Psychological Research (1999) 62: 131–153

ORIGINAL ARTICLE

Irving Biederman · Suresh Subramaniam
Moshe Bar · Peter Kalocsai · József Fiser

Subordinate-level object classification reexamined

Received: 28 August 1998 / Accepted: 7 September 1998

Abstract The classification of a table as round rather than square, a car as a Mazda rather than a Ford, a drill bit as 3/8-inch rather than 1/4-inch, and a face as Tom have all been regarded as a single process termed "subordinate classification." Despite the common label, the considerable heterogeneity of the perceptual processing required to achieve such classifications requires, minimally, a more detailed taxonomy. Perceptual information relevant to subordinate-level shape classifications can be presumed to vary on continua of (a) the type of distinctive information that is present, nonaccidental or metric, (b) the size of the relevant contours or surfaces, and (c) the similarity of the to-be-discriminated features, such as whether a straight contour has to be distinguished from a contour of low curvature versus high curvature. We consider three, relatively pure cases. Case 1 subordinates may be distinguished by a representation, a geon structural description (GSD), specifying a nonaccidental characterization of an object's large parts and the relations among these parts, such as a round table versus a square table. Case 2 subordinates are also distinguished by GSDs, except that the distinctive GSDs are present at a small scale in a complex object so the location and mapping of the GSDs are contingent on an initial basic-level classification, such as when we use a logo to distinguish various makes of cars. Expertise for Cases 1 and 2 can be easily achieved through specification, often verbal, of the GSDs. Case 3 subordinates, which have furnished much of the grist for theorizing with "view-based" template models, require fine metric discriminations. Cases 1 and 2 account for the overwhelming majority of shape-based basic- and subordinate-level object classifications that people can and do make in their everyday lives. These classifications are typically made quickly, accurately, and with only modest costs of viewpoint changes. Whereas the activation of an array of multiscale, multiorientation filters, presumed to be at the initial stage of all shape processing, may suffice for determining the similarity of the representations mediating recognition among Case 3 subordinate stimuli (and faces), Cases 1 and 2 require that the output of these filters be mapped to classifiers that make explicit the nonaccidental properties, parts, and relations specified by the GSDs.

I. Biederman (✉) · S. Subramaniam · M. Bar
P. Kalocsai · J. Fiser
Hedco Neuroscience Bldg, MC 2520,
University of Southern California,
Los Angeles, CA 90089–2520;
e-mail: bieder@usc.edu;
Tel.: (213) 740–6094 (Office), (310) 823–8980 (Home);
Fax: (213) 740–5687

Introduction

In their seminal (1976) paper, Rosch, Mervis, Gray, and Boyes-Braem argued that visual classifications are initially made at a "basic" level. We can roughly think of the basic level as that level of classification that people spontaneously employ to name a picture of an object, such as a "chair" or an "elephant." This is the level that maximizes cue validity in that it represents the best compromise in maximizing two quantities: (a) the distinctiveness between classes, and (b) the informativeness within a class. For example, elephants are highly distinctive from other basic level entities such as dogs, mice, tigers, and chairs. We also gain an enormous amount of information knowing that something is an elephant and not just an "animal." We do obtain more information from the *subordinate* classification in knowing that a particular elephant is an Asian elephant rather than an African elephant, but the modest amount of additional information in the subordinate classification comes at a considerable cost in distinctiveness: African elephants are not particularly distinctive from Asian elephants. The more abstract *superordinate* level, that of mammal (often termed "animal"), provides a high level of distinctiveness in that mammals are generally highly distinctive from fish or birds, but we lose greatly in informativeness in that we do not know which particular mammal is being specified.

The tradeoff between distinctiveness and informativeness can be appreciated with consideration of the subordinate-basic-superordinate level triples of Ford-car-vehicle or copperhead-snake-vertebrate.

Although the representations mediating basic-level classifications have been the subject of considerable theorizing (e.g., Biederman, 1987; Murphy & Brownell, 1985; Murphy & Smith, 1982; Rosch et al., 1976; Tversky & Hemenway, 1984), there has been little systematic work in exploring the perceptual representations mediating subordinate-level classifications, despite the importance of such representations in achieving expertise for the performance of many visual tasks, such as taking inventory, recognizing a person by his or her face, furniture shopping, bird watching, automobile identification, and target identification.

In this article we present an analysis of the perceptual requirements for basic- and subordinate-level visual classification of objects on the basis of their shape.[1] The analysis suggests that to a large extent, subordinate-level object classifications employ the same type of shape information as that employed at a basic level: *geon structural descriptions* (GSDs) (defined below). This information is typically, but not always, present at a smaller scale when subordinate classifications are required. The (possibly) special case of face individuation is considered separately. Prior to the analysis, we first consider the ambiguity that has resulted in failing to consider more thoroughly the perceptual basis for visual classification. We then review the evidence for the representations that we propose and suggest a neurocomputational basis for these representations.

Current status of subordinate-level classification

The literature on categorization posits only one level – the subordinate – at a less abstract level than the basic (or entry) level of classification. The employment of a single, common term has perhaps obscured the great variability in the perceptual demands required for distinguishing among different objects within a basic-level class (Biederman & Gerhardstein, 1993; 1995; Tversky & Hemenway, 1991). One consequence of the lack of differentiation of subordinate-level classifications is that investigators studying this domain have tended to overgeneralize the implications of their results. For example, the enormous difficulty and view dependence in attempting to distinguish among a set of bent paper clips is taken as characteristic of all subordinate-level classifications (Poggio & Edelman, 1990). In addition, the individuation of faces is often uncritically accepted as a prototypical subordinate-level object classification task, although there is some reason to believe, as will be argued later, that the identification of faces and the subordinate-level classification of objects may be mediated by different kinds of representations (e.g., Baylis, Rolls, & Leonard, 1987; Biederman & Kalocsai, 1997; Moskovitch, Winocur, & Berhrmann, 1997; Scalaidhe, Wilson, & Goldman-Rakic, 1997; Tanaka & Farah, 1993; but see Gauthier & Tarr, 1997).

What are Rosch's levels of classification levels of?

Objects are named faster with their basic-level than with their subordinate-level terms (Rosch et al., 1976). Part of the advantage of the basic level derives from the greater availability of basic-level names compared to subordinate-level names. Rosch et al. showed that basic-level terms appear first in the child's vocabulary, have fewer syllables, and are used much more frequently to refer to an entity than the subordinate-level terms. Basic-level concepts enjoy a visual advantage over superordinate level concepts as evidenced by Rosch et al.'s demonstration that members of a basic-level class, such as sofas, tended to have more similar shapes than members of a superordinate-level class, such as furniture. Rosch et al.'s assessment was made by superimposing silhouettes of exemplars from the different levels and noting that the basic-level composite image remained more identifiable than a composite of superordinate-level exemplars. The shape consistency enjoyed by the basic over the superordinate level and the perceptual advantage it confers are largely a consequence of the availability of common parts at the basic level (Tversky & Hemenway, 1984). Rosch et al. also noted that pictures from the same subordinate-level class were significantly more similar than pictures from a basic-level class, although that increase in similarity was not as large as an increase of basic- over superordinate-levels similarity. The lower shape variability of the subordinate-level class would be expected to make visual classification at that level easier, but this effect might be offset by the greater salience, in general, of the perceptual information required to distinguish entities at the basic level and, more importantly, the presence of more highly similar but different exemplars at the subordinate level.

For the most part, the study of levels of classification has not distinguished perceptual classification from naming or name-verification. Often we become aware that the two are not equivalent when we are unable to name an object (a "gizmo") or person that we otherwise recognize.

Part of the reason why naming an image at the basic level is faster than at the subordinate level is undoubtedly due to the greater frequency of occurrence and the fewer syllables, on average, of basic level *names*. Both word frequency and the number of syllables have been shown to correlate with naming reaction times (see Humphreys, Price, and Riddoch, in press for a recent review). Independent of the ease of access of an object's name, the basic-level advantage in naming, as noted

[1] Of course, visual information other than shape can be employed for classifying objects, particularly at a subordinate level, such as surface properties (e.g., color, texture, materials), position in a scene, and movement characteristics.

earlier, could also be a consequence of the perceptual information required for basic-level classifications simply being more discriminable or salient than the information required for subordinate-level classifications (Rosch et al., 1976). In addition, a representation of the concept of the object, mediating access to its name, might also be more readily activated in the case of basic-level concepts, a result that might be expected from the presumed greater frequency of making basic-level distinctions. Rosch et al. did not emphasize whether the basic level was the level that was perceived, conceived, or verbalized first, compared to the subordinate level. From Rosch et al.'s discussion, it would appear that both faster perceptual classification *and* faster access to names or concepts were intended.

A framework for expressing this distinction between perceptual and post-perceptual (e.g., name) classification is contained in the neural network for object recognition proposed by Hummel and Biederman (1992; described later). Hummel and Biederman posited an *object layer* (Layer 7 in Fig. 10), with individual units representing a structural description of the object (i.e., its parts and the relations among the parts). Units representing the object's name or other semantic information, such as its concept, where it is found, or how much it cost, could be associated with the object unit, but the ease with which the object unit could be activated from a particular image did not necessarily imply anything about the strength of the connections from the object unit to the name and semantic units. Thus, presentation of a "nonsense" object composed of several clear, simple parts in a novel arrangement could quickly recruit an object node and be rapidly activated on subsequent exposures without any activation of a unit for that object's name. It is the activation of the object node that corresponds to what Biederman (1987) termed *primal access* – the initial activation of a perceptual representation of the object.

How should variability among subordinate members of a basic-level class be represented?

Although it is possible for some basic-level classes, such as staplers, to posses only moderate degrees of shape variability, other classes, such as lamps, allow enormous shape variability. Indeed, the aforementioned gain that Rosch et al. (1976) noted in shape similarity of groups of subordinate-level objects compared to groups of basic-level objects could derive from the variability between different subordinates within the same basic-level class. For the low shape-variability class, it would be reasonable for classification to be initially at the basic level, with scrutiny required to distinguish among the subordinate members.

For a class of high shape-variability, such as birds, Jolicoeur, Gluck, & Kosslyn (1984) confirmed a conjecture of Rosch et al.'s (1976): Atypical members of a category were more quickly classified at their own "entry level" rather than at the basic level. For example, penguins and ostriches were classified faster as "penguins" or "ostriches," respectively, than they were classified as "birds."[2]

Biederman (1987), distinguishing between visual and verbal classification, proposed an even stronger modification of object classification. Visually, that is, with respect to an activation of an object node prior to the elicitation of a name, *all* visual classification is made at a subordinate level, given that the different subordinates were sufficiently distinct to initially activate their own units. In terms of the Hummel and Biederman (1992) network, these units would be in Layer 7 (L7). For a low variability class such as staplers, there may be only a few distinctive perceptual representations – that is, only a modest number of different sets of L7 units would be needed to handle the small number of exemplars (with different units in a set needed to represent different geons/relations for different views) – and hence the subordinate and the basic-level classes would be the same. For a high variability class such as lamps, there would be many different visual representations, that is, different L7 units, one set for ginger-jar lamps and another for pole lamps, for example, even though the same term, "lamp," is the first name that tends to be activated by any image of a lamp. Both kinds of lamps activate the same name, but they need not have. Pole lamps, like ostriches, *could* have had their own entry-level term, but do not. That is, one may say "lamp" faster than "pole lamp" to a picture of a pole lamp, but the visual representation that is activated for the pole lamp would not be the same as that activated for the ginger-jar lamp, as evidenced by the reduced priming for different shaped exemplars (e.g., Bartram, 1974).

There is actually little mystery in this common many-to-one, image-to-class mapping. Nothing prevents different perceptual representations, such as different shaped lamps, from activating the same class or name. In general, the facility with which the same name is activated from different perceptual representations will vary according to a host of factors such as the "typicality" of the object for its class and the pose or degree of view "canonicality" of the object (Palmer, Rosch, and Chase, 1981). From the current perspective, canonicality could arise from the ease of activation of a perceptual representation from the image, its distinctiveness, and the strength of the association between the perceptual representation and the name. A detailed analysis of the perceptual demands of subordinate-level classification is clearly critical for understanding these phenomena.

---

[2]In this paper, the term "basic level" will be retained, but with the entry-level qualification that the basic level need not accommodate highly atypical instances. That is, highly atypical instances of a basic-level class become their own basic-level class, even though technically they are members of the basic-level class. An indicant of such membership is whether the term "technically" itself can be applied to a classification. If it can, then we have an atypical member that forms its own entry-level class. Thus, it is acceptable to say "technically, a whale is a mammal," but it would not be acceptable to say "technically, a bear is a mammal."

Perceptual requirements for basic- and subordinate-level visual shape classification

Geon-structural descriptions

Nonaccidental Properties (NAPs; Lowe, 1984) are qualitative properties of images (hence they are 2D) that tend not to change with small rotations of the object in depth. With respect to image edges corresponding to discontinuities in surface orientation and depth, these properties can be expressed contrastively, such as whether an edge is straight or curved, approximately parallel or nonparallel, or the type of vertex that is formed from the cotermination of edges.[3] NAPs can be distinguished from *metric* properties, such as aspect ratio, degree of curvature, or the different acute angles between two segments of a bent paper clip, that do vary with the object's orientation in depth.

A GSD is a 2D representation of an arrangement of parts, each specified in terms of its nonaccidental properties (geons) and nonaccidental relations between the parts. (See Hummel & Biederman, 1992, for a more precise definition.) To the extent that the objects have part structures where, across objects, the parts (or geons) differ in NAPs and/or invariant relations with respect to the other geons (e.g., end-to-end vs. end-to-middle connected; smaller-than vs. about-same-size vs. larger-than; above vs. side-of vs. below), then they can be said to have distinctive GSDs (Biederman & Gerhardstein, 1993; Hummel & Biederman, 1992). The nonaccidental specifications, to the extent that they are distinctive and can be resolved, allow invariance in the recognition of an object at different orientations in depth.

In a minimal case of subordinate-level discriminations, one object may be distinguished from another by a speck that is present in one case but absent in the other (presence-absence is a nonaccidental property, Jacobs, 1997). If both objects had specks, then they would be distinguished by a GSD if one was pointy and the other rounded, for example, but not if they were both highly irregular and so could not be distinguished by a nonaccidental difference. In these examples, note that a generalized cylinder (Binford, 1971) is not explicitly activated, merely a surface. Any surface (or line) could be an aspect or feature of a generalized cylinder (Dickinson, Pentland, & Rosenfeld, 1992) but they could, of course, just be planar entities, as Biederman (1987) allowed in his presentation of geon theory.

[3]Because of perspective convergence, parallel lines in the object will converge if extended in depth. However, there is a strong perceptual bias to interpret *approximately* parallel lines as parallel if, given uncertainty as to the true slant of the lines in depth, they could be parallel (Biederman, 1987). Jacobs (1997) has argued that there are an infinite number of nonaccidental properties for configurations of five or more points. However, those nonaccidental properties that are salient are those that are defined for a minimal number of features (typically not more than three points).

The role of distinctive GSDs in basic- and subordinate-level object classifications

Edelman and Bülthoff (1992) studied recognition of a set of ten objects, each consisting of five elongated cylinders joined end-to-end. The objects differed only in the angle of their joins, so they looked like bent paper clips. Such objects are extraordinarily difficult to distinguish when they are presented at an orientation in depth that differs from an originally studied view. The new orientations must themselves be learned. Biederman and Gerhardstein (1993, Exp. 5) showed that substituting a different geon for the middle cylinder of each of the paper clips, so that they now resembled geon "charm bracelets," rendered them readily identifiable from arbitrary orientations in depth. That is, the distinctive GSDs allowed the objects to achieve near-depth invariance without learning. Biederman and Gerhardstein (1993) argued that differences in distinctive GSDs form natural boundaries between concepts. When such differences are absent, as they are with a set of bent paper clips, people do not spontaneously distinguish them, and it is doubtful that any culture would create basic- or subordinate-level distinctions across such stimuli (Biederman & Gerhardstein, 1995).

GSDs versus "view-based" accounts

Before considering recent evidence supporting GSDs, we will first consider a class of theories that have been termed "view-based" by their proponents. There have been a variety of such proposals, but, as applied, these theories essentially assume a template in which neither NAPs nor parts nor GSDs are provided any special status (e.g., Edelman & Bülthoff, 1992; Ullman, 1996). The essential bit of data cited in support of these theories is the cost in recognition time or accuracy when an object is viewed at an orientation in depth that differs from a previously experienced pose and the reduction in costs when that new pose is presented again. According to one author (Tarr, 1995), for nearby views, the costs reflect direct interpolation of the template between familiar views or extrapolation to a new view, but a more costly normalization process, akin to mental rotation, is required for recognition at greater rotation angles.

The issue, in our judgment, is not whether object recognition is view-based. *All* object recognition is view-based, hence the use of quotation marks around the term "view-based" in this section's title. As one of us has noted previously (Biederman & Gerhardstein, 1995), "view-based" is only worth arguing when the alternative is extra sensory perception (ESP). That performance may improve with repeated presentations of the new views does not imply anything about the representation mediating such views, whether it is a template or a GSD. The real issue that needs to be decided is *representation*. Even if we just consider the costs of rotation to a new orientation, it is clear that different sets of stimuli produce

dramatically different costs for a given rotation angle (Biederman & Gerhardstein, 1993). The costs in recognition of a depth-rotated object could, as many view-based theorists have attempted to demonstrate, reflect distortions of a template. However, the rotation could, as well, produce changes in an object's geon structural description because, for example, the geons and their attributes (e.g., coarse changes in aspect ratio and orientation) are occluded, or new geons are revealed, or the relations among the geons are altered (Biederman & Gerhardstein, 1993). All of this must be understood in terms of a resolution function specifying the time requirements to determine, at some level of accuracy, an object's parts and relations under given presentation conditions (e.g., duration, contrast, noise). This point was made by Biederman and Gerhardstein (1993) in discussing the need for a principled quantitative analysis to determine the rotation costs for activating GSDs: "Such an analysis would have to include a resolution function in that a part need not completely appear or disappear as a result of an orientation change before the change will begin to affect performance" (p. 1180).

As noted in discussing the Biederman and Gerhardstein (1993) study of the effects of adding a distinctive geon to each member of a set of bent paper clips, the gigantic effect in this domain of research are the differences in rotation costs depending on the kind of information that distinguishes the stimuli to be identified. When the stimuli cannot be distinguished by GSDs, enormous rotation costs are evident. In fact, performance accuracy in same-different matching tasks for such stimuli are often below chance! (see Biederman & Bar, in press; 1998.) The presence of distinctive GSDs allow an enormous reduction in rotation costs to where they are small, if not absent.

Evidence cited in support of view-based accounts

In failing to provide any explanation for the extraordinary large benefit offered by distinctive GSDs, view-base theories are in danger of making a claim – that recognition requires perception – which distinguishes no theory of shape recognition. However, we can ignore whether "view-based" is a vacuous theoretical position and consider two empirical claims that some view-based theorists have raised as a challenge to geon theory: (a) the presence of rotation costs for stimuli that differ in GSDs, and (b) the lack of a sizable difference in rotation costs between stimuli that do and do not differ in GSDs. We will consider each of these claims in turn.

Are there sizable rotation costs for stimuli that differ in GSDs?

A number of studies have documented rotation costs where the rotations occluded some geons and revealed others or produced accidental or near accidental views (e.g., Humphrey & Kahn, 1992; Srinivas, 1993). Michael Tarr and his associates (Tarr, 1995; Tarr, Bülthoff, Zablinski, & Blanz, 1997; Haywood and Tarr, 1997; Tarr, Williams, Haywood, & Gauthier, 1998) have recently reported rotation costs when accidental views were, presumably, controlled. The magnitude of these costs were small, as they were in the Biederman and Gerhardstein (1993) experiment, relative to the rotation costs incurred with stimuli that do not differ in GSDs, as discussed in the next subsection. For example, Tarr et al. (1998) studied the recognition of rendered single geons adapted from Biederman and Gerhardstein's (1993) Exp. 4 with line drawings. [One of Tarr et al.'s (1998) nine experiments did use line drawings.] If the slope of the plot of reaction time against rotation angle (from 0° to 90°) is expressed in °/s, then the rotation rates for these stimuli ranged from approximately 750°/s for a naming task to 3,600°/s for a Match-to-Sample task with a Go/No-Go response. For some experiments, Biederman and Gerhardstein reported flat functions or effective orientation costs of only 5,000°/s.

Do these slopes, shallow as they are, represent fundamental view-dependence as would be expected from, for example, mental rotation or the extrapolation or interpolation of templates, or do they represent variations in extracting GSDs at different rotation angles? Although the possibility of a template-like representation cannot be definitively ruled out to account for some of these rotation costs, there are a number of factors other than template mismatching that easily could have contributed to these costs. Despite the attempt at avoidance of accidental views, many of the views were, in fact, near accidents that required, for example, determination of whether a single small contour was straight or slightly curved, as Biederman and Gerhardstein noted. In Biederman & Gerhardstein's (1993) Exp. 4 (Go No-Go, match-to-sample of single geons), whereas most of the distractors never elicited a false alarm, some had false alarm rates of 60 to 100%! Whereas Biederman and Gerhardstein's subjects were induced to respond quickly and evidenced almost no rotation costs, but a 15–20% false alarm rate, Tarr et al.'s (1998) subjects responded far more slowly but with a false alarm rate of only 5% and a (modest) slope of 2,250°/s. It is likely that subjects in the Tarr et al. (1998) experiments were taking the time to resolve the small differences in contour needed to reject a near distractor.

Biederman and Bar, (in press; 1998) have noted a number of other artifacts in experiments that have reported rotation costs with stimuli that differ in GSDs. The essential point here is that rotation in depth tends to produce drastic changes in the 2D image that can differentially affect the perceptibility of the parts. Rendered images, as compared to line drawings, can easily yield lower contrast and noisy illumination and shadow contours at the orientation and depth discontinuities important for resolving the geons. This appears to be especially true in the Tarr et al. (1997) and Haywood and Tarr (1997) experiments. As an object is rotated in

depth, these effects can vary for different geons. In same-different matching tasks, transients are produced when rotated stimuli no longer occupy the same regions of the screen so the absence of a transient is a reliable cue that unrotated stimuli are the same. Biederman and Bar (1988) showed that the rendering effects can be reduced by increasing stimulus presentation durations. The impact of the transient is reduced by shifting all stimuli, even when they are not rotated. The shift increases the difficulty of the 0° rotation condition relative to the positive rotation conditions, thus producing lower rotation costs. Biederman and Bar argued that these transient shifts were the reason why, in the Haywood and Tarr and Tarr et al. (1997) studies, a rotation from 0° to a slight angle, say 30°, produced greater costs than rotations from greater angles, say from 60° to 90°. The opposite would be expected from the template extrapolation/mental rotation routines argued by Tarr (1995). Biederman and Bar (in press) reported virtually no effect of rotation in a same-different matching task for rendered novel two-geon objects that differed in their GSDs when longer exposure durations and shifted positions were employed.

There is independent evidence that these types of resolution variations may be sufficient to produce the observed rotation costs. Curiously, most such experiments have studied relatively small rotation angles, up to 90° and, in some cases, only to about 30° (Haywood & Tarr, 1997). From a "view-based" perspective, a rotation of 180° or mirror reflection of a bilaterally symmetrical object would be expected to produce enormous rotation costs, relative to these slight rotations angles. The opposite, however, occurs. Mirror reflections incur *no* cost in priming in people (Biederman & Cooper, 1991a, b; Stankiewicz, Hummel, & Cooper, 1998) and monkeys (Logothetis, Pauls, Bülthoff, & Poggio, 1994). If one assumes that the object is bilaterally symmetrical, then an algorithm developed by Vetter and Poggio (1994) can match mirror reflected images without a costly normalization procedure. However, the application of such an algorithm would produce no costs for rotation to any angles.

Do distinctive GSDs offer a benefit in reducing rotation costs?

There would seem to be no question that when a set of objects lack distinctive GSDs, the ability to recognize them from arbitrary viewpoints would be much worse than when distinctive GSDs are present. Thus, Rock and DiVita's (1987) subjects were at near chance levels in recognizing which of two smooth complex novel wire objects they had seen previously. Anyone who has tried to recognize a rotated bent paper clip from among other bent paper clip distractors, of the kind studied by Edelman & Bülthoff (1992), quickly realizes the extraordinary difficulty in performing such a task compared to charm bracelets.

However, are GSDs an appropriate representation to characterize this advantage of stimuli that differ in geons and relations compared to those that do not so differ? Tarr et al. (1997) performed a same-different matching task with rendered versions of the Biederman and Gerhardstein's (1993) charm bracelets and a comparable set of paper clips. As would be expected from the previous discussion, the charm bracelets were far easier to recognize under rotation: At 90° the d' for the charm bracelets with a single distinguishing geon was approximately 3.0; for the paper clips it was 0.5. Tarr et al. (1997) also included charm bracelets with three or five different geons (sampled from a set of ten). The additional geons reduced performance so that with five different geons the d' at 90° was 2.0 (still markedly greater than the d' for the paper clips). Tarr et al. (1997) interpreted this last result as evidence against GSDs, insofar as the additional geons did not facilitate performance. However, this interpretation is mistaken. Hummel and Biederman (1992) had argued that their network would not be able to distinguish a linear array of three geons, much less five. Part of the reason is that with single place predicate relations of the kind assumed by Hummel and Biederman (e.g., a cylinder side-of another geon), the inner orders of geons are not distinguished. More generally, with the identical set of side-of relations for all geons in all stimuli and ten five-geon subsets of the same ten geons, the vectors describing the different objects would be highly similar, thus reducing their discriminability. The consequence of this is that one would have to employ complex rules to distinguish the stimuli, such as: if the middle geon is a cylinder and one of the end geons is a wedge, then if the geon on the other side of the cylinder is a brick, it is object A, but if that geon is a cone, it is object B. Biederman and Gerhardstein (1993) explicitly argued that stimuli from such sets are not distinguishable by GSDs.

Biederman and Bar (in press) reported a critical same-different experiment comparing the detection of two-geon stimuli (see Fig. 3) that differed in a nonaccidental or in a metric property of a single part. Subjects did not know which part, if any, would be changed, nor in which manner (geon or metric). The second stimulus was always shifted with respect to the first, even when it was identical and not rotated. Subjects saw a given pair of stimuli only once. Rotations of $\approx$60° produced virtually no costs in detecting geon differences, but the detection of metric differences was below chance.

Last, Biederman and Bar (1998) investigated the same-different matching of a set of bent paper clips of the kind studied by Tarr et al. (1997). There were striking differences in the miss and false alarm rates at the identical rotation angles for individual pairs of images. Put simply, if the images projected by the first and second stimuli differed in a qualitative feature, such as an arrow vertex for one and a near linear array for the other, then the subject tended to respond "different,"

producing high miss rates (>50%) when the objects were the same and relatively low false alarm rates when the objects were actually different (27%). When the images did not differ in a qualitative feature, then the subjects tended to respond "same," producing high false alarm rates (as high as 88%!) when the stimuli were different and low miss rates when they were the same (as low as 4.6%). Despite a large number of paper clip experiments, to our knowledge, this article is the first published revelation of these enormous differences which cannot be handled by models that do not distinguish these features. GSDs would appear to be an apt representation not only for charm bracelets but for these paper clip stimuli as well.

Confirmation of this conclusion can be found in the observations of view-based proponents (in the restrictive sense of "view-based") themselves. In training monkeys to respond to a particular object at varied orientations, Logothetis et al. (1994) noted:

> … When the wire-like objects had prominent characteristics, such as one or more sharp angles or a closure, the monkeys were able to perform in a view-invariant fashion, despite the distinct differences between the two-dimensional patterns formed by different views. … The animals easily learned to generalize recognition to all novel news of basic objects [such as a teapot or spaceship]. … The objects were considered 'basic' because of their largely different shape from the distractors. … The monkeys had never seen these objects before. … So, their remarkable performance may be the result of quickly learning…some characteristic features of the objects, for instance, the lid's knob or the handle of the teapot, or some relationship between such features and a simple geometrical shape, endowed with an axis of symmetry (p. 411).

Recent evidence for the role of GSDs in basic- and subordinate-level classification

Biederman (1987; 1995) presented a number of studies documenting the empirical support for GSDs. Briefly, these can be grouped into studies demonstrating the importance of parts, on the one hand, and simple, nonaccidental shape differences, on the other. The latter class of results includes those which demonstrate that other variation is of little or no consequence in object recognition.

Evidence for a parts-based representation

A. Verbal descriptions of objects

When asked to list the characteristics of basic-level objects, such as chair, car, or elephant, people list the parts of the object (Tversky & Hemenway, 1984; Rosch et al., 1976). Both Rosch et al. and Tversky and Hemenway (1984) noted that an object's parts often provide much of the functionality of an object class. Tversky and Hemenway (1991) also noted that familiar subordinates are also distinguished by their parts. It is not obvious how one maps a template representation of an object to a representation of the object's parts.

B. Priming complementary images

Biederman and Cooper (1991b) produced pairs of complementary contour-deleted line-drawings of common objects in which every other vertex and line was deleted from each part so that the same parts could be activated from either member of a pair. The members of a complementary pair would produce an intact original line drawing if superimposed. After naming one of the briefly-presented images in an initial priming block, subjects named either the identical image, its complement, or a same name-different shaped exemplar. There was a considerable advantage of both the identical and complementary images in both reaction times (RTs) and error rates compared to the different shaped exemplars, indicating that a large portion of the priming was visual, and not just verbal or conceptual. Most important, the identical and complementary conditions were equivalent, showing a lack of contribution of the lower level features (lines and vertices) in the representation of the object (Biederman & Cooper, 1991b). This same study also ruled out any role for a top-down effect in visual priming in this experiment by showing that there was no visual priming between parts-deleted complements in which half the parts were removed from each image. Biederman and Cooper (1991b) noted that the equivalence of identical and complementary feature images were not limited to a priming effect: Despite the complete non overlap in features (lines and vertices), members of a complementary pair, at first glance, *look* equivalent. Scrutiny is required to appreciate that they are different. Biederman and Cooper (1991b) concluded that the lines and vertices were required to activate the representation of the parts, but the representation that mediated priming specified the parts, not the particular vertices and lines present in the original image. Later we review research (Kalocsai & Biederman, 1997) indicating that only a portion of these differences can be accommodated by routines for smooth continuation.

C. The advantage in recognition of recoverable versus nonrecoverable images

Biederman (1987) showed that line drawings of common objects that had undergone *nonrecoverable contour deletion* – the deletion of contour that prevented recovery of the parts – were unrecognizable, whereas the same, or even a greater amount of contour deletion, would allow recognition as long as the parts were *recoverable*. This study also eliminates a possible role of direct matching of spatial filter components, as Fiser, Biederman, and Cooper (1997) showed that recoverable and nonrecoverable images were equally similar to the original intact images (see Fig. 8 below). As two or three parts can be deciphered in the recoverable image, the recoverable images enjoy an advantage in identifiability, even when they are partially occluded so that they only have

approximately 50% of the contour of the nonrecoverable images (Biederman, 1987).

D. The priming and matching invariance of depth-rotated stimuli that differ in GSDs

The invariance in shape priming or matching over depth-rotated stimuli that differ in GSDs, described previously, is evidence against the role of global shape (Biederman & Bar, in press; 1998; Biederman & Gerhardstein, 1993).

E. Independent processing of parts

Recently, Koen Lamberts and his associates (Lamberts, 1998; Lamberts & Freeman, in press; Freeman & Lamberts, 1998) have presented strong evidence for the parallel independent sampling of an object's parts in the earliest stages of categorization. Subjects had to classify a particular image of a four-part lamp (top, shade, stem, and base), say, into one of two four-lamp categories, defined by different shapes for each of the parts, such as a square or round base. The parts could vary in salience, and a given shape for a part was more characteristic of one of the categories than the other; for example, three of the four lamps in category A had square bases. Using a deadline procedure in which subjects had to emit an instantaneous response to an unpredictable signal, these investigators showed that a lamp with a high salience part biased toward category A and with lower salience parts that unequivocally indicated that the lamp actually belonged to category B would be responded to as an A under very short deadlines but as a B under longer deadlines. More importantly, the quantitative data were fit by a model that assumed parallel and independent processing of the object's parts, as posited by geon theory (Biederman, 1987; Hummel & Biederman, 1992).

Evidence for the importance of simple, nonaccidental differences in object parts

The studies just reviewed in the previous section documented the role of parts in object recognition. We now review evidence suggesting that the representation of the parts can be modeled as geons in that NAPs are important and recognition depends less on metric and irregular contour variations.

A. Greater salience of NAPs over metric properties in sequential name matching

Subjects in Cooper and Biederman's (1993) experiment viewed sequentially presented pairs of line drawings of simple objects composed of two parts. Each image was shown for 100 ms, followed by a 100-ms mask. Subjects judged whether the two images had the same or different name. On half the same trials, the aspect ratio of one of the parts (e.g., the cylindrical base of a lamp) would vary. In the other half of the same trials, the geon for that base would change from a cylinder to a brick, but the aspect ratio would remain the same, as illustrated in Fig. 1. Compared to the "standard" object (Fig. 1), the differences in aspect ratio were scaled to be moderately greater that the differences in geons, as assessed by the Lades et al. (1993) system (described later) and also as assessed by a simultaneous same-different matching for physical identity. Despite the greater scaled similarity of the geon changes, matching two images differing in a

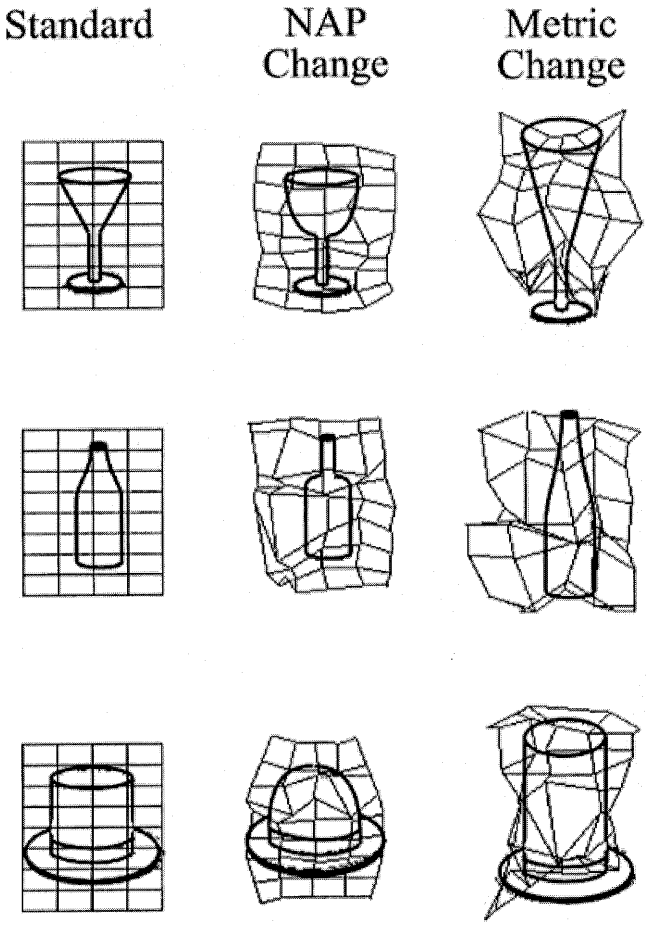

Fig. 1 Sample object stimuli from Cooper and Biederman (1993). Given the standard object on the left, a NAP of only a single part was changed in the objects in the middle column (NAP condition), and that same part was lengthened in the metric condition illustrated by the objects in the third column. Whereas the difference between metric and standard images were more readily detected when performing a simultaneous physical identity matching task (Are the objects identical?), in a sequential object matching task (Do the objects have the same name?), a change in a NAP resulted in far more disruption than a change in a metric property. The magnitude of the metric changes were slightly larger than the NAP changes, according to the Lades et al. (1993) model. The distortion of a regular lattice calculated by that model is proportional to the dissimilarity between two images

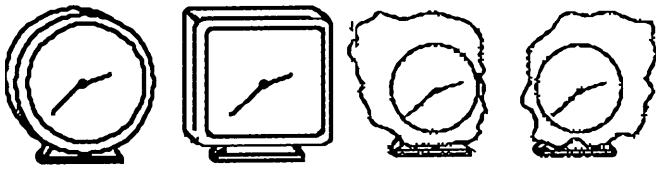

Fig. 2 Sample stimuli from Cooper et al. (1995). The two regular clocks on the left are slightly more similar in shape according to the Lades et al. (1993) model than the two irregular clocks on the right. Nonetheless, basic-level matching was more disrupted by a change from one regular exemplar to the other than between one irregular exemplar and the other

geon, such as a cylinder-base lamp and a brick-base lamp as "lamps," resulted in longer RTs and higher error rates than matching two images with a part differing only in aspect ratio. Longer RTs and higher error rates would be expected if the two images to be judged "same" were perceptually more dissimilar.

B. Insensitivity to irregular shape variations

In a design similar to that of Cooper and Biederman's (1993), Cooper, Subramaniam, and Biederman (1995) showed that sequential name matching of simple objects was unaffected by changes in the irregularities of a part (Fig. 2), but that changing a nonaccidental characteristic of a regular part resulted in elevated RTs and error rates. For example, instead of the base of a lamp being a cylinder or a brick, it was a highly irregular free-form mass of approximately the same aspect ratio as the regular parts. Subjects judged whether two object images, each presented for 100 ms and each followed by a 500-ms mask, had the same or different name. On same trials the images could differ in a geon, from brick to cylinder, for example, or in the shape of an irregular part. (The magnitude of the geon and irregular part changes were scaled according to the Lades et al., 1993, model of spatial filter similarity.) Although there was no effect of a change in the shape of an irregular part, that the part *is* irregular is coded nonetheless in that a change from a regular to an irregular part or vice versa resulted in large decrements in matching performance.

C. Depth invariance conferred by differences in GSDs but not metric properties

Biederman and Bar (in press; Fig. 3) replicated Biederman and Gerhardstein's (1993) result that depth-rotated stimuli could be recognized without cost when the distractors differed in a geon. They further showed that in the identical paradigm where the detectability of geon and metric differences were equated at 0° differences in orientation, a depth rotation of about 60° drove detection of metric differences to below chance accuracy. Subjects in this experiment had only one trial ("one shot") with a given stimulus pair and could not predict if the objects would differ and, if so, what part would differ and in what

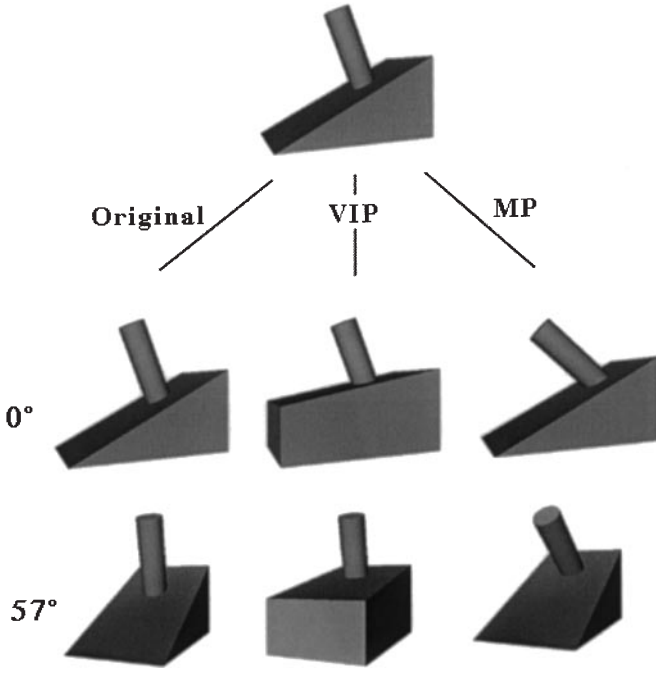

Fig. 3 Sample two-geon stimuli and trial types in the Biederman and Bar (in press) experiment. On half the trials the stimuli were same, and an half they were different. When different, the stimuli could differ in a NAP of a part or a metric property of that part. The stimuli were calibrated so that metric and geon differences were equally detectable at 0° rotation. Rotation in depth (average 57°) resulted in no cost in detecting the geon differences but produced performance for detecting the metric differences to below chance levels of responding

way. The advantage of the geon-differing pairs in this task refutes Tarr and Bülthoff's (1995) claim that nonaccidental properties only confer an advantage when observers know in advance where and what they are.

These studies document the importance of a NAP characterization of even a single object part in mediating subordinate level recognition. The same experiments document a lack of an effect of variations in aspect ratio, irregular contours, and metric information, in general, even when equated to geon differences according to an early filter similarity space according to the Lades et al. (1993) model. We also note that GSDs provide a representation format that can be directly mapped onto the linguistic units that people employ in describing objects and reasoning about them (Tversky & Hemenway, 1984; Hummel & Holyoak, 1997).

D. Neural tuning

The moderately complex feature tuning of cells in the inferior temporal (IT) cortex of the macaque, reported by K. Tanaka (1993), to a considerable degree can be characterized as preserving NAPs. Similarly, Logothetis, Pauls, & Poggio (1995) allowed that a small set of distinctive features might have been what was mediating the responses of the IT cells that were responding to a

particular view of an object. (See, also, the previous discussion of the studies of Logothetis et al., 1994; 1995). Given that an object can be recognized with little or no cost under dramatic changes in viewpoint variables of position, size, and orientation, while at the same time the subject retains explicit episodic memory for the specifications of the view, such as the object's original position, size, and orientation (e.g., Biederman & Cooper, 1991a, 1992; Cooper, Biederman, & Hummel, 1992), neural evidence for both types of representations should be found.

Implications for subordinate-level classification

In the Cooper and Biederman (1993) and Cooper, Subramaniam, and Biederman (1995) experiments, the critical differences in the stimuli concerned a single part. Tversky and Hemenway (1984) and Murphy and Lassaline (1997) noted that a difference in a single part is often sufficient for subordinate level distinctions. Consequently, the information emphasized by distinctive GSDs – nonaccidental differences and simple parts – likely plays an important role in the recognition performance for such stimuli, just as they do in basic-level naming.

Three subordinate level object cases

The degree to which subordinate-level entities differ in GSDs and the scale of these distinctive GSDs (if present) vary on continua. However, there are three important qualitative landmarks on these continua that may determine the type of processing that is employed to distinguish among subordinate objects.

We suggest here a taxonomy of subordinate-level object classification tasks based on the perceptual requirements for achieving a given classification. (The possible special case of face individuation will be discussed later.) We can describe the perceptual requirements for a particular subordinate-level classification of an object in terms of (a) the kind of information that is available for distinguishing among the various subordinates, NAPs versus metric properties, (b) the scale (size) of the contours or surfaces that are to be distinguished, and (c) the magnitude of the to-be-discriminated differences, that is, the similarity of the values that have to be distinguished.

We will argue that people are predisposed to employ NAPs rather than metric properties, even when the information distinguishing the NAPs is at a small scale. Only when there are large differences in metric properties will such information be employed. In such cases, the nominally large metric difference will often produce a NAP difference, either in a geon property or in the relations among geons. For example, consider an object part that is very small in one case and very large in another. The difference might be readily distinguished as smaller-than versus larger-than some other part rather than the absolute size of the part.

Although a given subordinate-level classification can fall anywhere on the three attributes, the bias toward using NAPs and the ease of using such information, even when the differences are just modest, suggest three prototypical cases. Two of these subordinate subclasses, those that employ NAPs at a large or small scale, Cases 1 and 2, respectively, account for the vast majority of subordinate-level classifications that can be made quickly and accurately in everyday and technical lives. To a large extent, NAPs are the visual information required to achieve basic-level classifications; thus, as discussed earlier, there is little to perceptually distinguish basic-level and Case 1 subordinates, except that the former have unique names. Case 3 requires discriminating similar values of metric information and such classifications are performed rarely, primarily in technically constrained tasks.

Case 1. Large part or relation differences or very large differences in metric values

Consider, first, the differences between such subordinate pairs as a grand piano and an upright piano, a pole lamp and a ginger jar table lamp, and a round table on a central pedestal and a square table with four legs. In these examples, the GSDs differ in both geons and relations, and consequently, discriminating among these instances will tend to be fast, accurate, viewpoint-invariant, and readily labeled. Even if only a single large geon were to distinguish the instances, as with a square and a round table, both with four legs, then easy and invariant subordinate discrimination would be evidenced (Biederman & Gerhardstein, 1993).

Large differences in metric properties can also lead to a different viewpoint-invariant characterization of the object as occurs when the relative size of the parts and their qualitative aspect ratios (i.e., whether the axis is longer, approximately equal to, or shorter than the cross section of the geon) vary. In these cases a different object classification can be invoked, as with a nail and a tack. For the tack, the axis of the shaft is approximately equal to the diameter of the head; it is longer for the nail. This variation would produce different GSDs (Hummel & Biederman, 1992) and, consequently, relatively easy subordinate discriminations.

Very large metric differences that do not change relations can also be readily discriminated, such as those of a yardstick and a 12-inch ruler or a truck and a toy truck. However, when metric and viewpoint-invariant differences are scaled according to same-different judgments of physical identity of simultaneously presented stimuli, it is clear that differences in NAPs are far more salient than differences in metric properties, and only the NAP-differing stimuli reveal immediate viewpoint invariance (Biederman & Bar, in press; 1998).

Case 2. Differences in GSDs at a small scale

A second basis for distinguishing among subordinate classes arises with those cases in which a small viewpoint invariant difference is employed to distinguish among otherwise highly similar and complex entities, such as when we employ the logo or name to distinguish a Mazda from a Ford or attempt to distinguish a bull from a cow. (Letters have evolved and logos are designed to differ in NAPs.) Scientists studying a pod of sperm whales use the nonaccidental aspects of the pattern of tears and nicks in the trailing edge of the tail to distinguish the individuals in the pod (Nature TV program, 1996). In these cases, a basic-level classification is first performed – that the object is a car, a bovine, or a whale – and then a search is undertaken for the distinguishing geon differences, such as the logo, udders, or nicks, to determine the subordinate classification. Sanocki (1993) makes a similar point in a priming study in which the brief presentation (33–67 ms) of the external outline of a house or vehicle immediately prior to a briefly presented stimulus is shown to facilitate the identification of the subordinate type, despite the external outline being common to all the subordinates. The critical subordinate level information – knowing where to look for a small viewpoint-invariant difference – is thus based on an initial basic-level classification of the object, a house or vehicle in Sanocki's task. Depending on the difficulty of locating the critical information and determining its shape, such tasks could be more difficult to perform than Case 1 subordinates. Subordinate discriminations of this type promise the greatest gains from minimal training: All one has to do is to learn where to look for a NAP difference (Biederman & Shiffrar, 1987).

Case 3. Small metric differences

The third case is that in which the critical information for distinguishing among subordinate entities is fundamentally metric. Such information would primarily include small differences in aspect ratio or curvature of a region, such as distinguishing between a 1/4-inch drill bit and a 3/8-inch drill bit. These types of subordinate discriminations are the most difficult of the three cases, particularly when the objects are free to rotate in depth. It is not clear that humans have much facility for performing such classifications or ever spontaneously do so (Biederman and Bar, in press; Biederman, 1995; Biederman & Gerhardstein, 1993, 1995; Miller, 1956).[4]

[4]This discussion as to the employment of small metric differences is limited to object classification. There is extensive evidence that different kinds of representations mediate object classification and motor interactions (Milner & Goodale, 1995; Biederman & Cooper, 1992). Thus, the dorsal pathway, extending from V1 to the posterior parietal cortex, which is crucial for visual control of motor interaction with objects, is well-tuned to metric properties. The ventral pathway, extending from V1 to V2 to V4 to IT, is crucial object classification and is sensitive to GSDs.

There is often considerable mysticism concerning expert perceptual discriminations (Biederman & Shiffrar, 1987). Tanaka and Taylor (1991) reported that bird (but not dog) experts were able to more quickly reject false instances at a subordinate level (oriole?) than at a basic level (bird?). However, the distractors (i.e., false instances) were selected to "maximize the visual contrast between target pictures" (p. 472). That is, the distractor for the oriole would be a duck, not a tanager! Because the picture pairings were repeated, it is also possible that subjects were able to exploit not only the class differences in shape, such as between an oriole and a duck, but also any fortuitous differences in pose and setting.

A number of investigators have attempted to study Case 3 subordinate recognition with stimuli resembling bent paper clips (e.g., Bülthoff & Edelman, 1992), differing only in the angle between the segments. The learning that is evident when people study such complex, metrically varying sets of stimuli at particular poses, however, may well be based on qualitative configurations. For example, a normally viewpoint-invariant characteristic, such as the approximate cotermination of the endpoints or approximate parallelism of two segments, might be present only for a narrow range of orientations (Biederman, & Gerhardstein, 1993; Biederman & Bar, 1998). Subjects may learn to associate a set of these configurations with a given paper clip to distinguish it from other bent paper clips, as discussed previously. This type of subordinate classification might thus be properly regarded as an instance of Case 2 rather than Case 3.

For metrically varying stimuli which do not afford qualitative features, a similarity space for stimuli may be determined, a priori, from a representation based solely on the pattern of activation of a lattice of multiscale, multiorientation Gabor-type filters (e.g., Lades et al., 1993; Biederman & Kalocsai, 1997). Thus, RTs and error rates for discriminating between a pair of complex, random appearing blobby shapes (created by Shepard & Cermak, 1973) or a pair of highly similar faces is strongly and negatively correlated with such a similarity measure (Biederman & Subramaniam, 1997; Biederman & Kalocsai, 1997). As stimuli are selected that are less and less similar, so they can be distinguished by differences in part structures or viewpoint-invariant properties – when they become instances of Case 2 rather than Case 3 discriminations; the spatial filter similarity space is no longer relevant (Fiser et al., 1997; Biederman & Bar, 1998; Biederman & Subramaniam, 1997).

It is easy to imagine sets of objects, all from the same basic level class, in which the information specified by Case 1 is sufficient to distinguish among some of the subordinates, that specified by Case 2 to distinguish among others, and the fine metric differences specified by Case 3 required to distinguish still others. An easy-to-difficult hierarchy would likely be manifested in the subordinate classification of complex objects, whereby more difficult stimulus discriminations would only be engaged to the extent that they were required.

The ecological frequency of the three classes of subordinate level discriminations

Biederman and Gerhardstein (1993) asserted that Case 3 subordinate classifications (small metric differences) were made only rarely by people in their everyday lives. (This claim is made with respect to the representations mediating the *classification* of objects, rather than with the representations mediating the *motor interactions* with those objects.) Tarr and Bülthoff (1995) have argued that the absence of precise statistics of the frequency of performing the three cases of subordinate discriminations prevents one from assessing the importance of GSDs relative to metric variations. Although exact frequency counts are unavailable for the different kinds of classification, it is not too difficult to determine orders of magnitude estimates for the frequencies of the different classes.

Entry-level classifications, which govern to a large extent how we understand our visual world and how we select objects for interaction, are almost always conveyed by large differences in GSDs, documenting the high frequency at which objects are classified through discriminations of the first kind. What would be an upper-bound estimate of the possible rate of such identifications? Accurate identification of object pictures displayed by rapid serial visual presentation (RSVP) techniques can be revealed with presentation rates of 10 pictures per second (Potter, 1976). Barring fatigue and boredom, this rate would allow a capacity of 576,000 objects per 16-hour day. Only a third of this value would be achieved if we were limited to the normal scanning eye fixation rate of 3 fixations/second, although this reduction in the upper bound of the possible rate of object identification would be more than balanced by the high frequency at which we generally appreciate not a single object but many objects interacting to compose a real-world scene.

It is, of course, difficult to estimate actual numbers, as many classifications are implicitly made without an overt response, such as when we select a chair in a room so that it is near a table and a lamp, but with a view of the door and window. Although we may not explicitly name or motorically interact with the object, Smith and McGee (1980) showed that the classification of a picture of an object is fast, obligatory, and automatic. A large number of the objects in a scene are thus identified.[5] Probably our highest everyday rates of object recognition are achieved when quickly changing television channels (''channel surfing'') and our lowest rates achieved when we perform some repetitive activity in a restricted environment, such as reading or playing racquetball. Informal, subjective observations by the first author (IB) suggest that basic level classifications – often of several objects and their interactions – are performed at least once per saccade. If we take two objects per fixation and one fixation per second to be lower-bound estimates, then we are performing approximately 57,600 classifications of the first type in a 16-hour waking day. Mental comparison of complex objects differing metrically is made only a few times per day.

The discussion of the frequency of these classifications of the first type has, until this point, not distinguished entry-level from subordinate level classifications. According to geon theory (Biederman, 1987; Biederman, & Gerhardstein, 1993; Hummel & Biederman, 1992), *all* object representations are minimally specified at the level of Case 1 insofar as the representation specifies the parts and relations comprising a geon structural description. An alteration in a geon or relation will activate a different GSD, although not necessarily a different entry-level classification, as with the aforementioned square and round table. Whether the representation specified the small scale or metric information required for Case 2 and Case 3 classifications undoubtedly will vary with the necessity to use that information.

Subordinate-level classifications of the second type, as when we determine the make of a car from its logo, are made far less frequently than those of the first type. IB's diary account suggests that such discriminations may be made at the rate of several per hour, although the rates can certainly increase to several per minute in certain activities, such as when reshelving books in a library or classifying collections of butterflies.

It is only rare that we make subordinate-level classifications of the third type, those based on fine metric differences. Although there are an abundance of metric differences among sedans, which we can appreciate by superimposing one image on top of another, we almost never employ such information in determining the manufacturer. Instead, we ignore these metric variations and seek out the name or logo of the car. If we try to distinguish identical models of chairs in our office, we look for a distinguishing stain or scratch. *Birds of North America* (Robbins, Bruun, Zim, & Singer, 1983), commonly referred to as ''The Golden Guide,'' presents nonaccidental descriptors to distinguish highly similar birds, such as ''a curved plume'' versus a ''straight plume'' for distinguishing among western U.S. male quail.

The extreme reluctance to employ metric variations for representing differences among highly similar objects almost amounts to an aversion. Biederman and Gerhardstein (1995) recount how a figure was mistakenly described as three different orientations of the same bent paper clip. In a large number of presentations, preprints, and a publication, the error – it was actually three different paper clips rather than one – was never detected. The aversion to making metric judgments is likely based on a realistic appraisal of one's own capacity. A human can easily judge which one of two side-by-side lines is the longer, but object identification is an absolute judgment

[5]Although identification of many objects in a well-formed scene from a single glance is achieved (Biederman, 1987; Biederman, Mezzanotte, & Rabinowitz, 1982), there is very poor memory for those objects or their attributes unless they are explicitly attended (Rensink, O'Regan, & Clark, 1996).

task, requiring memory and categorization, and the speed and accuracy of such judgments for metric variations is severely limited (Miller, 1956). In summarizing a large body of research, Miller noted that the capacity for errorless classification of a unidimensional metric property was 7 ± 2. Often, such judgments are made slowly and deliberately, with much more time required for their execution than that for object classification. At an uncertain orientation in depth, the accuracy of metric judgments declines precipitously (Biederman & Bar, 1998). Despite a lifetime of exposure to rulers, few of us can accurately – much less, quickly – judge, within a centimeter, the length of the stapler on our desks. Similarly, we have all seen particular angles, formed by junctions of pipes, paper clips, etc., rotate in depth. Yet it does not seem possible for us to generalize that experience when performing in a bent paper clip experiment.

That people exploit small viewpoint-invariant (rather than metric) differences when such information is available in making classifications among highly similar entities was demonstrated by Biederman and Shiffrar (1987) in a task in which subjects had to determine the sex of day-old chicks based on pictures of their genitalia. This task, which was reputed to be the most difficult visual learning task known – presumably requiring years to master – could be learned with less than a minute's instruction as to where to look to find a structure (the "eminence") and determine a simple viewpoint-invariant difference as to whether it was convex (male) versus concave or flat (female).[6]

Similarly, Biederman and Shiffrar (1987) noted that a proposed training program (Kotas & Bessemer, 1980) for teaching military personnel how to distinguish NATO from Russian tank, emphasized distinctive features that were, in fact, viewpoint-invariant part differences. Figure 4 shows four NATO and three Russian tanks redrawn from Kotas and Bessemer's report. A simple, viewpoint-invariant rule allows easy classification of these images into the two categories: If the rear of the turret is completely curved (i.e., not straight or notched), then it is a Russian tank.

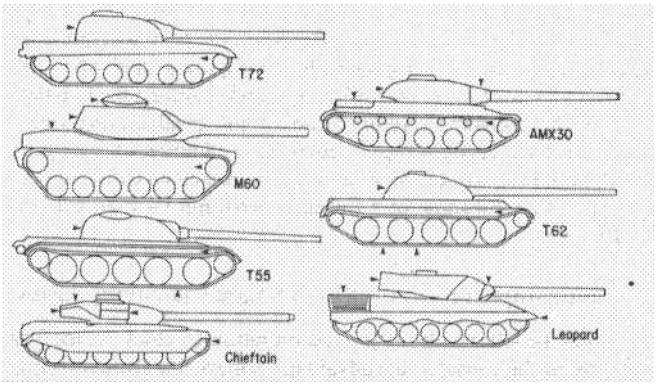

Fig. 4 (From Biederman & Shiffrar, 1987, redrawn from Kotas & Bessemer, 1980). NATO and former Soviet Union (FSU) tanks. The rear of the turrets of the FSU tanks, designated by the T in their names, are all completely rounded. (From Kotas and Bessemer, 1980. Adopted with permission)

The bird typing, chicken-sexing, tank-classification, and auto identification tasks can thus be solved by a similar strategy. First, a determination is made of the basic-level class of the image, for example, that it is chick genitalia, a tank, or a car. Often, of course, this information is known at the outset of the task. Then, a smaller region is isolated, such as the eminence, rear of turret, or logo (or name), and a classification based on viewpoint-invariant differences is made of that region. These are all instances of Case 2 subordinate level discriminations. If viewpoint-invariant differences are not present (or known), then an organism attempting a classification would have to resort to discriminating metric properties (Case 3).

Often, the ability to make absolute judgments is regarded as a special talent or learned skill, such as "perfect pitch" or the ability to make fine discriminations in an industrial inspection task, rather than as a regular characteristic of everyday recognition. In any event, the representation that is at the core of many of the "view-based" models – a template specifying precise metric values – would seem to be an inappropriate model to account for human object recognition performance.

Representations mediating subordinate-level classifications: Metric templates or geon structural descriptions?

As suggested previously, recent theorizing on shape representation has coalesced around two theoretical perspectives: metric templates and invariant parts. We here consider these theoretical positions in somewhat more detail. Both classes of theories assume that an image of an object is initially represented in terms of the activation of a spatially arrayed set of multiscale, multioriented detectors, such as the arrangement of simple cells in V1 hypercolumns.

[6]Axel Larsen (Personal communication, 1998) has raised an alternative possibility that, rather than searching for nonaccidental differences, the great difficulty in *discovering* the basis of the discrimination for chick sexing is that there is "a natural inclination to use metric alignment as a basis for recognition." Our own view is the opposite: that people generally search for a nonaccidental feature, even (especially) at a small scale. Metrical alignment when done mentally is difficult, and there is some evidence that people do not spontaneously do it, insofar as a large number failed to detect that an oft-shown image labeled as different views of the same bent paper clip was actually three different clips (Biederman & Gerhardstein, 1995). The presence or absence of "accidental" qualitative features for depth-rotated stimuli dominate performance in matching metrically varying objects such as bent paper clips (Biederman & Bar, 1998). What may appear as metric alignment could be the matching of corresponding parts. We do not know why chicken sexing took so long to discover, but undoubtedly the lack of quick feedback. We do not know why chicken sexing took so long to discover but undoubtedly the lack of quick feedback (unless the chick is sacrificed on the spot) contributed to the difficulty. Secondary sexual characteristics do not appear until the chick is one month old.

Metric templates

Metric templates (e.g., Edelman, 1995; Lades et al., 1993; Poggio & Edelman, 1990) map these values (a) directly onto units in an object layer, with each unit representing a different stimulus (Lades et al., 1993; Edelman, 1995) or (b) onto hidden units which, over experience with the stimuli, can be trained to differentially activate or inhibit object units in the next layer (Poggio & Edelman, 1990). Metric templates preserve the 2D retinotopic spatial positions and metrics of the inputs, without explicit specification of edges, viewpoint-invariant properties, parts or relations among parts. The distribution of the pattern of activation over the detectors becomes the representation which can then be compared to new patterns (e.g., Lades et al., 1993; Poggio & Edelman, 1990; Edelman, 1995).

A biologically inspired, highly successful face recognition system developed by Christoph von der Malsburg and his associates (Lades et al., 1993; Wiscott, Fellous, Krüger, & von der Malsburg, 1997; Biederman & Kalocsai, 1997) suggests a theoretical perspective from which many of the phenomena associated with the present discussion of template models might be understood. We focus on this system because the representation is motivated by the early spatial representations of human vision and, as will be discussed, the model's determination of shape similarity correlates extraordinarily well with human psychophysical similarity of complex shapes and faces where distinctive GSDs are not available. The representation does not make explicit the critical information in GSDs – parts, nonaccidental properties, and explicit relations – which are, presumably, determined later in the ventral pathway. The quantitative specification of early shape similarity provided by the model thus provides a basis with which to assess and evaluate the contribution of GSDs.

As diagrammed in Fig. 5, the fundamental representation element is a column of multiscale, multiorientation spatial (Gabor) kernels with local receptive fields centered on a particular point in the image. Each column of filters is termed a "Gabor jet" and each jet is presumed to model aspects of the wavelet-type of filtering performed by a V1 hypercolumn. As illustrated in Fig. 6, Lades et al. (1993) posited a two-layer network. The input layer is a rectangular lattice of Gabor jets. The pattern of activation of the 80 kernels (5 scales × 8 orientations × 2 phases, sine and cosine) in each of the jets is mapped onto a representation layer, identical to the input layer, that simply stores the pattern of activation over the kernels from a given image. An arbitrary large number of images can be stored in this way to form a gallery.

Matching a new image against those in the gallery is performed by allowing the jets (in either the probe or a gallery image) to independently diffuse (gradually change their positions) to determine their own best fit, as illustrated by the arrows on the jets in the input layer. This allows a matching of two images that may have moderately different orientations and expressions. The similarity of two images is taken to be the sum (or mean) correlation in corresponding jets of the magnitudes of activation values of the 80 corresponding kernels. The correlation (range 0 to 1) for each pair of jets is the cosine of the angular difference between the vectors of the kernels in an 80-dimensional space. (If the values are identical, the angular difference will be 0 deg and the cosine [= correlation] will be 1. The greater the differences in the vectors, the greater the angle, and the lower the cosine.) The correlations over the jets are summed to get a total similarity score. The degree of deformation

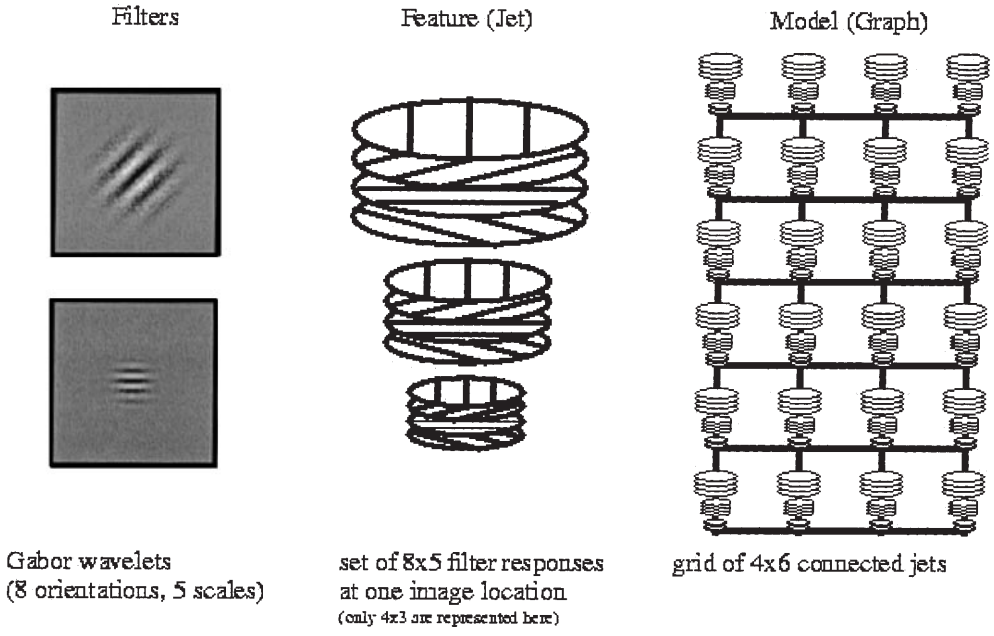

Fig. 5 Illustration of the input layer to the Lades et al. (1993) network. The basic kernels are Gabor filters at different scales and orientations, two of which are shown on the left. The center figure illustrates the composition of a jet, with the larger disks representing lower spatial frequencies. The number of jets, scales, and orientation can be varied. (From Biederman & Kalocsai, 1997)

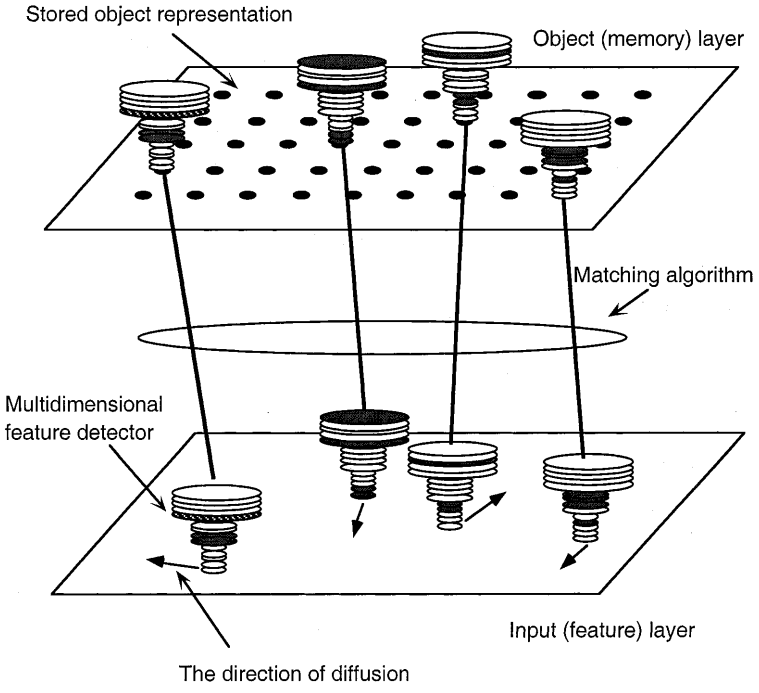

Fig. 6 Schematic representation of the Lades et al. (1993) two-layer spatial filter model. The model first convolves each input image with a set of Gabor kernels at five scales and eight orientations arranged in a 5×9 lattice. (Any or all of these parameters can be varied.) The set of kernels at each node in the lattice is termed a "Gabor jet." The activation values of the kernels in each jet along with their positions are stored for each of the images to form a "gallery." The drawing shows the diameters of the receptive fields to be much smaller than actual size in that the largest kernels had receptive fields that were almost as large as the whole face. (From Fiser et al., 1997)

of the lattice of original positions typically provides a visual measure of the similarity of two images.

Given a test image against a number of stored images, the most similar image is taken to be the recognition choice. Over modest changes in pose and expression and modest-sized galleries of faces (a few hundred), the Lades et al. (1993) model does a good job at recognizing faces, with recognition accuracy that can exceed 90%. Recent extensions of the model (Wiskott et al., 1997), described in the section on Face Recognition, in which each jet is centered on a facial landmark, such as the temporal corner of the right eye, can achieve 95% accuracy in galleries of several thousand faces with greater variations in input conditions (e.g., pose, expression, lighting). The Lades et al. model can be readily applied to the similarity scaling of objects as well as faces, so it has the potential to serve as a device for the scaling of both kinds of stimuli. Figure 7 (from Kalocsai, Biederman, & Cooper, 1994) shows three pairs of images scaled by the model. The images are of the same individual and the pairs differ in pose and expression. The greater the deformation of the lattice, the lower the scaled similarity. The lower the scaled similarity, the more difficult it was for human observers to judge that two images were of the same person.

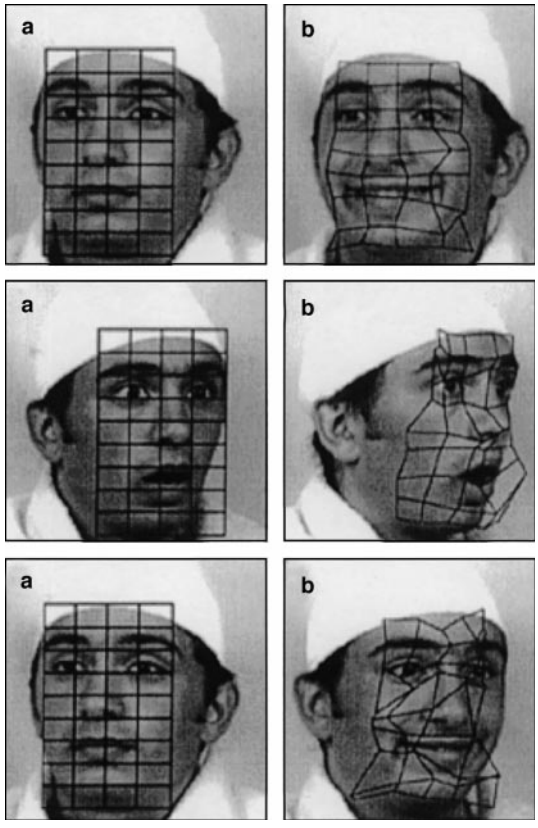

Fig. 7 Sample images from the Kalocsai et al. (1994) experiment with the Lades et al. (1993) lattice deformations superimposed over different pairs of images of the same person. The positioning of the lattice over an original image is shown in the left-hand column (*a*) and the deformed lattice is shown in the right-hand column (*b*). Top, middle, and bottom rows show changes in expression, orientation (60°), and both expression and orientation, respectfully. The similarities as determined by the Lades et al. (1993) model correlated highly with performance in matching a pair of images when there were at different orientations and expressions (Kalocsai et al., 1994)

The Lades et al. (1993) system as an object recognizer

As effective as the Lades et al. (1993) system is as a face recognizer, the model evidences shortcomings as an object recognizer in that it does not distinguish among the largest effects in object recognition. One such effect is the difference in the recognizability of contour-deleted stimuli where the geons can or cannot be recovered from the image, as shown in Fig. 8 (Kalocsai & Biederman, 1997). Whereas with sufficient exposure duration, recoverable stimuli can be recognized nearly perfectly, median accuracy of recognition of nonrecoverable stimuli is zero. Fiser et al. (1997) investigated whether the Lades et al. system would reveal this difference in recognizability. The intact versions of each of the 48 images from Biederman (1987) were used as the gallery. The similarity of the recoverable and nonrecoverable to the intact versions was then assessed with the Lades et al. system. Figure 8 shows an example of this matching. Overall, there was no difference in the similarity of recoverable and nonrecoverable images against the intact versions, indicating that the system was completely insensitive to the contour variations that produce the enormous psychophysical difference in recognizability.

The Lades et al. (1993) system is also insensitive to another aspect of shape representation: the greater salience of nonaccidental differences in shape compared to metric differences, as illustrated in Fig. 1 from the Cooper and Biederman (1993) experiment. Compared to the "original" object, the differences in aspect ratio were made to be moderately greater that the differences in geons, as assessed by the Lades et al. system. Despite the greater similarity of the geon changes, matching two images differing in a geon, such as a cylinder-base lamp and a brick-base lamp as "lamps," resulted in longer RTs and higher error rates than matching two images with a part differing only in aspect ratio.

What is the justification for using the Lades. et al. (1993) system as a scaling device for "early" representations of shape similarity between objects? The model captures essential aspects of the multiscale, multiorientation filtering within circumscribed receptive fields that is characteristic of the tuning of many of the cells in the earlier cortical stages in the ventral pathway (viz., V1, V2, and V4). In this sense, the model offers a scaling device for determining the spatial similarity of shapes as specified in the earlier stages. The Lades et al. system, however, is insensitive to the information that is specified by the representations mediating object recognition, such as edges, parts, and nonaccidental properties (Biederman, 1987). This information is presumed to be made explicit in stages subsequent to those that are performing spatial filtering, perhaps as specified by the complex feature cells described by K. Tanaka (1993). The stage analysis here is not strict: Kobatake and Tanaka (1994) showed that in areas V2 and V4 there are some complex feature cells coexisting with the spatially tuned cells which predominate in those areas.

Consistent with the preceding interpretation of the role of Gabor-jet similarity in human shape recognition were the results of a physical identity, simultaneous object matching experiment of Cooper and Biederman (1993), which contrasted with the results of the sequential, conceptual matching experiment in the investigation described previously. In this experiment, subjects viewed a display of two object pictures (the same object images as in the sequential name matching experiment described earlier) presented at diagonal quadrants of the display, so subjects could not use simple symmetry as a cue to same trials. The objects always had the same name, but the subjects had to judge whether they were *physically* identical or not. On half the trials the images were identical. On the other half, they differed by a geon (each with the same aspect ratio) or in the aspect ratios of the same geon. On the different trials, subjects were faster at detecting the aspect ratio differences than the geon differences, in line with the similarity scaling from the Lades et al. (1993) model which, as noted previously,

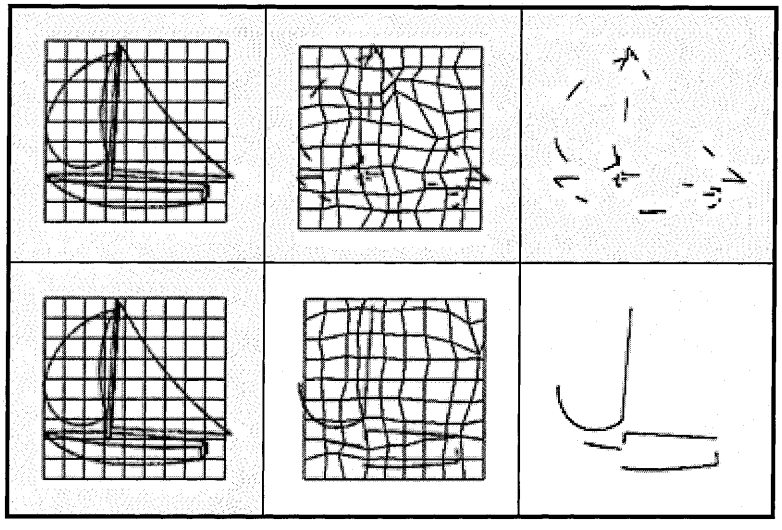

Fig. 8 Examples of the grid distortions in matching recoverable and nonrecoverable images to the original intact images performed by Fiser et al. (1997). (Figure from Kalocsai & Biederman, 1997)

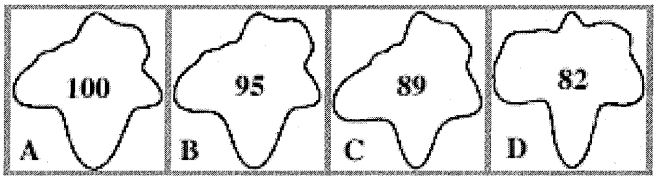

Fig. 9 Four examples of the free-form shapes from Shepard and Cermack (1973). The numbers superimposed over the shapes indicate the similarity values as calculated by the Lades et al. (1993) system of each of the shapes (*A*, *C*, and *D*) to shape *A*. Higher values indicate greater similarity

assessed the differences in aspect ratio to be greater than the differences in geons. This task could presumably be done with a representation that could be specified solely by a spatial filter representation because (a) the task required no memory, as the to-be-compared images were presented simultaneously, and (b) the task required no object classification, as the subjects only had to judge physical identity of objects from the same class. That the results of the simultaneous physical matching task were consistent with the Lades et al. similarity orderings – easier detection of aspect ratio differences than geon differences – provides some confirmation of the relevance of this model as a scaling device for early shape representations.

Spatial-filter similarity and stimuli not distinguished by GSDs

Even stronger evidence for the relevance of the Lades et al. (1993) model as a scaling instrument for early shape representations is provided by a recent study by Biederman & Subramaniam (1997) in which subjects performed physical same-different judgments on pairs of sequentially presented free-form, asymmetrical, blobby novel shapes devised by Shepard and Cermack (1973) and illustrated in Fig. 9. Shepard and Cermack generated 81 of these shapes by varying two parameters to produce a toroidal two-dimensional space (''toroidal'' because the space curves around onto itself). Biederman and Subramaniam reported that the correlation between Lades et al. similarity values, illustrated in Fig. 9, for pairs of shapes which had similarity values greater than 82 and RTs and error rates on different trials was .95 and .96, respectively. The similarity values were taken to be a percentage of the maximum similarity for a pair of identical shapes, which would be 100. Maximum similarity for a pair of different shapes was 95. For example, subjects would be slower to judge that A and B in Fig. 9, with a similarity of 95, were different stimuli compared to A and C, with a similarity of 89. That is, without any free parameters, the similarity values generated by the Lades et al. model provided an excellent measure of psychophysical similarity of complex shapes. The value of 82 turned out to be the point below which subjects in Shepard and Cermak's experiment started to reliably report qualitative differences between pairs of shapes. When Biederman and Subramaniam ran an experiment that included the full range of shapes (down to a minimum similarity of 68 between a pair of shapes), the results were clearly bilinear, with values below 82 showing only a weak correlation between wavelet similarity and RTs and error rates (which were near or at 0%). Inspection of the stimuli revealed that dissimilar shapes typically had nonaccidental differences, such as the left lobe being tapered to a rounded end in one shape, as in A in Fig. 9, and parallel to a squarish end in another, as in D.

Spatial filter similarity and stimuli that do differ in GSDs

This extensive treatment of a domain where the Lades et al. (1993) system does provide a good account of psychophysical similarity should not obscure the fact that such conditions will be rare in noncontrolled settings. As noted earlier, almost always some nonaccidental differences will exist in discriminating among highly similar entities, and it will be rare that an observer has to appeal to a wavelet type of representation.

Geon versus irregular part differences. The Cooper et al. (1995) object name-matching experiment described previously (Fig. 2) had a design similar to that of Cooper and Biederman's (1993), except that instead of the stimuli varying in aspect ratio, the stimuli could vary in the contours of an irregular part (that resembled, somewhat, the 1973 Shepard and Cermack shapes shown in Fig. 9). The magnitude of the difference in irregularities were scaled by the Lades et al. (1993) system to be slightly more dissimilar than the difference in the regular parts. Despite the greater wavelet dissimilarity of the irregular parts, a change in a geon resulted in far more disruption in judging that the two objects had the same name than did a change in the shapes of the irregular parts. In fact, there was no effect of a change in the shape of an irregular part. This result thus parallels that of Cooper and Biederman's in showing that when performing object classification, nonaccidental differences (viz., a difference in geons) are far more salient than differences in metric properties or variations in irregular shapes. This study also replicates the Cooper and Biederman result in showing that the Lades et al. Gabor-jet type of similarity measure does not account for psychophysical similarity in object recognition. Specifically, contour variation that is part of a highly irregular region is likely regarded as texture rather than an aspect of shape that is specified in a representation of an object. It is not that the texture itself is not noted. Biederman and Subramaniam (1997) also showed that there was considerable disruption in matching if the change was from a regular part shape to an irregular part shape or vice versa.

Can smooth continuation account for the advantage of recoverable over nonrecoverable images and the equivalence of complementary feature images? For the recoverable images in the recoverable-nonrecoverable studies (Biederman, 1987) and the complementary image studies (Biederman & Cooper, 1991a), contour was often deleted in the middle of a segment. It will be recalled that Fiser et al. (1997) showed that the Lades et al. (1993) model did not distinguish recoverable from nonrecoverable images but did distinguish original and complementary images. Could a simple routine for smooth continuation produce a large advantage for the recoverable images with respect to their similarity to the original intact images and render complementary images equivalent? Kalocsai and Biederman (1997) assessed the Lades et al. similarity of recoverable-nonrecoverable images to the original intact versions and the complementary image pairs to each other after an extension field was applied to the Gabor kernels. The extension field extends the direction of activation of each of the kernels so that a kernel that is centered in the gap created by midsegment deletion would receive activation from both sides. The magnitude of the extension activation was Gaussian tuned and fell away with differences in orientation.

Imposition of an extension field did increase the similarity of the recoverable images to the original intact versions but not necessarily enough to account for the extraordinary differences in identifiability. The advantage of recoverable over nonrecoverable is maintained even when half of the recoverable image is deleted. Under such conditions, the Lades et al. (1993) model would assess the nonrecoverable image as more similar to the original. For both narrow ($\pm 15°$) and wide ($\pm 45°$) extension fields, only about a third of the differences between complementary images could be accounted for by smooth continuation.

Neural net implementation of extracting GSDs from objects

Hummel and Biederman's (1992) neural net implementation of geon theory (Fig. 10) can provide a framework for understanding how an invariant structural description can be extracted from the image of an object. At the top layer of the model (layer 7), individual units represent an invariant perceptual description of the object in terms of a binding of one or more units in layer 6. (The invariance holds over part aspects, i.e., the same representation will be activated as long as the same parts can be readily discerned in the image.) Each layer 6 unit represents a *geon feature assembly* (GFA), which binds the output of units representing a single geon, the attributes of that geon (e.g., aspect ratio, 2-D orientation), and the pairwise relations of that geon to other geons, such as Above, Larger-than, End-to-end connected. The units in L7 compete to self-organize to a particular pattern of output from L6 in that connections from a given L6 pattern to a particular L7 cell are strengthened if the cell fires shortly after the presentation of the L6 pattern. The details of this self-organization are beyond the scope of this paper, but a competitive function within L7 is a "vigilance parameter" (Grossberg, 1986), which tends to strengthen the inhibition from the maximally active L7 cell to other activated L7 cells. This produces a "winner-take-all" effect in that the most strongly activated unit succeeds in coding a given pattern of L6 output. Thus, two non-identical images from two instances of the same basic-level class with highly similar GFAs would tend to activate the same L7 unit. Very different GFAs, even if the objects had the same basic level name, would tend to activate different L7 units. One way in which slightly dissimilar instances of a basic-level class, i.e., different subordinates, might become differentiated is to suppress the inhibition from the most strongly activated L7 cell. This would allow more weakly activated cells to strengthen their connections to different subordinate patterns of L6 cells.

In this network, primal access corresponds to the activation of an L7 cell, which will generally be driven by a particular GSD. Two different L7 cells could be associated with the same name, but this point should not obscure the fundamental assumption that primal access would be at the level of a distinctive GSD.

As noted earlier, GSDs assume the same input layer (viz., a lattice of multiscale, multioriented filters), as assumed by template theories, but activate intermediate representations that make explicit the part structure (e.g., geons and the relations among geons) based on edges marking orientation and depth discontinuities and a NAP specification of these edges (e.g., Hummel & Biederman, 1992). There are two major differences between the two classes of theories:

1. In the deformable templates models, the connection weights to the units in the hidden layer are learned during the course of the experiment for that particular set of stimuli, whereas in the invariant parts models, the routines by which part structures and viewpoint-invariant properties are determined are assumed to have been developed over the course of infancy or evolution and are thus largely invariant to the particular set of stimuli that are selected for the experiment. Attentional processes, however, may allow selectivity to particular stimulus attributes.

2. In the deformable templates models, a coordinate space for the location of the various features is preserved. The relative positions of the features are implicit in their positions in the coordinate space. Transformations are required to achieve translation or scale invariance, and an additional theory would be posited to allow verbal description of the object, for example, "the shade is above the base." In the invariant parts theories, a structural description is activated that has relation units that *explicitly* specify the relations, such as top-of, among the parts. Mapping to language requires no additional perceptual processing. As the structural

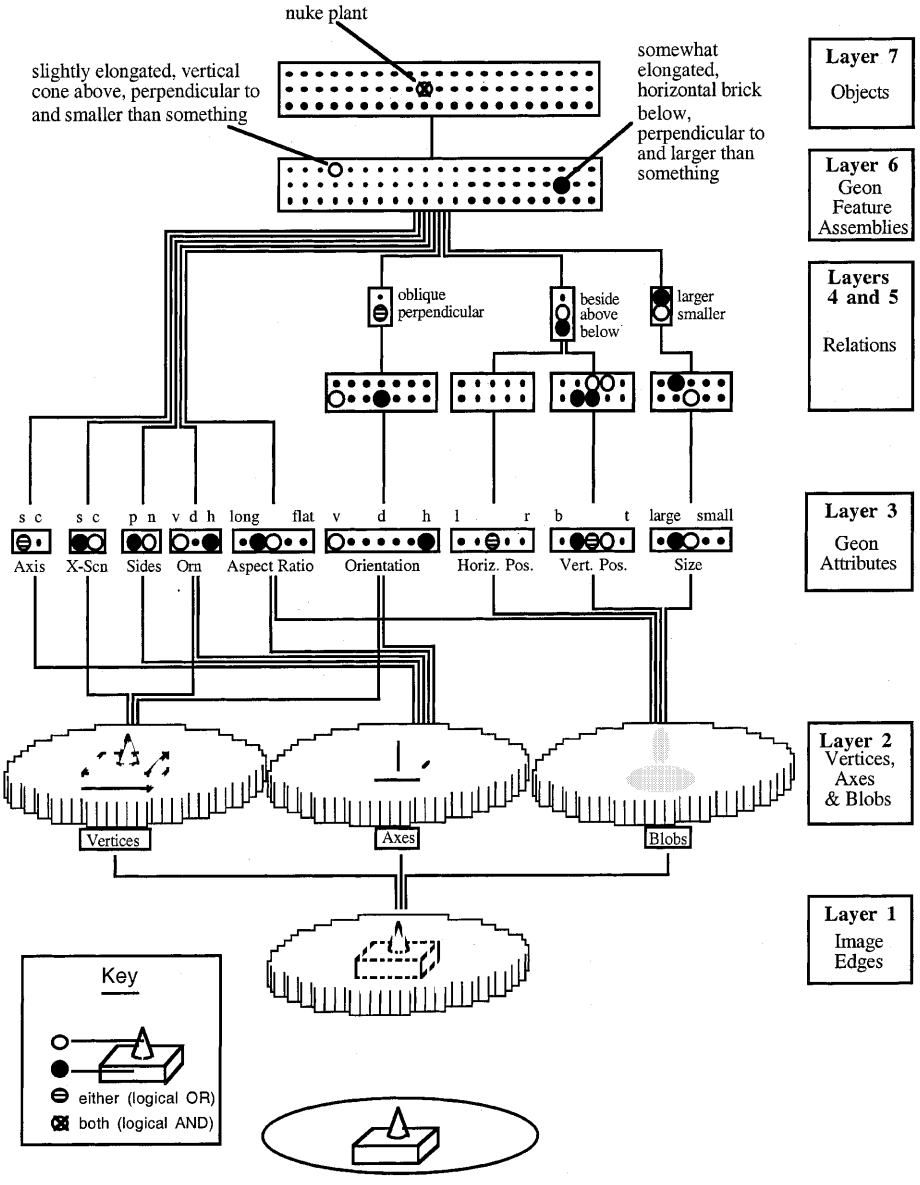

Fig. 10 The architecture of the Hummel and Biederman (1992) neural net implementation of geon theory indicating the representation activated at each layer by the image in the Key. In Layers 3 and above, *large circles* indicate cells activated in response to the image, and *dots* indicate inactive cells. Cells in Layer 1 represent the edges (specifying discontinuities in surface orientation and depth) in an object's image. Layer 2 represents the vertices, axes, and blobs defined by conjunctions of edges in Layer 1. Layer 3 represents the geons in an image in terms of their defining dimensions: axis shape: straight or curved; cross-section shape (*X-Scn*): straight (*s*) or curved (*c*); whether the Sides are parallel (*p*) or non-parallel (*n*); coarse orientation (*Orn.*): vertical (*v*), diagonal (*d*), or horizontal (*h*); aspect ratio: elongated (*long*) to flattened (*flat*); Fine orientation (*Orientation*): vertical (*v*), two different diagonals (*d*), and four different horizontals (*h*); horizontal position in the visual field (*Horiz. Pos.*): left (*l*) to right (*r*); vertical position in the visual field (*Vert. Pos.*): bottom (*b*) to top (*t*); and size: *small* (near 0% of the visual field) to *large* (near 100% of the visual field). Layers 4 and 5 represent the relative orientations, locations, and sizes of the geons in an image. Cells in Layer 6 respond to specific conjunctions of cells activated in Layers 3 and 5; cells in Layer 7 respond to complete objects, defined as conjunctions of cells in Layer 6. It is the activation of a Layer 7 cell that would qualify for primal access

description is invariant over changes in position or scale, no transformations are required to achieve such invariance. In addition, if the same surfaces are present in the image, the representation will be largely invariant to rotation in depth.

Tests of the subordinate-level hierarchy

The previous sections have suggested a hierarchy for distinguishing among subordinate-level entities in which large-scale GSDs (Case 1), if not sufficient, are employed to determine the locus of a small-scale GSD (Case 2) and, if necessary, finer metric information (Case 3). Would such a hierarchy reflect the similarity space when observers were attempting to identify stimuli in a large and varied set of objects? Specifically, would trained observers attempt difficult part discriminations only to the degree that easier discriminations (viz., of large

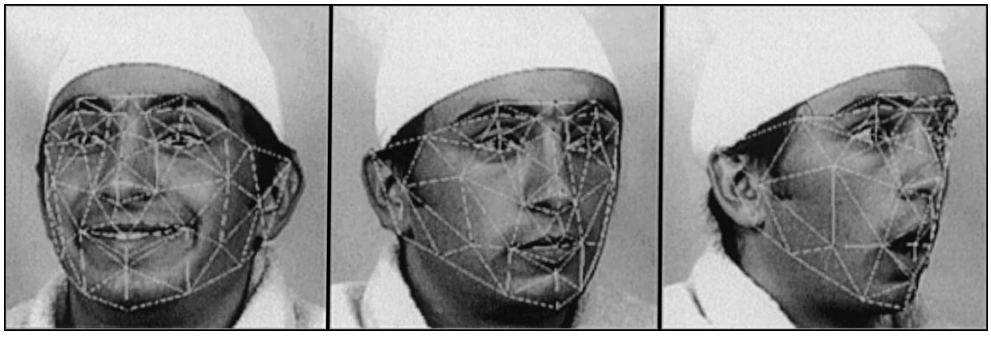

Fig. 11 Illustration of the mapping of jets onto fiducial points (*the vertices of the triangles*) on three images of the same person at different orientations and expressions. (From Biederman & Kalocsai, 1997)

viewpoint-invariant differences) would suggest that they would be diagnostic to a particular classification? Would a similarity tree that made such an assumption predict performance with a large and varied set of instances?

O'Kane, Biederman, Cooper, and Nystrom (1997) recently reported a test of such a subordinate-level hierarchy. In this investigation, trained observers attempted to identify military vehicles, 15 in one experiment, 9 in another, from infrared images. An inverted similarity tree (with the trunk on top) was constructed in which the top level distinguished groups of vehicles according to whether viewpoint-invariant differences between large parts, such as tracks versus wheels, were apparent. This top level thus constituted a Case 1 discrimination. Depending on the outcome of the first test, subsequent features, corresponding to Cases 2 and 3, could be selected for finer distinctions. At the bottom of the tree were metric distinctions, such as long versus short gun, which were only relevant for distinguishing among two of the vehicles.

The similarity tree was created subjectively, after viewing high-resolution side views of each of the vehicles at close range, without any knowledge of how the vehicles appeared in the actual experimental trials where they were shown at low resolutions, distant ranges, and uncertain orientations. The observers, trained military personnel who were familiar with viewing targets in infrared displays, were never instructed on the similarity tree. Their pretraining simply consisted of viewing the high-resolution, close-range photos. A measure of similarity between a pair of vehicles was defined as the nodal distance between the pair in the tree. The values ranged from 1 to 9 for the 15-vehicle experiment and 1 to 7 for the 9-vehicle experiment. The dependent variable was the confusion rate as a function of nodal distance. In both experiments, the rate of confusions between a pair of vehicles correlated .97 with a negative exponential of the nodal distance between vehicles!

Face recognition

Biederman and Kalocsai (1997) have recently argued that the pairwise similarity values for human faces, as determined by the Lades et al. (1993) system, are psychophysically valid. They reported an experiment in which subjects judged whether two briefly presented pictures of faces, sequentially presented, were of the same or a different person. The pairs of images on a given trial were always of the same sex and approximately the same age, with no striking distinguishing characteristics. The hairline – a potentially nonaccidental shape cue available on a coarse scale – was occluded. On both positive and negative trials, the expressions could be the same or different (either neutral or angry). The similarity values of a pair of faces were negatively correlated with different RTs and error rates and positively correlated with same RTs and error rates. Kalocsai et al. (1994) studied the effects of differences in orientation in depth and expression in judging whether two sequentially presented face pictures were of the same or a different person (Fig. 7). They found a strong negative correlation between the dissimilarity of the face images according to the Lades et al. model and the error rates and RTs in judging that the images were of the same person.

Wiscott et al. (1997) have recently extended the Lades et al. (1993) system so that each of the jets are centered on a particular facial landmark, termed a fiducial point, such as the left corner of the mouth, as illustrated in Fig. 11. The assignment of jets to landmarks is done automatically, based on a calibration sample of approximately 70 faces for which the jets were originally assigned to fiducial points by hand. The jet diffuses to a point that most closely matches one of the jets for that fiducial point. The matching of the two representations, the probe and an image in the gallery, is done in the manner described by Lades et al. This version of the face-recognition system shows great success, greater than 95%, in matching an image of a face to a person in a gallery of several thousand faces.[7] By the incorporation of different poses in the calibration set, as illustrated in Fig. 11, the system has a capacity to recognize faces over a wider range of poses than that shown by previous systems.

[7]The system won a recent national U.S. competition among face recognition systems (Phillips & Rauss, in press).

As noted previously, the Lades et al. (1993) system fails to reflect some major effects apparent in object recognition, such as the difference in recognizability between recoverable and nonrecoverable images, as noted by Fiser et al. (1997). Essentially, these shortcomings derive from characteristics of distinctive geon structural descriptions – edges, parts, nonaccidental properties, and relations among parts – that are not made explicit in the direct matching of spatial filter activation values assumed both by the Lades et al., and Wiscott et al. (1997) systems. Would evidence for face-like representations emerge if observers had to discriminate between highly similar objects? In particular, would a dependence on the original spatial filter values be evidenced? Two results suggest that basic-level classification of objects is not dependent on the direct matching of filter values. First, recall that the recoverable and nonrecoverable images were equally similar to the original, intact images with respect to those values (Fiser et al., 1997) yet there was an enormous difference in recognizability. Second, complementary pairs of contour-deleted line drawings of common objects, in which each member of a complementary pair had every other vertex and line from each part deleted (so the images of a pair, when superimposed, would make an intact original image), primed the other member of the pair as well as they primed themselves (Biederman & Cooper, 1991b), despite considerable differences in their similarity. (That is, the identical images had similarity values of 100%, but the complements had values that were considerably less than that [Fiser et al., 1997]).

Biederman and Kalocsai (1997) reported a direct test of this possibility that matching for objects as well as faces were performed with the filter values. They prepared complementary pairs of images in the Fourier domain of a set of gray-level images of highly similar chairs and faces. The complements were created by Fourier-filtering each original image into eight scales (spatial frequencies, SFs) and eight orientations. Each member of a complementary pair was assigned the contrast of every other scale and orientation. If the 8 scales × 8 orientations were laid out as a checkerboard, with the rows being the SFs and the columns being the orientations, one member of each pair would be assigned the red squares, and the other member the black squares. Subjects performed a sequential matching task in which they judged whether two sequentially presented images were of the same chair (in one experiment) or the same person (in another). The similarity of the images on Different trials and the similarity of members of a complementary pair (on Same trials) were approximately equal for the faces and chairs. Interest centered on the Same trials, whereby on half of them the images were identical and on the other half the images were complementary. For chairs, there was absolutely no difference between Identical and Complementary same trials, despite the difference in kernel activation values. For faces, there was a sizable increase in RTs and error rates when matching complements, as opposed to identical images. The results of this experiment thus confirm that the matching of faces is dependent on preservation of the spatial filter values, whereas there is sizable invariance over these values for objects.

Actually, object priming reveals even more striking invariance over specific filter values than was demonstrated in the Biederman and Kalocsai (1997) experiments. Whereas Biederman and Kalocsai used a checkerboard arrangement in which all scales and orientations were present in each member of a complementary pair, Fiser and Biederman (1995) created complements in which one member of a pair was low-passed and the other member high-passed, with an octave separation between the spatial frequencies. Nonetheless, there was invariance in the priming such that it made no difference whether the primed image was identical to the first or was of a different spatial frequency.

Conclusions and implications

Subordinate-level categorization should not be viewed as a single, homogeneous process but as a form of visual cognition that accommodates rich variation in the perceptual processing that is required. Nonetheless, this variation is more properly regarded as a variation in the scale rather than in the kind of information that must be distinguished. For the vast majority of subordinate-level classifications, the classification is based on distinctive GSDs. Except for faces, subtle metric differences rarely form the basis of a subordinate class.

A benefit of representing members of subordinate-level classes in terms of geon structural descriptions is that differences among the members can be readily communicated. The representation of an object in terms of its parts, their relations, and their viewpoint-invariant properties – aspects of an image represented by GSDs – appears to be not only fundamental for efficient viewpoint-invariant perception, but readily accessible to cognition and language.

That most important shape differences can be expressed by GSDs means that GSDs can provide a framework for training people on how to distinguish among highly similar objects. Indeed, every identification book on animals, birds, or leaves conveys the critical features as part of a hierarchical similarity tree. The training of subordinate instances by GSDs allows invariant recognition despite rotation in depth and other viewpoint variations. Different GSDs offer readily available perceptual boundaries, whether or not a culture has chosen to coin common linguistic expressions for these distinctions.

Acknowledgements This research was supported by grants ARO DAAHO4-94-G-0065, ARO DAAG55-97-1-0185; ONR N00014-95-1-1108; and NMA202-98-K-1089. We thank Lawrence W. Barselou and Gregory L. Murphy for insightful and helpful discussions.

References

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

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1987). Functional subdivisions of the temporal lobe neocortex. *Journal of Neuroscience*, *7*, 330–342.

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

Biederman, I. (1995). Visual object recognition. In S. F. Kosslyn and D. N. Osherson (Eds.), *An invitation to cognitive science*: Vol. 2. *Visual cognition*. (2nd ed., pp. 121–165). Cambridge: MIT Press.

Biederman, I., & Bar, M. (in press). One-shot viewpoint invariance in matching novel objects. *Vision Research*.

Biederman, I., & Bar, M. (1998). Same-different matching of depth-rotated objects. *Investigative Ophthalmology & Visual Science*, *39*, 1113.

Biederman, I., & Cooper, E. E. (1991a). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, *20*, 585–593.

Biederman, I., & Cooper, E. E. (1991b). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*, 393–419.

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 121–133.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for 3D viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 1162–1182.

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*, 1506–1514.

Biederman, I., & Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London B*, *352*, 1203–1219.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143-l77.

Biederman, I., & Shiffrar, M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual learning task. *Journal of Experimental Psychology: Human Learning, Memory, and Cognition*, *13*, 640–645.

Biederman, I., & Subramaniam, S. (1997). Predicting the shape similarity of objects without distinguishing viewpoint invariant properties (VIPs) or parts. *Investigative Ophthalmology & Visual Science*, *38*, 998.

Binford, T. (1971, December). Visual perception by computer. Invited paper at IEEE Systems Sciences and Cybernetics Conference, Miami, FL.

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

Cooper, E. E., & Biederman, I. (1993). Metric versus viewpoint-invariant shape differences in visual object recognition. *Investigative Ophthalmology & Visual Science*, *34*, 1080.

Cooper, E. E., Biederman, I., & Hummel, J. E. (1992). Metric invariance in object recognition: A review and further evidence. *Canadian Journal of Psychology*, *46*, 191–214.

Cooper, E. E., Subramaniam, S., & Biederman, I. (1995). Recognizing objects with an irregular part. *Investigative Ophthalmology & Visual Science*, *36*, 473.

Dickinson, S., Pentland, A., & Rosenfeld, A. (1992). 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 771–784.

Edelman, S. (1995). Representation of similarity in 3D object discrimination. *Neural Computation*, *7*, 407–422.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in: The recognition of familiar and novel views of 3D objects. *Vision Research*, *32*, 2385–4000.

Fiser, J., & Biederman, I. (1995). Priming with complementary gray-scale images in the spatial-frequency and orientation domains. *Investigative Ophthalmology & Visual Science*, *36*, 475.

Fiser, J., Biederman, I., & Cooper, E. E. (1997). To what extent can matching algorithms based on direct outputs of spatial filters account for human shape recognition? *Spatial Vision*, *10*, 237–271.

Freeman, R. P. J., & Lamberts, K. (1998). Feature salience and the time course of perceptual categorization. Manuscript submitted for publication.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Spatial Vision*, *37*, 1673–1682.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Vol. 1. Speech recognition*. San Diego: Academic Press.

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

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

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466.

Humphrey, G. K., & Kahn, S. C. (1992). Recognising novel views of 3-D objects. *Canadian Journal of Psychology*, *46*, 170–190.

Humphreys, G. W., Price, C. J., & Riddoch, M. J. (in press). From objects to names: A cognitive neuroscience approach. *Psychological Research*.

Jacobs, D. W. (1997). What makes viewpoint invariant properties perceptually salient? Unpublished manuscript, NEC Research Institute, Princeton, NJ.

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

Kalocsai, P., & Biederman, I. (1997). Biologically inspired recognition model with extension fields. *Proceedings of the 4th Joint Symposium on Neural Computation* (pp. 116–123). University of California, San Diego.

Kalocsai, P., Biederman, I., & Cooper, E. E. (1994). To what extent can the recognition of unfamiliar faces be accounted for by a representation of the direct output of simple cells. *Investigative Ophthalmology & Visual Science*, *35*, 1626.

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*, 856–867.

Kotas, B. L., & Bessemer, D. W. (1980). *Comparison of potential critical feature sets for simulator-based target identification training*. Final Report, U. S. Army Research Institute for the Behavioral and Social Sciences, Fort Knox Field Unit.

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

Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 695–711.

Lamberts, K., & Freeman, R. P. J. (in press). Building object representations from parts: Tests of a stochastic sampling model. *Journal of Experimental Psychology: Human Perception and Performance*.

Logothetis, N. K., Pauls, J., Bülthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*, 401–414.

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

Lowe, D. (1984). *Perceptual organization and visual recognition*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. Oxford: Oxford.

Moscovitch, M., Winocur, G., Behrmann, M. (1997) What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*, 555–604.

Murphy, G., & Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 70–84.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. In K. Lambert and D. Shanks (Eds.), *Knowledge, concepts, and categories*. (pp. 93–131). Cambridge, MA.: MIT Press.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, *21*, 1–20.

Nature (1996, October 13). Sperm whales: The real Moby Dick. Thirteen/WNET and BBC-TV, *Nature*. Los Angeles: KCET.

O'Kane, B., Biederman, I., Cooper, E. E., & Nystrom, B. (1997). An account of object identification confusions. *Journal of Experimental Psychology: Applied*, *13*, 21–41.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance, IX*. Hillsdale, NJ: Erlbaum.

Phillips, P. J., & Rauss, P. (in press). The face recognition technology (FERET) program. *Proceedings of the Office of National Drug Control Policy, CTAC International Technology Symposium*.

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

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 509–522.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1996). To see or not to see: The need for attention to perceive changes in scenes. *Investigative Ophthalmology & Visual Science*, *37*, 213.

Robbins, C. S., Bruun, B., Zim, H. S., & Singer, A. (1983). *Birds of North America*. New York: Golden Press.

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

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

Sanocki, T. (1993). Time course of object identification: Evidence for a global-to-local contingency. *Journal of Experimental Psychology; Human Perception and Performance*, *19*, 878–898.

Scalaidhe, P. O., Wilson, A. W., & Goldman-Rakic, P. S. (1997). Areal segregation of face-processing neurons in prefrontal cortex. *Science, 278*, 1135–1138.

Shepard, R. N., & Cermak, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-from stimuli. *Cognitive Psychology*, *4*, 351–377.

Smith, M. C., & McGee, L. E. (1980). Tracing the time course of picture-word processing. *Journal of Experimental Psychology: General, 109*, 373–392.

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

Stankiewicz, B. J., Hummel, J. E., & Cooper, E. E. (1998). The role of attention in priming for left-right reflections of object images: Evidence for a dual representation of object shape. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 732–744.

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

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*, 457–482.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science, 262*, 685–688.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study of the role of view-point dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin & Review*, *2*, 55–82.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1494–1505.

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

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, *1*, 275–277.

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

Tversky, B., & Hemenway, K. (1991). Parts and the basic level in natural categories and artificial stimuli: Comments on Murphy (1991). *Memory & Cognition*, *19*, 439–442.

Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT.

Vetter, T., & Poggio, T. (1994). Symmetric 3D objects are an easy case for 2D object recognition. *Spatial Vision*, *8*, 443–453.

Wiscott, L., Fellous, J.-M., Krüger, N., & von der Malsburg, C. (1997). Face recognition by elastic graph matching. *IEEE Pattern Recognition and Machine Intelligence*, *19*, 775–779.