation must contain information; however, they do not need to contain as *much* information as in a strictly feature-based representation because the configural properties contain information as well. This means that the set of features used to identify an object can be less precise, that is, more generic, when structure is included. This provides two potential benefits. First, the total set of features used to recognize all objects can be much smaller because the same set of features can be reused to represent multiple objects. Note that this benefit applies even if one construes features as localized receptive field responses. Our best guesses about neural coding suggest that there is only a limited repertoire of selectivities within the primate visual system (Logothetis & Sheinberg, 1996); thus, it behooves vision theorists to account for how these relatively simple neural codes may be combined to represent the wide range of objects we encounter and recognize. Second, utilizing structure may allow recognition to proceed in cases where the features in the image are *inherently* "generic"—due, for example, to image noise, occlusion, or a rotation in depth. What follows is an attempt to define the particular role(s) of these two notions of structure in somewhat more detail.

A. Category-Specific Structure

Category-specific structure can be viewed as a form of "features-plus" encoding since it involves imposing configural information over a set of features that may already be diagnostic for a category. Feature-only representations are, in and of themselves, appealing for several reasons. Most feature-based strategies capitalize on the redundancy across images of objects in a particular category in order to identify a member of that category (e.g., Ullman et al.,

2002). For example, the category ''faces'' may depend on the presence of nose, eyes and mouth, which are visually similar across different examples of the category. Recognizing objects based on features, rather than whole images, carries a number of likely benefits. First, the number of potential stored images is smaller than in purely holistic systems—in which each object is stored as a unique image or set of images—since multiple objects can be represented on the basis of the same set of features.[4] Second, extraneous information can be ignored and attention may be paid only to informative features. This allows feature-based strategies to overcome image variability and effectively generalize across different images. For example, while different houses vary considerably from one another in terms of their global appearance, doors and windows tend to be fairly similar across houses. This means that in theory, you might be able to recognize a wide variety of images as houses, despite their large differences, on the basis of a small number of shared stored features.

However, feature-plus (i.e., structural) encodings provide a number of potential benefits beyond feature-only encoding. Ignoring relational information between features limits the power of a representational system in several ways. First, feature-based strategies may *overgeneralize* in cases where features are present but in the wrong configuration. For example, an image may contain two different objects that contain the diagnostic features of some third object. In addition, feature-based strategies are also limited in terms of the number of differentiated objects they can distinguish; to the extent that some class of objects is defined on the basis of some particular set of features, members *within* that category cannot be distinguished from one another. Instead, further delineation within a class defined by some set of features requires a new set of features (e.g., contrast the local features used to categorize faces as faces used by Ullman et al., 2002, with the more global features used to individuate faces used by Zhang & Cottrell, 2005).

Category-specific structural approaches can serve to overcome these problems. Overgeneralization is dealt with by *constraining* the possible space of images containing the appropriate features to those in which the appropriate relations hold between features (e.g., Amit & Geman, 1997). Structural encoding allows you to define relations with varying degrees of specificity.

[4] Although this claim might appear contrary to the ''multiple-views'' approach advocated by the second author (Tarr, 1995), this theory actually makes a case for ''views'' as sets of viewpoint-dependent features, not holistic entities or templates. Indeed, Perrett, Oram, and Ashbridge's (1998) ''evidence-accumulation'' model over multiple local features has been our preferred explanation for why systematic effects of viewpoint are observed in visual recognition tasks (Tarr, 2003).

For example, encoding configuration in terms of "coarse" or qualitative relations (such as discussed earlier) rather than in terms of specific metrical distances allows for a great deal of generalization flexibility while still preserving certain basic constraints. Alternatively, relations can be defined in a more sophisticated fashion based on the typical variability within a category. For example, we recently conducted a series of experiments in which subjects judged pairs of objects to be more similar when that were produced by *articulating* their parts in a physically plausible manner—rotated around their intersection or "joint" with the object—as compared to other objects consisting of the same parts but articulated in a nonbiological manner—rotated around the part's endpoint (Fig. 4). These results suggest a means of achieving recognition, that is, "invariant to articulations." Critically, this approach depends on explicitly modeling the spatial relations between object parts.

Category-specific structure also allows much higher representational power on the basis of the same feature set. As is well known from the categorization literature, a given object can belong to many hierarchically arranged categories, for example, the same object can be identified as an animal, as a

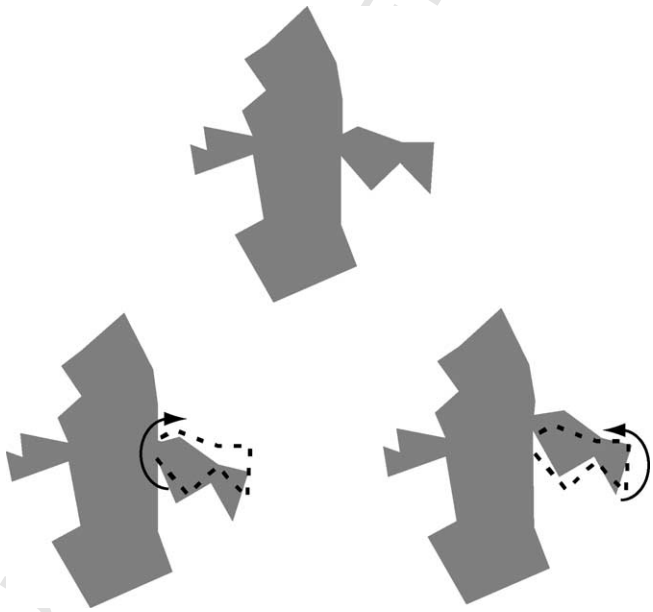

Fig. 4. The two bottom shapes represent valid (left column) and nonvalid (right column) articulations of a part of the top object. In the valid articulation, the part rotates around the endpoint where the part joins the "body"; in the nonvalid articulation, the part rotates around an endpoint on the opposite side.

bird, as a duck, and as Daffy Duck (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). To the extent that some set of features is diagnostic for a particular category, it cannot distinguish *between* members of that category. Category-specific structure can serve to ''push down'' the same set of features into more specific categories by imposing configural information over the shared set of features. This sort of structural encoding will typically rely on a metric (or ''second order'') encoding that allows you to define specific distances between features, rather than rough topological relations. This can dramatically reduce encoding cost for differentiating between members of a class of visually similar objects; even though there will always be some encoding cost to specifying the relations themselves, this will usually be much smaller than the cost of establishing new features for each new partition.

The most extreme examples of the potential economy of relational encoding are cases in which an observer may want to distinguish between *every* member of a particular class of similar objects in order to perform individual-level recognition (e.g., face recognition). In a feature-only system, each individual must be distinguished on the basis of a new feature, a clearly inefficient strategy when these are computationally costly features such as images. Relational information imposed on the original set of generic features, however, can provide the needed differentiation without introducing new features. For example, in Zhang and Cottrell (2005) informative spatial relations between parts are handled by creating larger noncompositional features. In our view, regardless as to whether this works or not, this is highly inefficient in that you may potentially need to create new features for each object to be recognized. A more efficient strategy is to actually encode the relations between *generic* features across the class that are potentially most informative with regard to individuation within that class. An example of this kind of discriminatory ability is so-called perceptual expertise (Bukach, Gauthier, & Tarr, in press)—the ability of experts to distinguish between members of a visually homogeneous object class at a specific (e.g., species or individual) level which untrained observers cannot distinguish reliably. Of course, the most prominent example of perceptual expertise in the visual domain is face recognition; different faces are visually similar to one another as compared with other objects, in part because they typically contain the same basic parts (eyes, nose, mouth, and so on), these features tend to be visually similar to one another across faces, and they appear in the same qualitative configuration across faces. That is why, as mentioned earlier, generic features (e.g., equating across individual noses) can serve as an effective strategy for detection. However, to the extent that the features are generic, they will not be able to distinguish between individual faces. A structural account—one that differentiates between individual faces on the

basis of spacing between the features—on the other hand, could theoretically encode many different faces on the basis of this same set of features and their relations.

This reasoning suggests that facial recognition—as opposed to face detection—is particularly well adapted to a structural recognition strategy. There is a sizable body of evidence that has been taken to suggest that facial recognition—as opposed to face detection—relies on structural encoding (Gauthier & Tarr, 2002; Maurer, Le Grand, & Mondloch, 2002; Tanaka & Farah, 1993; although see Riesenhuber, 2004). For example, in Tanaka and Farah's (1993) study, changing the horizontal position of the eyes affected the recognition of the nose and recognition of face parts in isolation was poorer than the recognition of the same parts embedded in the appropriate face. Both of these effects are consistent with the idea that the representation of an individual face includes spatial relations between parts; consequently, perturbing those relations or removing them altogether hinders recognition. Of course, although it is not our favored explanation, both effects would also be found if faces were encoded as unitary templates or as collections of large features (e.g., one feature encompassing both the nose and the eyes; Zhang & Cottrell, 2005). Thus, such results, although consistent with structural models, do not actually provide definitive evidence for such models. Similarly, one of the most often cited pieces of evidence for configural coding in face recognition—the "face inversion effect" (Yin, 1969)—may also be explained by the poorer match to larger *viewpoint-dependent* features or by the disruption of the processing of structural information. More specifically, the face inversion effect is typically defined as disproportionately (relative to nonface objects) poorer performance for both encoding and recognizing faces when inverted 180° in the picture plane. If the particular configural relations used to support individual face identification are view dependent, for example, "above" or anchored relative to gravitational upright then picture-plane inversions will disrupt one's ability to derive such relations. Alternatively, if the eyes, nose, and mouth are encoded as one large feature then rotated versions of faces will necessarily produce mismatches between input images and such global features. Furthermore, such higher-order features will typically contain less symmetry around the horizontal axis as compared to "standard" features (e.g., eyes and mouths which are typically fairly symmetrical around the horizontal axis), making it more likely that larger features would be susceptible to inversion effects.

There are other reasons to question whether these and related results actually provide support for structure in face recognition. First, the generality of the distinction between "featural" and "configural" processing in terms of the face inversion effect has been disputed based on both psychophysical and computational grounds (Riesenhuber, Jarudi, Gilad, & Sinha, 2004). Second, there are important questions concerning the interpretation

of some of these results and their relation to "everyday" face recognition. For example, a subset of studies demonstrating the use of configural information relies on a task in which pairs of highly similar images are compared after a brief interval, the subject's task is judging whether the two faces are the same or different (Le Grand, Mondloch, Maurer, & Brent, 2001; Maurer, Le Grand, & Mondloch, 2002). These face stimuli (e.g., "Jane" and her sister; Le Grand et al., 2001) would probably be judged as being the same individual by a naïve or casual observer. Thus, it is possible that observers in these studies are converging on a structural strategy—explicitly using spacing information—that is not engaged in typical facial recognition.

Finally, a potential red herring in these studies is that they often assume *a priori* that the features of a face are confined to nameable parts such as the eyes, nose, and mouth and that the spacing between them can be manipulated without changing the features of the face. For example, in Tanaka and Farah's (1993) study, they used stimuli consisting of simple "MacAMug" line drawings, containing contour information only for eyes, nose, and mouth; no surface texture or albedo (surface reflection) information was shown. Yet, as already discussed, plausible theories of facial recognition will almost certainly include features that encompass several of these "basic" features (Riesenhuber et al., 2004; Zhang & Cottrell, 2005) to form "higher order" features (e.g., a single feature could contain an eye and part of the nose), as well as surface information (Moore & Cavanagh, 1998; Vuong, Peissig, Harrison, & Tarr, 2005). In either case, line drawings of nameable face parts alone would constitute degraded stimuli, potentially prompting idiosyncratic strategies.

AU:4

B. Category-Generic Structure

Category-specific structure refers to cases in which the set of features is bound to a particular category. This presents the possibility that such features are extracted *after* category membership has been determined. For example, as mentioned earlier, it is possible for the extraction of the eyes, nose, and mouth of a face to take place *after*—and even be dependent on—the identification of the image as a face. Conversely, in the case of category-generic structure, it is possible that structural information can be utilized in order to determine category membership in the first place, based on much more generic features that are in and of themselves *not* diagnostic for any category. That is, structural information exclusively defines category membership, regardless of local feature appearance. Evidence for this kind of structural encoding can be found in a number of very simple examples. First, consider the case of the "smiley face" emoticon (Fig. 5A) which can easily be recognized on the basis of very sparse features. It is very clear that a strategy

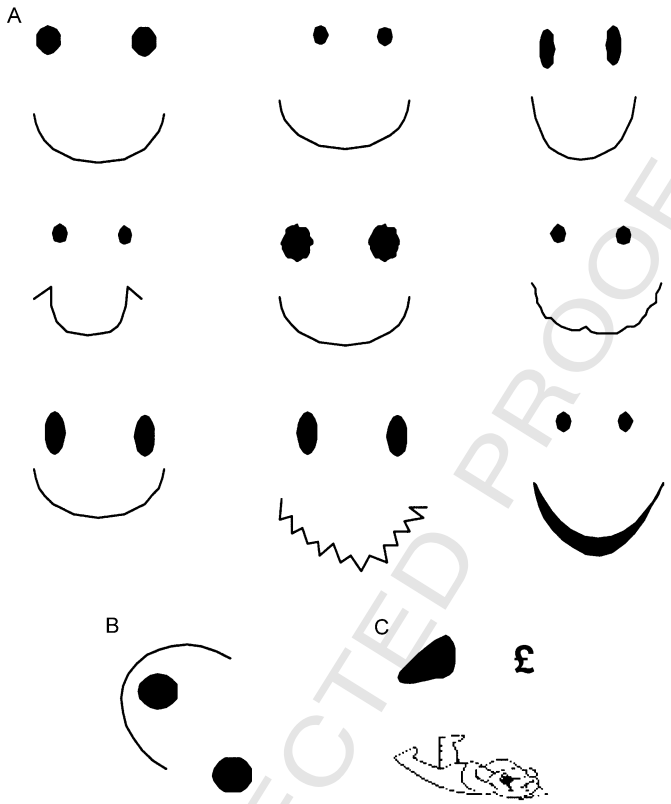

Fig. 5. (A) Valid smiley faces. (B) A scrambled smiley face. (C) A valid face configuration with bad features.

operating purely on the basis of the localized features in this image will not recognize the image as a face (the features are simply dots and curved contours). Such features, even taken together, contain very little information as they are consistent with almost an infinite set of objects. For example, the image in Fig. 5B contains the same features but is clearly *not* a face; the configuration of the features clearly matters. However, configuration is not the *only* thing that matters; as Fig. 5C makes obvious, not just *any* features in the appropriate configuration constitute a smiley face either. As required in our construal of structural representations, both the features and their relations appear to be critical. However, the range of potential features is very large and does not appear to be specific to particular objects. Of course, one might counter that our ability to recognize such images might depend on

a template-like representation in which both the appearance of local features and their relations matter. However, this alternative becomes somewhat less plausible when considering the wide variety of images—some of which you have probably never encountered before—that still look like a smiley face. As mentioned earlier, holistic or pure template representations are typically poor at generalizing to novel images.

Another counterargument to this particular example is that a smiley face is a highly stylized image and recognizing it may be more like interpreting *symbols* in a manner akin to text reading (which, as discussed earlier, is most likely structural) than to object recognition. However, the same basic phenomenon can be demonstrated for other, less familiar types of images a well. For example, it is easy to recognize the various images in the left-hand column of Fig. 6 (two faces and a pair of stylized people). However, as

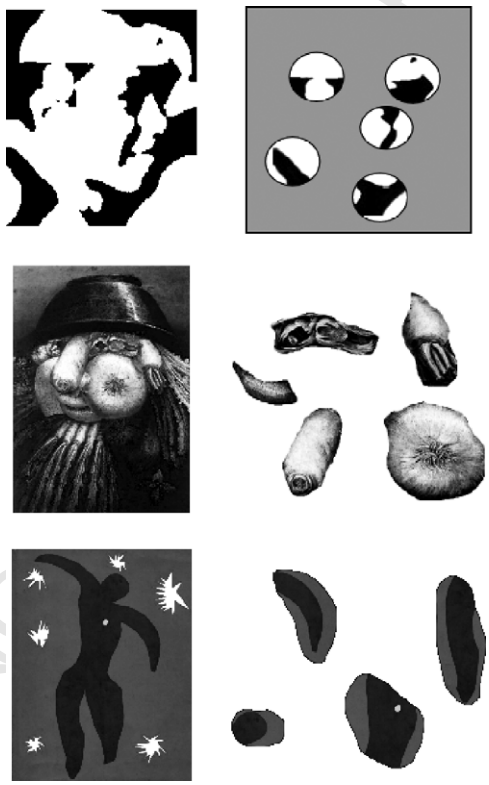

Fig. 6. The left column shows three easily recognizable objects. The right column shows that they are unrecognizable by features alone.

should be clear from the images in the right-hand column, the individual local features of these images are *not* recognizable ~~as a single object~~ outside of the configural context of the entire image (try inverting the page to produce a similar effect for the left-hand images).

Like the smiley face images, these examples pose a challenge to pure feature-based accounts. In particular, it does not appear likely that a local feature-matching scheme would arrive at the "correct" interpretation of the individual features in these examples (i.e., the one that you ultimately end up with after recognizing the whole). However, as with the smiley faces, you cannot arbitrarily replace these features with others—such as one vegetable with another (i.e., shape does matter)—and retain the same identity, just as you cannot replace the letters of a word and retain the same word. Instead, we suggest that functionally generic features, such as those in Fig. 6, do help to constrain the possible space of objects in the image but do not single out a unique object or part (e.g., the fact that it is a leek serving as the nose is irrelevant; on the other hand, it is an elongated cylindrical shape—replacing it with a mango would probably not work). For example, as illustrated in Fig. 7 the circular feature is a candidate for many different parts (e.g., a wheel, an eye) while the elongated oval feature may be a candidate for other parts (e.g., a dish antenna, a mouth). In a pure feature-based account any

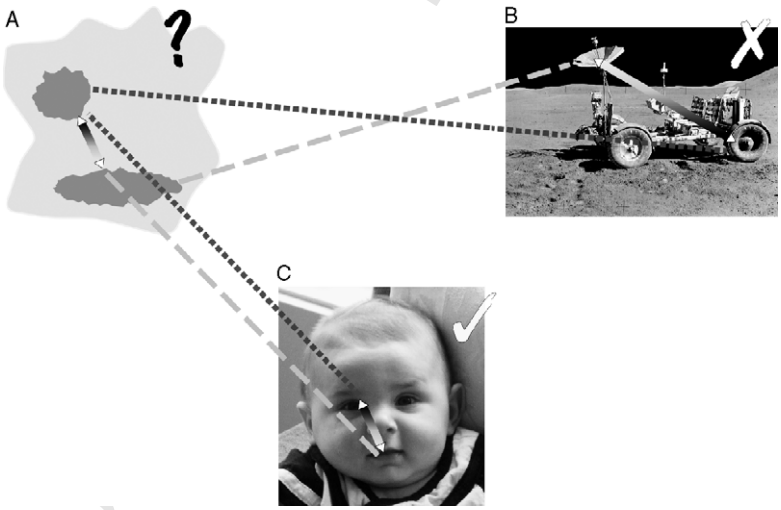

Fig. 7. (A) An ambiguous object containing generic features. (B) A potential match that contains the same features but in a different configuration. (C) A potential match that contains the same features in the same configuration. It is the structural information that distinguishes between these two candidates.

hypothesis for the identity of the entire object is likely to be based on the joint probability of the co-occurrence of these parts, given some object in the distal scene. For highly generic features, this calculation would be unlikely to yield the correct answer with any reliability since many objects are likely to match the generic features (Fig. 7B). However, in a structural encoding scheme, the hypothesized object would be the one that maximizes the probability of that set of features *and* their particular configuration (Fig. 7C). This provides additional constraint that may serve to identify the object with greater accuracy (e.g., Amit & Geman, 1997). This configural weighting could be instantiated in one of several ways. For example, in a ''global'' method, the goal would be to choose the *object* label that maximizes the likelihood of observing the features and their relations. The face in Fig. 7C might be chosen because the class ''faces'' maximizes the probability of these generic features in this particular configuration. Alternatively, a ''local'' method might have the goal of choosing, on a feature-by-feature basis, the individual object *parts* that maximize the likelihood of those features taking into account their occurrence in the image relative to *the position of other features*. This calculation could be realized by treating the features as interacting weights—for instance, as a Markov random field (Geman & Geman, 1984) in which individual features influence others based on their local spatial proximity or relational probabilities calculated over extended spatial ranges (Roth & Black, 2005). In such a scenario, the final object label would ''emerge'' based on the weightings of the local, interacting, features. The face in Fig. 7C might be chosen because eye and mouth are likely interpretations of the generic features given their spatial relationship. Regardless of whether local or global methods are used, the key point is that relational information may provide the leverage to achieve robust recognition based on generic features.

IV. Conclusions

It is our contention that the primate visual system is ''built'' to represent structure at multiple levels of processing. That is, from retinal processing to scene understanding, information about both what things are and where they there are is represented in one form or another. What is less clear is whether structural information in a truly compositional form is utilized in accomplishing the ''holy grail'' of visual abilities, recognition. Structural object representations once seemed like an obvious approach to recognition, yet even nominally structural models have emphasized the nature of features at the expense of a clearer explication of structural codes. At the same time, a number of challenges to such models have stymied the development of

better-specified theories of structure in both theoretical and experimental work. In the end, the particular features of the representation may be less crucial than the implementations of structure in successful models of visual recognition. At a minimum, simply relying on features, without their relational information, amounts to ignoring highly valuable information that is readily available to the visual system. As we have discussed, such information can serve to economize visual encoding by allowing the same set of features to be used far more flexibly as well as supporting robust recognition under conditions where feature-based strategies are likely to fail. Thus, our view is that the advantages of including structure are such that the question is not really *whether* structure is critical to object recognition but rather *how* structure is represented in a manner that makes visual recognition possible in the first place.

Acknowledgment

AU:5 This research was supported by NGA Award #HM1582-04-C-0051 to both authors.

References

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

Arguin, M., & Saumier, D. (2004). Independent processing of parts and of their spatial organization in complex visual objects. *Psychological Science, 15*, 629–633.

Barenholtz, E., & Feldman, J. (2003). Visual comparisons within and between object-parts: Evidence for a single-part superiority effect. *Vision Research, 43*(15), 1655–1666.

AU:6 Barenholtz, E., & Feldman, J. (in press). Determination of visual figure and ground in dynamically deforming shapes. *Cognition.*

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In "The mechanisation of thought processes." London*:* Her Majesty's Stationery Office.

Baylis, G. C., & Driver, J. (1994). Parallel computation of symmetry but not repetition in single visual objects. *Visual Cognition, 1*, 377–400.

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

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects. *Vision Research, 39*, 2885–2899.

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

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

Black, M. J., & Roth, S. (2005). *The Receptive Fields of Markov Random Fields.* Paper presented at Cosyne, Salt Lake City, UT, March 17–20.

AU:7 Bukach, C., Gauthier, I., & Tarr, M. J. (in press). Beyond faces and modularity. *Trends in Cognitive Sciences*.

Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M. A., et al. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain, 123*(Pt. 2), 291–307.

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

Edelman, S., & Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science, 27*, 73–109.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: MIT Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critique. *Cognition, 28*, 3–71.

Gauthier, I., & Tarr, M. J. (2002). Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance, 28*(2), 431–446.

AU:8 Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721–741.

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26*(1), 86–112.

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

Hayward, W. G., & Tarr, M. J. (2000). Differing views on views: Comments on Biederman & Bar (1999). *Vision Research, 40*, 3895–3899.

Hinton, G. (1979). Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science, 3*, 231–250.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157–185). Hillsdale, NJ: Lawrence Erlbaum.

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

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2001). Early visual experience and face processing. *Nature, 410*(6831), 890 (Correction *Nature* 412, 786).

AU:9 Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience* (Vol. 19, pp. 577–621). Palo Alto, CA: Annual Reviews.

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Lawrence Erlbaum.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260.

AU:13 Mel, B. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation, 9*, 977–804.

Moore, C., & Cavanagh, P. (1998). Recovery of 3D volume from 2-tone images of novel objects. *Cognition, 67*(1–2), 45–71.

Navon, D. (1977). Forest before the trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*, 353–383.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature, 423*(6941), 752–756.

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*(1,2), 111–145.

Riesenhuber, M., Jarudi, I., Gilad, S., & Sinha, P. (2004). Face processing in humans is compatible with a simple shape-based model of vision. ~~*Proc Biol Sci,*~~ *271*(Suppl. 6), S448–S450.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

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.

Roth, S., & Black, M. J. (2005, June). Fields of experts: A framework for learning image priors. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (Vol. 2, pp. 860–867).

Schyns, P. G., & Rodet, L. (1998). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(3), 681–696.

Singh, M., & Hoffman, D. (1998). Part boundaries alter the perception of transparency. *Psychological Science, 9*, 370–378.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46A*, 225–245.

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. (2003). Visual object recognition: Can a single mechanism suffice? In M. A. Peterson and G. Rhodes (Eds.), *Perception of faces, objects, and scenes: Analytic and holistic processes* (pp. 177–211). Oxford, UK: Oxford University Press.

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., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey, and machine. *Cognition, 67*(1–2), 1–20.

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

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*, 682–687.

Vuong, Q. C., Peissig, J. J., Harrison, M. C., & Tarr, M. J. (2005). The role of surface pigmentation for recognition revealed by contrast reversal in faces and Greebles. *Vision Research, 45*(10), 1213–1223.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. ~~*Prog Neurobiol,*~~ *51*(2), 167–194.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141–145.

Zhang, L., & Cottrell, G. W. (2005). Holistic processing develops because it is good. In B. G. Bara, L. Barsalou, and M. Bucciarelli (Eds.), *Proceedings of the 27th annual cognitive science conference*. Mahwah: Lawrence Erlbaum.

~~Further Reading~~

~~Tarr, M. J., & Kriegman, D. J. (2001). What defines a view? *Vision Research, 41*(15), 1981–2004.~~

Comp. by: Nadaraja Date:12/6/06 Time:17:53:28 Stage:First Proof File Path://
Spsind002s/Production/PRODENV/0000000001/0000000989/0000000016/
0000086026.3D Proof by:Sangeetha QC by: ProjectAcronym:bs:PLM
Volume:47005

Author Query Form

Journal: The Psychology of Learning and Motivation, 47
Article No.: Chapter 5

Dear Author,

During the preparation of your manuscript for typesetting some questions have arisen. These are listed below. Please check your typeset proof carefully and mark any corrections in the margin of the proof or compile them as a separate list. This form should then be returned with your marked proof/list of corrections to Elsevier Science.

Disk use

In some instances we may be unable to process the electronic file of your article and/or artwork. In that case we have, for efficiency reasons, proceeded by using the hard copy of your manuscript. If this is the case the reasons are indicated below:

☐ Disk damaged ☐ Incompatible file format ☐ LaTeX file for non-LaTeX journal

☐ Virus infected ☐ Discrepancies between electronic file and (peer-reviewed, therefore definitive) hard copy.

☐ Other: .....

We have proceeded as follows:

☐ Manuscript scanned ☐ Manuscript keyed in ☐ Artwork scanned

☐ Files only partly used (parts processed differently:.....)

Bibliography

If discrepancies were noted between the literature list and the text references, the following may apply:

☐ The references listed below were noted in the text but appear to be missing from your literature list. Please complete the list or remove the references from the text.

☒ Uncited references: This section comprises references which occur in the reference list but not in the body of the text. Please position each reference in the text or, alternatively, delete it. Any reference not dealt with will be retained in this section.

Query Refs.	Details Required
Author	**Further Reading section:** Some of your references may appear in a "Further Reading" section at the end of the chapter. These are references that were found to be uncited in the text, and have been called out to your attention. This is only temporary. To ensure that the called out references appear in the reference list, please cite them in the appropriate spot within the text. We appreciate your cooperation.

Query Refs.	Details Required	Author's response
AU1	Marr & Nishihara, 1978 is not listed in the reference list. Please check.	

Comp. by: Nadaraja Date:12/6/06 Time:17:53:28 Stage:First Proof File Path://
Spsind002s/Production/PRODENV/0000000001/0000000989/0000000016/
0000086026.3D Proof by:Sangeetha QC by: ProjectAcronym:bs:PLM
Volume:47005

AU2	There is no such reference as Baylis & Driver, 1995 in the reference list, rather Baylis & Driver, 1994 is present. Please check the change in year as per the reference list.	
AU3	Riesenhuber, 2004 is not listed in the reference list. Please check.	
AU4	Please check the change in spelling of the second author as per the reference list.	
AU5	This information has been shifted from the front page footnote to the Acknowledgment section as per style. Please check.	
AU6	Please provide the updated status of the reference.	
AU7	Please provide the updated status of the reference.	
AU8	Please check the insertion.	
AU9	Please check the change	
AU10	Please provide complete journal name.	
AU11	Tarr & Kriegman, 2001 is not cited in text. Please check.	
AU12	Please provide complete journal name.	
AU13	Please check page range.