
Introduction

Shape is the major route by which we gain knowledge about our visual world. All contemporary theories of shape-based object representation (e.g., Hummel & Biederman, 1992; Kobatake & Tanaka, 1994; Serre, Wolf, & Poggio, 2005) assume a hierarchy of features by which the initial Gabor-like filtering that is characteristic of V1 (the first cortical visual stage) cell-tuning is ultimately transformed through a series of stages to a point where cell-tuning is better described by “moderately complex” features with receptive fields (r.f.s) that often cannot be analyzed into their linear components (Kobatake & Tanaka, 1994; Tanaka, 1993). These later stages are the inferior temporal cortex in macaque (IT) and, in humans as determined by fMRI and lesions, likely the lateral occipital complex (LOC). Along with the increase in r.f. nonlinearity in IT and LOC, the cells exhibit a high degree of invariance so the response is only moderately changed to variations in viewing conditions. In this chapter we will review recent evidence, both behavioral and neural, that sheds light on the nature of the object representation that results in a view-invariant representation and leads to its unique phenomenology and psychophysics.

In addition to invariance, a striking aspect of the representation of shape is that, subjectively, at least, it appears to be structured. But a major class of theories of object representation (e.g., Riesenhuber & Poggio, 2002; Serre, Oliva, & Poggio, 2007) denies any role for such structure. Such accounts have been likened to a “bag of features” in which perceptual organization and a distinction between parts and relations are eschewed. The alternative (e.g., Hummel & Biederman, 1992; Hummel & Stankiewicz, 1996; Humphreys & Riddoch, 1987; Winston, 1975) is a structural description (SD) in which units for parts and relations are distinguished. This chapter will focus on the issues of invariance and representation. Indeed, they cannot be considered separately as it is the representation that may or may not allow invariance.
Two Theoretical Issues: Invariance and the Nature of the Representation

Invariance

Given that all contemporary theories posit a hierarchy of features—rather than, say, the direct matching of pixels or Gabor kernels—what are the issues that remain? One, of course, is how invariance can be achieved. An object seen at one position, orientation in depth, direction of illumination and size, can be often recognized with little or no cost when seen at another viewpoint where these variables can undergo considerable changes in their values. The Riesenhuber and Poggio (2002) model achieves a degree of position and size invariance, as does the Hummel and Biederman (1992) network. I will concentrate on invariance to rotation in depth and direction of illumination in this chapter as in these cases the spatial (i.e., Fourier or Gabor) composition of the image, characteristic of its initial (i.e., V1) cortical representation, is altered, often considerably.

Invariance to viewpoint, classically referred to as “shape constancy,” is, perhaps, the most striking phenomenon about the perception of shape. Despite the dramatic changes in the image of the surfaces of an object as it rotates in depth, the perception of the object’s shape is unchanged. This achievement is automatic and immediate, requiring no apparent cognitive effort, a process that has been referred to as “discounting the image.” Somewhat surprising, what does engage cognitive effort is the recovery of the shape of the projected image, as evidenced by the formal training in drawing classes that is required for most of us to achieve this skill. Left to his or her own devices, a child draws the rim of a cup viewed at a slant as a circle (its true shape), despite the image being elliptical. How does this achievement come about?

A bit of recent (personal) history and commentary on that history with respect to invariance to rotation in depth. Theories of object recognition are often termed controversial, particularly in accounting for the effects of rotation in depth. The apparent controversy has been phrased in terms of whether vision is “view-based,” requiring familiarity with a particular view for recognition to be achieved at that viewpoint without cost (e.g., Poggio & Edelman, 1990), or invariant. But what is the controversy? As we have noted previously (Biederman & Bar, 2000), all accounts of vision have to be view based. The alternative is extra sensory perception (ESP)! Moreover, “view based” is an effect—an assertion that there is a cost of rotation in depth. It cannot be posed as an alternative to recognition-by-components (RBC, or geon theory), which is a theory of representation.

The empirical issue has been defined in terms of whether there is a zero vs. some cost in matching or recognition of an object when it is presented at an orientation in depth other than the orientation of its initial presentation. Again, all accounts of vision would have to say that under most circumstances there should be some cost. In an extreme case, one cannot know what the back of a house looks like from looking at its front, aside from generalization from the viewing of similar houses. But what about the quite common intermediate case where some of the perceptual information
from one view can be discerned from another? In my 1987 *Psychological Review* paper, I proposed a “geon recovery principle” by which a similarity function could be computed, akin to Tversky’s (1977) aspects of similarity measure, that was a positive function of the number of geons that were present in both views and a negative function of the geons that were present in view A but not in view B and the number of geons that were present in view B but not in view A. There would have to be weightings of these geons in terms of their perceptibility (resolution) due to foreshortening and self-occlusion as well as their diagnosticity to the object/response class, and so forth. In an extreme case, a single pixel from a geon that would receive one classification if it had a straight axis and another classification if it had a curved axis would be insufficient to allow errorless classification. I also speculated that the same function might define the similarity between any two objects or an intact object and a version missing some parts, and so forth. There is considerable evidence that the overlap in parts does predict, at least qualitatively, rotation costs (e.g., Biederman & Gerhardstein, 1993), but a full quantitative account is still lacking. It is unfortunate, in my opinion, that so much ink has been spilled on attacking a position—zero rotation costs for all conditions of rotation— that no one ever held. Essentially what I had been advocating was a view-based theory with respect to the parts or surfaces that were in the image.

*Nonaccidental vs. metric properties.* Geon theory is a particular instantiation of SDs (i.e., geon structural descriptions, GSDs) (Biederman, 1987; Hummel & Biederman, 1992). GSDs place heavy reliance on nonaccidental properties (NAPs). NAPs are qualitative properties of (in the case of shape) orientation and depth discontinuities, which are largely unaffected by rotation in depth. For example, whether a contour is straight or curved is unlikely to change as the object rotates in depth. In contrast to NAPs, much image variation is metric (MPs, for metric properties), such as degree of curvature or the length of a contour. Whereas small differences in MPs are registered with difficulty, differences in NAPs provide a ready basis for distinguishing one object’s parts and relations from another (e.g., Biederman & Bar, 1999) and likely serve as the critical information for achieving invariance (e.g., Biederman, 1987). Neither a classification of contour by NAPs nor explicit parts nor explicit relations are specified by view-based templates or current bag of features accounts.

NAPs offer a solution to the inverse optics problem: that there are an infinite number of possible real-world 3D shapes or volumes that could produce any one 2D image. However, we almost always come up with the correct interpretation. Consider the nonaccidental property of collinearity (straightness). If a straight contour is present in the image, the perception of the 3D entity is that of a straight contour even though there are an infinity of points that could have projected that contour that are not collinear. So we are biased to interpret a collinear image as a collinear edge in the real world. It would be an accident of viewpoint if non-collinear points happened to project a straight contour. Essentially the visual system discounts the possibility of accidents when we have an image that could be collinear, smoothly curved, parallel, or symmetrical, which is why these properties are termed “nonaccidental.” The same holds true when there is a vertex in the image formed by the cotermination of two or three contours. These biases produce the simplicity of the
interpretation of shape that is the hallmark of Pragnanz. Given uncertainty of the slant of a surface, if an image could be that of a parallelogram we see it as a parallelogram even when we know it to be a trapezoid, as with the trapezoidal window (King, Meyer, Tangney, & Biederman, 1976). This is the solution to the trapezoidal window illusion. So the bias toward simplicity is that we do not assume, say, an asymmetrical shape when a symmetrical shape can explain the image (King et al., 1976). These nonaccidental biases lead to an interpretation of simple shapes (insofar as parallel, symmetrical shapes are simple) that constitute the geons.

**Invariance to surface features: The role of orientation and depth discontinuities**

Another distinction between geon theory and HMAX is that object recognition is achieved in geon theory through a determination of the shape of an object as defined by its orientation and depth discontinuities. HMAX just takes the image as is—surface characteristics such as color and texture as well as the object’s shape. Thus geon theory would tend to minimize the differences between a photograph of an object and its line-drawing rendition. For HMAX, this would be an enormous difference.

**Nature of the representation: Feature hierarchies vs. structural descriptions**

As noted previously, all theories posit a hierarchy of features by which Gabor-like units with local r.f.s are ultimately mapped onto highly nonlinear units with a high degree of invariance. An excellent example of nonlinear units can be found in the V4 L-vertex cells discovered by Pasupathy and Connor (1999). An L vertex, as its name implies, is formed by the cotermination of two (non-collinear contours) at a common point. The V4 L-vertex units are each tuned to a particular angle (e.g., 60°), at a particular position in the visual field and at a particular orientation as defined by the bisector of the angle of the vertex. These units are nonlinear in that they are activated neither by a contour matching the bisector of the angle nor by a single leg of the vertex. Both legs are required.

But is a set (or bag) of features consisting of vertices and lines sufficient, by itself, for understanding object recognition with a specification of the relations among the features? Some accounts (e.g., Riesenhuber & Poggio, 2002) would answer this question in the affirmative. Indeed, their HMAX model does an impressive job in assigning new object instances into previously learned object categories.

An alternative theoretical approach also assumes a feature hierarchy but maps the features onto a structural description (Biederman, 1987; Humphreys & Riddoch, 1987; Winston, 1975). An SD distinguishes parts and relations. It thereby allows reasoning about objects so that not only can the model determine that two images represent different objects, but how they differ. For example, given two objects, one with a cylinder on a brick and the other with a wedge on an identical brick, we can readily perceive that it is the top parts of the two objects that differ in shape, even if the bricks were not aligned horizontally.

As noted above, a major distinction between feature hierarchies and SDs is that relations are explicitly defined and distinguished from parts in an SD, as in Hummel
and Biederman’s (1992) JIM model but not in a feature hierarchy. Instead the relations in a model such as HMAX are implicit in a 2D coordinate space. A set of features (which in an SD might correspond to a part) might have coordinates that, if read out explicitly (as they are in an SD), could show that these features were above, or larger than, or connected end-to-end with another set of features but those labels, for example, “above,” “connected end-to-end,” bound to particular parts do not exist in HMAX but would be, presumably, the result of processing at some later stage.

Empirical Research

NAPs (geons) vs. MPs

GSDs specify both parts and relations. I will here concentrate on the NAP characterization of the parts.

The benefit conferred by NAPs, documented by Biederman and Gerhardstein (1993) and confirmed by Biederman and Bar (1999), is quite dramatic and is among the largest effects in shape recognition. This benefit does not appear to depend on exposure to the regular, simple artifacts that are so prevalent in environments in the developed world. Recently, Biederman, Yue, and Davidoff (2009) showed that the Himba, a semi-nomadic tribe in northeastern Namibia with minimal exposure to developed-world artifacts, showed markedly better performance in a match-to-sample task in matching single geons when the distractor differed in an NAP than an MP. In fact, the NAP advantage for the Himba was identical to that shown by University of Southern California undergraduates, suggesting that the connectivity subserving the NAP advantage develops from robust statistics that would hold with virtually any natural environment.

IT-tuning also shows greater sensitivity to differences in NAPs compared to MPs. Kayaert, Biederman, and Vogels (2003) showed that the IT cells in the macaque modulated their firing (up or down) more to a change in an NAP compared to a change in an MP when the differences in NAPs and MPs were equated by a measure of pixel energy.

Matching depth-rotated objects

Distinctive NAPs can confer an enormous benefit in attempting to determine whether two bent paper clips are the same or different when they are viewed at different orientations-in-depth (Biederman & Gerhardstein, 1993). These investigators substituted a different geon for each center segment of a set of 10 line drawings of bent paper clips. The addition of the distinctive geon dramatically reduced rotation costs (to 5,000°/sec) from a level with error rates so high that reaction times (RTs) were, essentially, uninterpretable.

View-based template accounts, in assigning no special status to NAPs or parts, would require familiarity with the specific views of novel objects, with only a modest generalization gradient around a nearby view. Some (Tarr & Bülthoff, 1995)
protested this demonstration, arguing that NAPs were of value only with a small set of known stimuli where people could anticipate a distinguishing NAP. That people would spontaneously exploit distinguishing NAPs was, indeed, one of the points that Biederman and Gerhardstein (1993) and Biederman, Gerhardstein, and Bar (1995) wished to make, but is familiarity required to get immediate viewpoint invariance with novel objects?

Moshe Bar and I (Biederman & Bar, 1999) compared directly the rotation costs for detecting differences in either an MP or an NAP in a same–different sequential matching task. We used novel, rendered two-part objects, such as those shown in Figure 17.1, presented at either the same or different orientations-in-depth. On half the trials the objects were identical; on half the trials they differed in either an MP, for example, aspect ratio, of a single part or an NAP, for example, straight vs. curved axis (producing a different geon) of a single part. The contrast of the object on the left and the third object in the right-hand column of Figure 17.1 illustrates a NAP.

Figure 17.1 An illustration of four trials in a same–different matching task of two-geon novel objects (from Biederman, 2000). The object on the left is S1, the first stimulus for all four trials. The four objects in the right column are possible S2s. The top object differs in both geons; the second and third in one geon; and the bottom object is the same, but rotated in depth. Observers should have no trouble accurately performing same–different judgments. Nor should they have any difficulty in describing the objects and how they differ from each other. Only the third and fourth S2s would have been trials in the Biederman and Bar (1999) experiment.
difference (straight- vs. curved-axis cylinder). The MP variation would have been a
cylinder with a different length (aspect ratio) or angle of attachment to the wedge.
The subjects saw a given stimulus sequence only once, so they could not predict
whether a part would change, or, if there was a change, which part would change
and how it would be changed.

How much of an MP difference is equivalent to a given NAP difference? A solu-
tion to this apples-and-oranges problem is critical for any principled answer to this
question. In the Biederman and Bar (1999) experiment the MP and NAP differences
were selected to be equally discriminable, as assessed by RTs and error rates, when
the objects were at the same orientation in depth (0° orientation difference). The
MP image differences were somewhat larger than the NAP image differences when
the images were scaled according to a wavelet-like similarity measure termed a “Gabor
jet” (Lades et al., 1993). A similarity measure based on Malsburg’s Gabor jets predicts
the psychophysical similarity of metrically varying stimuli (faces and teeth-like blobs)
almost perfectly (Yue, Biederman, Mangini, von der Malsburg, & Amir, 2012). Rotation
angles that averaged 57° produced only a 2% increase in error rates in detecting
the NAP differences but a 44% increase in error rates (to a level that was below
chance) in detecting MP differences.

Rotation costs, though small, are often apparent even with distinguishing GSDs
(Biederman & Bar, 1999; Biederman & Gerhardtstein, 1993; Hayward & Tarr, 1997;
Tarr, Bülthoff, Zabinski, & Blanz, 1997; Tarr, Williams, Hayward, & Gauthier,
1998). What might be producing these costs? It is possible, as noted by Biederman
and Gerhardtstein (1993), that an orientation-specific representation underlies these
costs. This representation may be of one of two types, given current theorizing: (a)
an episodic representation that binds view information to an invariant representation
of shape, as detailed by Biederman and Cooper (1992, see below), that could be
employed on some percent of the trials to mediate performance (though not neces-
sarily object perception), or (b) that there are viewpoint-specific representations
directly mediating object perception. But before the latter alternative is accepted
merely on the basis of some costs with distinguishing GSDs present, other possible
bases for the costs must be ruled out.

Look again at Figure 17.1 and consider what one would have to do to make it dif-
ficult, under rotation, to determine that the third S2 object was different from the first
and the fourth S2 object was the same. One way would be to render the object in such
a way that it would be difficult to determine if the distinguishing geon was curved or
straight. Biederman and Bar (1999), in a critical review of those studies reporting high
rotation costs, noted that low resolution of the distinctive geon was a common char-
acteristic in such studies. Biederman and Bar (1998) showed that factors that increased
the discriminability of distinguishing geons in rendered images, such as avoiding near-
accidents or using increased exposure durations, greatly reduced rotation costs.

There is another, subtler, effect that could have contributed to the apparent costs
of rotation in the Tarr et al. (1997, 1998) and Hayward and Tarr (1997) same-
different matching studies. On rotated Same trials and all Different trials in a Same-
Different matching task, a difference signal might be produced by the change in
luminance of specific display positions. This signal may be related to Nowak and Bul-
lier’s (1997) finding that marked changes—a transient—in a stimulus produce a signal
that rapidly propagates through the ventral pathway all the way to frontal cortex. (Because of the intervening mask, the difference signal would be between S2 and an actively maintained representation of S1, as noted by Biederman and Bar, 1999.) No difference signal would be produced when S1 and S2 are the identical, unrotated, object in the same position. So the subject could readily use the absence of a difference signal to respond Same on unrotated (0°) trials, artifically lowering RTs on such trials with the consequence that the slope of the RT X Rotation Angle function would be increased. Biederman and Bar (1998) showed that the effect of this artifact in increasing rotation costs could be greatly reduced by merely shifting S2 with respect to S1 on all trials, so that the difference signal was always present. The translation, by producing a difference signal on all trials, served to raise the RTs and error rates for 0° trials relative to rotated trials. This had the effect of greatly reducing the apparent costs of rotation. Biederman and Bar’s (1999) experiment, which found near-invariance over rotation, also translated S2 with respect to S1 on all trials.

Observations about bent paper clips as experimental stimuli

Many of the studies documenting large rotation costs have employed stimuli resembling bent paper clips. The central motivation for devising such stimuli (and others of similar design) was that they would be unfamiliar, so that the learning of different poses could be studied. However, one must consider an obvious characteristic that accounts for much of the extraordinary difficulty in classifying members of sets of such stimuli: The members of such sets are not distinguished by GSDs.

The absence of distinguishing GSDs in the standard set of bent paper clips means that the critical information for everyday shape recognition is absent from these stimuli, so the relevance of such objects to normal recognition can be questioned. Some bent paper clip devotees have suggested that their stimuli are relevant for subordinate-level recognition, such as the difference between different kinds of tables. However, a review of the vast majority of subordinate-level classifications that people make in their lives suggests that it is extremely rare that distinguishing GSDs are not available. A square table can be distinguished from a round table without appeal to metric information and certainly without engaging in mental rotation. Biederman, Subramaniam, Bar, Kalocsai, and Fiser (1999) note that NAPs of small regions, rather than metric templates, are specified for discriminating among highly similar classes such as birds on the same page in the bird guides.

Spontaneous appeal to nonaccidental properties

Think of how you would discriminate two different chairs of the same manufacturer’s model. Without fail, visitors to my office look for a distinctive scratch or stain or other such nonaccidental differences, at a small scale. They never consider what would be readily expressed by a metric template—the configuration of the whole chair or, in selecting a small feature, those that might differ metrically (at a modest scale).

A recent study (Amir, Xu, & Biederman, 2011) assessed the spontaneous employment of nonaccidental properties in distinguishing highly similar subordinate exemplars. We showed bird-naïve subjects a series of gray-level displays, each with five
birds representing exemplars of five closely related subspecies found on the same page in the bird guide (Robbins, Bruun, & Zimm, 1983). The following problem was posed to them: You are on the phone with your friend who has these five birds in a cage but she doesn’t know the names (designated by number on the display). What do you tell her so that she can know the names? Virtually all the descriptors generated by the subjects (entered by keyboard) were nonaccidental, for example, “#2 has a dark patch on the belly and #4 has a curved head feather,” which could be readily seen from any pose that projected that surface. So successful were the descriptors in distinguishing the birds that the experimenters could achieve 100% accuracy in using them to designate the target bird. In fact, the descriptors were essentially the same descriptors as those published in the bird guide, so the naïve subjects were generating the guide used by experts!

The objects shown in Figure 17.1 meet the criteria of being unfamiliar, yet in retaining distinctive geons they allow one to study how such information might be used. Although a set of paper clips lack distinctive GSDs, their projections often provide an accidental or near-accidental characteristic that people try to interpret in terms of GSDs (Biederman & Bar, 1998; Biederman & Gerhardtstein, 1993, 1995). For example, the bottom S2 object in Figure 17.2 resembles an arrow that would

**Figure 17.2** Illustration of four trials in a same-different matching task (from Biederman, 2000) for bent paper clips of the kind that could have been run by Edelman and Bülthoff (1992). Only the bottom S2 is identical to S1 (but rotated in depth).
normally be produced by actual cotermination of segments but is here an accident of viewpoint. Biederman and Bar (1998) observed that when there were such qualitative differences in appearance—typically well captured by differences in GSDs—miss rates in a same–different matching task were extremely high, exceeding 80%. When S1 and S2 were actually different clips but with similar GSDs, as in the upper three S2s of Figure 17.2, the false alarm rates were extremely high, again exceeding 80%.

As rotation angles increase from 0 to 90°, there is an increasing chance of changes in the qualitative characterization of such stimuli. The oft-reported increase in matching costs with increasing rotation angles may be more a consequence of an increasing chance for a change in an accidental GSD then in the rotation of a template. Consistent with this interpretation are the low rotation costs for 180° rotations. Such rotations often approximate mirror reflections under which the GSDs are preserved.

When GSDs are insufficient

There is no doubt that aspects of early cortical representation are well described by a 2D array of local filters at a variety of scales and orientations. The view expressed here is that the outputs of such a representation are mapped onto nonaccidental classifiers—such as units distinguishing straight from curved lines or various vertices produced by cotermination of end-stopped activity. A vector representing the activity of these nonaccidental classifiers (which, in JIM, are bound through correlated firing), in turn, activate units akin to Hummel and Biederman’s (1992) geon feature assemblies, representing single or pairwise combinations of geons and their invariant nonaccidental relations, such as:

**VERTICAL_CYLINDER_ABOVE_PERPENDICULAR_SMALLER_THAN_X.**

The output of such geon feature assemblies could readily map onto language, as evidenced by the manner in which people describe the objects in Figure 17.1, as well as memory structures supporting object cognition.

What if the stimulus does not have distinguishing parts and nonaccidental properties, as with the set of smooth blobby shapes studied by Shepard and Cermak (1973)? In such a case the nonaccidental classifiers would not be differentially activated to distinguish the members of the stimulus set and the observer would have to rely on whatever metric information distinguished the stimuli, in which case the similarity space would be that established by the early local, multiscale, multioriented Gabor-like filters (Biederman & Subramaniam, 1997; Yue et al., 2012). It should also be the case that discrimination among such stimuli should be more difficult than if distinctive GSDs were available (at the same level of spatial filter similarity), show more rotation costs, are difficult to articulate, and are not the basis of natural concept distinctions.

Discrimination performance among a set of highly similar faces shows such characteristics (Biederman & Kalocsai, 1997), as well as similar pairs of the Shepard and Cermak (1973) shapes (Biederman & Subramaniam, 1997) and objects with
irregular parts (Cooper, Subramaniam, & Biederman, 1995). See Biederman (1995) for a review.

**Can View-Based Accounts Incorporate Geons as a Unique or Diagnostic Feature?**

Given my earlier point that “view-based accounts assign no special status to NAPs,” one can ask whether view-based theorists could regard geons as some kind of unique or diagnostic feature extracted from a 2D view. The answer is “Of course.” But there is a serious problem with an account that holds that a unique or diagnostic feature will be employed for recognition. How does the perceiver know what is unique or diagnostic the first time he or she views an object? Consider, again, an individual seeing the nonsense object on the left side of Figure 17.1. The coding of that object would, roughly, appear to be a vertical cylinder on top of a wedge. That is, the object is described in terms of its simple parts and the relations among these parts (Tversky & Hemenway, 1984). This type of representation, a GSD, may well be the default description that the visual system generates in the absence of explicit knowledge about the other to-be-discriminated objects. GSDs often convey the functionality—or affordances—of the object. Moreover, GSDs often readily map on to verbal descriptions and allow reasoning about objects. We can readily say how the four objects on the right side differ from the one on the left (or from each other).

I note that the GSDs readily map onto the phenomenology of object representation. When viewing the objects in Figure 17.1, we readily decompose the objects into two simple parts, one above the other. The decomposition is done by segmenting the objects at regions of matched concavities (according to the transversality regularity) (Biederman, 1987). These concavities can be described by the L-vertices. Cells tuned to such vertices were discovered in cortical area V4 by Pasupathy and Connor (1999) (described earlier). The consequence of such parsing is simple, symmetrical parts that tend to be “good figures” in the Pragnanz sense. The tendency toward Pragnanz is thus not of the whole object but of the object’s parts, as I described in 1987. Chairs, elephants, and airplanes can be bad figures in their entirety but be represented by good parts. The good parts render these bad objects readily recognizable under almost all viewing conditions.

The important question is not whether a representation is view based, but what that representation is (as, again, all representations are view based). The phenomena of (a) small rotation costs with distinctive GSDs when matching novel objects, (b) the sizable costs in recognizing new views of objects, such as a set of bent paper clips, that are not distinguished by GSDs (as discussed in the next section), and (c) the reduction in the costs in (b) from learning the new views have obscured the issue of representation, insofar as the nature of what was learned was often not considered. In allowing translational and scale feature invariance, HMAX resembles an earlier proposal by Bartlett Mel (1997). There is nothing in the Riesenhuber and Poggio model to suggest the enormous inferential leverage and invariance to rotation costs provided by distinctive NAPs or the difference in recognizability between recoverable
and nonrecoverable contour deletion. These models are, essentially, bags of features in that they do not posit explicit structures, such as parts and relations among parts, by which objects might be represented—and described. Instead, different arrangements of the parts merely produce new features. A potential serious shortcoming of such models is that it is not clear how well they would do with novel objects that are to be distinguished from unknown sets of other objects, such as with the task illustrated in Figure 17.1.

Recent Neural Evidence for GSDs

Parts in IT

It has long been known that macaque IT neurons are highly shape selective and that different neurons show different shape preferences. Tanaka (1993) demonstrated that these preferences could be elicited quite strongly to features of “moderate complexity,” typically composed of one or two parts. This level of complexity closely matches what would be expected from single geons, invariant shape features, and, most frequently, geon feature assemblies (Hummel & Biederman, 1992), in which two geons are bound in a specific relation.

What occurs in IT when a macaque views an object? Tanifuji and his associates (Tsunoda, Yamane, Nishizaki, & Tanifuji, 2001) have employed optical imaging to address this question. Viewing a complex object such as a fire extinguisher does not activate the whole region homogenously. Instead, several spots of activity are apparent. This group then recorded the activity of individual neurons within these spots, using Tanaka’s (1993) reduction technique in which parts of a complex stimulus are removed in an effort to determine the specific feature(s) driving the cell. For the most part, neurons within a spot tended to respond to a single part of the object, such as the hose or the barrel, without any reduction in activity compared to their response to the complete object, although, occasionally, the removal of parts of the object resulted in increased firing, suggesting inhibition of that neuron from the removed part. A neuron responding to the curved hose ceased to fire when the hose was straightened, consistent with a general finding that to a first approximation, Tanaka’s (1993) and Kobatake and Tanaka’s (1994), moderately complex features are viewpoint invariant. Consistent with this interpretation is Esteky and Tanaka’s (1998) results showing that metric variation, namely, changes in aspect ratio that would be produced by a rotation in depth, had only a minimal effect on IT cell activity.

NAPs vs. MPs in IT

Kayaert et al. (2003) tested macaque IT (area TE) neurons with the identical set of two-geon stimuli used by Biederman and Bar (1999) to determine if greater modulation in cell activity would be produced by a change in a geon compared to a change in an MP. They found that geon changes, despite their smaller image changes (as assessed by wavelet similarity measures), produced greater modulation (up or down) in cell activity. Moreover, when the original objects were rotated (i.e., those without
an MP or geon change), the modulation attributable to the rotation itself was highly correlated with the modulation produced by MP changes for that cell but completely independent of the modulation produced by the geon changes. Such a tuning pattern would be expected given the results of Biederman and Bar (1999) that only geon-changed stimuli were readily discriminable from the originals under rotation.

As noted earlier, Kayeart et al. (2003) more recently replicated the Vogels, Biederman, Bar, and Lorincz (2001) effect of greater modulation from NAP as compared to MP changes. They scaled the image changes by a pixel energy measure. MP image changes had to be approximately 50% larger than NAP changes to produce the same degree of modulation. Moreover, the amount of modulation produced by depth rotation was equivalent to the modulation produced by nonrotated MP changes when the two conditions were equated according to the magnitude of image change.

Recent neural results supporting a geon account of shape representation:

Simplicity of parts, coding of independent generalized cylinder dimensions, and sufficiency of orientation and depth discontinuities

Three assumptions of geon theory are that (a) the representation of parts tends to be simple, (b) that the geons are a partition of the set of generalized cylinders based upon nonaccidental differences in the generating function by which a cross-section is swept along an axis, and (c) that the information can be derived from orientation and depth discontinuities of the original image. Results from both recent single-unit studies and fMRI studies provide strong support for these assumptions. In the Kayeart, Biederman, and Vogels (2005) investigation, macaques passively viewed 2D regular (i.e., simple) and irregular shapes while neurons in area TE were recorded. A difference in a regular shape, say between a circle and a square, produced markedly more absolute modulation (i.e., change in firing, up or down) than a change in a highly irregular shape, where the two types of changes were matched with respect to pixel similarity. If the irregular shapes differed in an NAP (namely, with round vs. straight contours), then the cells modulated more, suggesting that the sensitivity to NAPs can be witnessed even with irregular shapes.

Kayeart, Biederman, Op de Beeck, and Vogels (2005) discovered that a population code of IT neurons represents independent dimensions of generalized cylinders. For example, a given cell might respond to a highly curved axis of shape independently of its taper, aspect ratio, or curvature of its sides. These cells were, to a great extent, tuned to one end of a dimension or the other, with very few cells preferring intermediate values. Thus a cell could respond predominantly to a highly curved axis while another cell would respond to a straighter axis, with the firing declining as the axis curvature was changed away from the maximally preferred value.

In the Kayeart et al. (2003) study demonstrating greater sensitivity of IT cells to NAP compared to MP changes, the preferences were unaffected by depicting the shapes as 3D volumes, 2D silhouettes, or line drawings. This suggests that the shape preferences are tuned to the orientation and depth discontinuities. Consistent with this result is a finding by Kourtzi and Kanwisher (2000) that adaptation of the fMRI BOLD signal when viewing a sequence of two object images is maintained when the image changes from a gray-level photograph to a line drawing of the same object.
A change in the object causes a release of the adaptation, that is, a larger BOLD signal. The lack of an effect of image variables—and a form of invariance—was demonstrated by Vogels and Biederman (2002) who showed that the preferences of IT cells to rendered 3D objects was largely maintained irrespective of changes in the direction of illumination, changes which produced large effects on the image itself.

The ease with which one can accept geon theory’s reliance on orientation and depth discontinuities based on neural, fMRI, behavioral, and phenomenological observation belies the challenge that this step poses to computer vision systems which still cannot extract a good line drawing from a gray-level image. That implementations may be on the horizon (or a distant horizon?) is demonstrated by Zerroug and Nevatia’s (1996) system that was able to extract geons from a single picture.

Familiarity

There have been a number of reports of TE cell preferences reflecting experimental manipulations of familiarity (e.g., Logothetis, Pauls, Bülthoff, & Poggio, 1994; Tanaka, 1996). There are two points to be made about such demonstrations. First, tens, if not hundreds, of thousands of trials are required to obtain such preferences (Logothetis, 1999). Second, as discussed previously, it is not unlikely that there are at least two functions of object recognition subserved by IT. One is to provide descriptions of objects, novel or familiar, such as what the reader experienced when first viewing S1 in Figure 17.1. Such a system is likely well developed by late infancy. The second function is to provide an episodic record of the perceptual experience with particular objects or scenes. It is possible that the cells found in the training experiments are those subserving this second episodic memory function. That there may be these two representations of objects was documented by Biederman and Cooper (1992) who showed that the priming of object naming was invariant with size changes but that such changes produced considerable interference on episodic old–new judgments of the shape of the object (in which size was to be ignored). Distractors in that experiment were objects with the same name but a different shape. Similar results were found for changes in position and reflection (Biederman & Cooper, 1991) and orientation (Cooper, Biederman, & Hummel, 1992). Although the first function probably supports the lion’s share of human object recognition, it would certainly be possible to employ the second problem to solve particular classification tasks. If I know that a chair is on the right and a table on the left, a flash of an object on the right could lead me to infer that it was a chair rather than a table. Such view information could be employed whenever there was difficulty in determining an object’s GSD.

Baker, Behrmann, and Olson (2002) trained macaques to classify four vertical batons, each with one shape one top and another shape on the bottom, into two classes. The assignment of batons to responses was such that the monkey could not use an individual shape to make a response (see Figure 17.3). For example, one of the two batons assigned to the left key could have a circle on top and a square on bottom while the other would have a triangle on top and a star on the bottom. The two right batons would be one with a circle on top and a star on the bottom while the other had a triangle on top and a square on the bottom. After about 25,000
Figure 17.3  Illustration of the kinds of (but not actual) stimuli from the Baker, Behrmann, and Olson (2002) experiment. On each trial the monkeys saw one of the four batons. The two left batons were assigned to the left key; the two right to the right key. Note that the individual shapes and their positions (top or bottom), by themselves, are insufficient to determine the correct response. The monkey must process the conjunction of the two shapes. After about 28,000 training trials, cells in IT were found that responded to the individual batons but three times as many were found that responded to the individual shapes.

trials, cells in IT were found that responded to an individual baton, but three times that many still responded to the individual shapes, irrespective of what else was assigned to it. These results indicate that specific combinations of features can be learned but that the dominant coding in IT seems to be the individual shape.

Structural descriptions

Despite the common assumption of SDs in cognition and their value in object reasoning as described previously, there has been, until recently, no direct evidence for them. Behrmann, Peterson, Moscovitch, and Suzuki (2006) describe a patient with a lesion of the left LOC (approximately) who, at first glance, is sensitive to the shape of parts but not at all to the relations among these parts. After learning four two-geon objects he was able to determine when a geon changed but was completely insensitive to a change in the relations. However, the patient presents simultagnosia so it is possible that he can process only a single geon at a time.

Lescroart and Biederman (2010) investigated the cortical sensitivity to different arrangements of parts in three-geon novel objects composed of one large geon to which two different smaller geons were attached in different relations. Variations in the relations between the parts were defined by the medial axis of the parts and for each of three medial axis structures there were three different compositions of geons. These nine objects were varied over six planar and depth orientations. Subjects viewed individual objects and classified each, in different runs, either as one of the three axes structures or one of the three geon sets. Whereas in cortical areas V1 and V2 a support vector machine classifier distinguished orientation better than axes structure, by V3 and LO, the lateral occipital cortex, the accuracy reversed with axes structure being more accurately classified. This held true even when subjects were discriminating geon sets. There was no question that subjects could discriminate the axes structures as performance was near ceiling. These results provide strong evidence for the neural correlate of relations. Given that LOC, the lateral occipital complex (composed of
LO along with the posterior fusiform cortex), is the first stage in the ventral pathway at which shape is distinguished from texture, this result suggests—in line with phenomenology—that the relations between object parts are perceived simultaneously with the shape of the parts.

Consistent with the multivoxel result is a study by Biederman and Hayworth (2004) who reported an fMRI study in which subjects viewed brief two-frame “flip movies” in which one part of a two-geon object cycled between two different shapes so that a cylinder on top of a brick could change into a pyramid and back again for several cycles. A 24-s block of trials consisted of three of such geon change movies with the particular shape change varying between movies. In another block the geon would retain its shape but vary its relations, such as the cylinder moving from vertically on top of a brick to horizontally to the side of the brick. The magnitudes of these image changes were equated with respect to pixel energies and, indeed, MT was equally affected by the different kinds of changes. For every subject for every voxel in LOC, greater activity was associated with a change in part shape compared to a change in the relations between parts. In fact the relations condition did not lead to greater activity than a control condition in which the object retained its shape but merely rotated in depth. However, a region of the intraparietal sulcus showed markedly greater activity to the relations condition than the part shape condition. An event-related design is currently being run to assess whether this pattern of results is maintained without the presence of motion.

This sensitivity to relations at the same locus where shape is determined holds true not only for the relations between parts but also for the relations between objects. Hayworth, Lescroart, and Biederman (2011) employed a fast fMRI adaptation design (similar to the Kourtzi and Kanwisher study described previously) to determine the locus of the neural correlate of relative position. In their design, subjects viewed a sequence of two frames, each with two unrelated objects, say an elephant above (and not in contact with) a bus in the first frame. In the second frame, the objects could be identical, or translated, left, right, up, or down. In the relation-change conditions, the bus would now be above the elephant, with the position change of the two objects equivalent to what it was when the pair were simply translated, maintaining their relative position. They found virtually no effect of (i.e., no release from adaptation) in LOC when the pair of objects was translated but a marked increase in the magnitude of the BOLD response if the relative positions changed. This result strongly implies that at the cortical locus where shape is distinguished from texture, relations between objects are explicitly coded.

**Conclusion**

The evidence suggests that GSDs provide a suitable representation from which the invariance of object perception can be achieved. By making explicit the distinction between parts and relations, this representation readily matches the phenomenology of object perception, which allows object analogy and reasoning. The representation also provides a solution as to the role of the Gestalt organizational principles, especially that of Prägnanz, in object recognition: Prägnanz is not a tendency of the whole
object, but of the object's parts. The representation of parts as simple shapes and volumes defined by qualitative properties of orientation and depth discontinuities confer their invariance. The neural correlates for this conception of object representation have received strong empirical support over the past two decades so we are now able to understand the neural basis of the phenomenology of object perception and conception.

Acknowledgment

This chapter represents an updating of Biederman (2007). This research was supported by NSF 0420794, 0426415, and 0531177.

References


