GEON THEORY AS AN ACCOUNT OF SHAPE RECOGNITION IN MIND, BRAIN, AND MACHINE

Irving Biederman
Department of Psychology, University of Southern California
Hedco Neuroscience Bldg., University Park, CA 90089-2850, USA

Eric E. Cooper
Department of Psychology, University of Minnesota
Elliott Hall, Minneapolis, MN 55455, USA

John E. Hummel
Department of Psychology, Franz Hall
University of California at Los Angeles, Los Angeles, CA 90024, USA

Jozsef Fiser
Department of Computer Science, University of Southern California
Hedco Neuroscience Bldg., University Park, CA 90089-2850, USA

Abstract

In a fraction of a second humans are able to comprehend novel images of objects and scenes. Indeed, the human represents the only existence proof that a general shape recognizer is even possible. Geon theory offers an account of this phenomenon characterized by four general assumptions: a) Objects are represented as an arrangement of simple convex or singly concave parts (geons), b) The geons can be distinguished by binary contrasts (differences) in viewpoint invariant properties, such as straight vs. curved, rather than metric properties such as degree of curvature, c) The relations among the geons are explicit, such as PERPENDICULAR-TO or TOP-OF, as part of a structural description, rather than implicit in a coordinate space, and d) A relatively small number of geons is sufficient. Recent research evaluating these assumptions is reviewed.

1 Introduction

Three striking and fundamental characteristics of human object recognition are its invariance with changes in viewpoint, its ability to operate on unfamiliar objects, its robustness in the face of occlusion or noise, and its speed, subjective

BMVC 1993 doi:10.5244/C.7.18

ease and automaticity. Geon theory (Biederman, 1987) offers an account of this extraordinary capacity. In this paper we review the current statement of the theory, its empirical status as an account of human object recognition, and some of the challenges that remain. First we present a brief discussion of the goals of this research.

2 What Should be Modeled in a Theory of Human Shape Recognition?

There is likely no single answer to this question in that humans can activate an apparently unbounded set of classes for any given object image and achieve this activation in a variety of ways. For example, if the goal is to distinguish among the contents of a bin of parts, model-based matching, in which the image is matched against an exact, metrically specified, object representation (the model) can be successful (e.g., Lowe, 1987; Ullman, 1989). The major problem is estimating the pose of the object. No concern need be paid to the extent to which the theory's performance resembles that of a human.

We have concentrated on modeling primal access (Biederman, 1987): The initial activation in a human brain of a basic-level representation of an image from an object exemplar, even a novel one, in the absence of any context that might reduce the set of possible objects. This commits us to take seriously the data (especially reaction times) obtained during real-time performance. By concentrating on basic- (actually entry-) level classification, we account for the kinds of classification by which humans gain most of their knowledge about their world, that something is a sofa or an elephant, for example. This basic level refers to the level of abstraction of visual concepts that maximizes between-category distinctiveness and within-category informativeness. Specifying the subordinate level class, for example that something is an African (vs. Indian) elephant or is a particular style of sofa, provides only a slight increase in informativeness at an enormous loss of distinctiveness,. That is, the differences among sofas are much smaller (and less significant) that the difference between a sofa and an elephant! Similarly, the superordinate level, that something might be an animal or an article of furniture, sacrifices informativeness with only a slight gain, if any, in distinctiveness. By modeling entry-level rather than basic level, we can treat an exemplar of a class that differs greatly in shape from others of that class. Thus penguin is considered a separate entry-level class from the class birds . Entry level terms for an object are the first to enter a child's vocabulary and are used at least ten times more frequently than other level terms to refer to the same entity (Biederman, 1987).

3 Geon Theory

Geon theory assumes that objects are represented as an arrangement of simple, viewpoint-invariant, volumetric primitives, termed geons, such as bricks, cylinders, wedges, cones, and squashes, and their curved axis counterparts, as illustrated in figure 1. The arrangement matters, as illustrated with the cup and the bucket. In the cup, the curved cylinder is connected in (two) END-TO-MIDDLE joins and is

SIDE-OF the cylinder. In the bucket, the curved cylinder, is ON-TOP-OF and is connected in (two) END-TO-END joins.

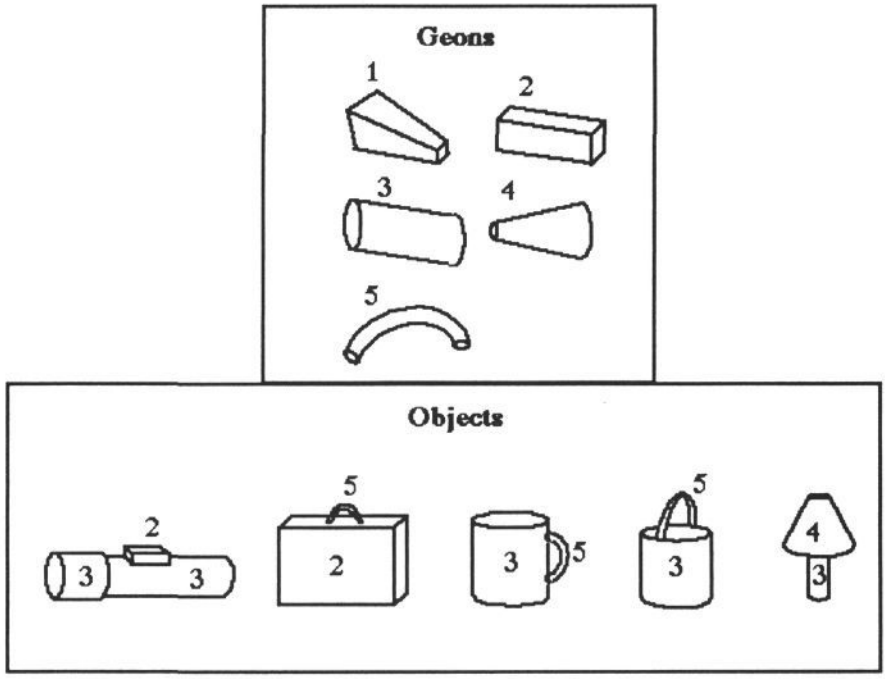

Figure 1: Five geons and five objects. Note that the pail and the bucket are composed of the same geons but in different relations. TOP-OF is a relation. If the page is rotated 180 the pail will resemble a cap and the lamp a trowel or shovel.

The viewpoint invariance derives from a classification of the edges corresponding to the orientation and depth discontinuities of the object's surfaces, according to viewpoint invariant contrasts (VICs). VICs are differences in nonaccidental properties, i.e., properties of edges that are unaffected (or largely unaffected) by rotation in depth, such as straight vs. curved, vertices formed at the cotermination of edges (L, Y, arrow, and tangent Y), and whether edges are parallel or nonparallel. In the current version of the theory there are 24 geons generated by these contrasts which serve as a partition of the set of generalized cylinders into convex or singly concave volumes. Double concavities are taken as candidates for parsing points by which a complex object is decomposed into its geons.

In a current theoretical effort, we are exploring a redefinition of the geons in terms of those volumes that minimize viewpoint uncertainty over the aspect graph (Koenderink & Van Doorn, 1979). For example, a cylinder can be identified as such from almost any viewpoint, except for those cases where it projects as an ellipse or quadrilateral. If the areas of the viewing sphere corresponding to each projection of a volume, ignoring aspect ratio and degree of curvature, are converted to probabilities, say .98 for the cylinder, and .01 for the ellipse, and .01 for the quadrilateral, then geons have the property of low uncertainty in the Shannon

sense, i.e., they yield low values for -p(i)log2p(i), where each of the p(i)s refers to the proportion of the surface of the viewing sphere occupied by each interpretation. A complex volume or shape would change its interpretation many times over the viewing sphere and thus have high uncertainty. The lowest uncertainty, zero, would be for a sphere. Obviously, low viewpoint uncertainty is equivalent to high viewpoint invariance. Although it is likely that the volumes that would pass the criterion of low viewpoint uncertainty would largely correspond to the current set of geons, this redefinition allows a more principled basis for the derivation of the geons. In particular, low viewpoint uncertainty, by minimizing change as different views of an objects are encountered, may yield optimal conditions for a self-organizing neural network to develop hidden units that function as geon detectors.

3.1 Relation of geon theory to other current models of biological object recognition.

All models of biological object recognition assume an input layer that can be approximated by a lattice of filters that cover the visual field. Each node in the lattice is occupied by a number of simple filters (typically modeled as a Gaussian damped sinusoid, termed a Gabor filter), each "tuned" to a particular orientation and spatial frequency at that position in the visual field, though there is considerable overlap in the coverage of the filters at nearby nodes. Each of the Gabor filters corresponds to a simple cell and each node corresponds to a single hypercolumn in the initial region of the cortex that receives visual inputs (area V1). There are approximately 2,100 hypercolumns in the V1 area of each hemisphere. At the object layer are units corresponding to the various object categories that the network is supposed to differentiate.

Most current theories of object recognition can be distinguished along two dimensions: a) the degree to which they assume intermediate representations between input and output, and b) whether they assume the representation is defined in a coordinate space or is a structural description.

Intermediate Representations. A recent review article by Dickinson, Pentland, and Rosenfeld (1992) provided an excellent account of this dimension of theorizing. Some theories assume no intermediate representation, mapping the output of the filters directly onto the object layer, as in the theories of Poggio & Edelman, 1990, and Buhmann, Lange, von der Malsburg, Vorbrggen, & Wrtz). At the other extreme, are those theories such as geon theory that assume volumetric primitives. Between these two extremes, are theories that specify lines (as in Lowe's, 1987, model) and surfaces, for example. Dickinson et al observed a tradeoff between the ease at which the representation could be determined from the image and the ease at which object representations in the data base could be successfully activated. Thus, filter values are easy to compute but are of very little help in selecting among objects given that viewing conditions can change. Geons are extremely difficult to extract but can be quite powerful in accessing object representations.

Coordinate Space. Another important dimension by which theories can be classified is whether they assume a coordinate space that preserves retinal proximities for the matching of input against stored representation (as in Poggio & Edelman, 1990) or whether a structural description, consisting of elements (such as parts)

and explicit relations among these elements, as in Hummel and Biederman (1992).

4 The Psychophysics of Entry Level Classification

4.1 Picture Priming Experiments

What are the data that we wish to model? We have recently completed a number of picture-priming experiments in which subjects name, as quickly and as accurately as possible, briefly (100-500 msec) presented line drawing of objects. The image is followed by a mask consisting of a random-appearing arrangement of contours. The pictures are then shown a second time, several minutes later (in a different order). There is marked facilitation (or priming) in the speed and accuracy of naming on this second block of trials. A part of the facilitation is visual (and not just verbal or conceptual) in that an image of the same basic level class but of another shape, such as another type of chair, is named more slowly than the original object.

4.2 Strong Invariance

What happens if the image, on its second presentation, is projected to another part of the retina (an equal distance from fixation as on its first presentation), or at another size, or at another orientation in depth than what it was when first presented, or reflected? Would there be any priming? A weak form of invariance would be supported if there was some priming, but less than if the object was at its original position, size, or orientation. Remarkably, the results clearly supported strong invariance–there was no effect of changing position, size, reflection, or orientation in depth (up to parts occlusion) (Biederman & Cooper, 1991a,b; 1992; Biederman & Gerhardstein, 1993). Biederman & Gerhardstein also showed that depth invariance could readily be achieved with nonsense objects (in a same-different matching task), indicating that invariance could occur in the absence of a familiar object model.

On computational grounds, the invariances seem entirely reasonable in that the alternative, a separate representation of an object for each of its image manifestations, would require a prohibitively large number of representations. The invariance in recognition speed, i.e., the strong invariance, moreover, is inconsistent with the hypothesis (such as that advanced by Ullman, 1989) that recognition is achieved through template transformations for translating, scaling, or rotating an image or template so as to place the two in correspondence, as such transformations would (presumably) require time for their execution, not to mention the formidable initial problem of selecting the appropriate transformation to apply to an unknown image. Transformational models can achieve weak invariance, but not strong invariance (unless they assume transformations with no time cost).

The phenomenon of strong invariance for position, reflection, size, and orientation in depth, may not be just a psychophysical curiosity but may reflect a fundamental partitioning in the way in which the brain handles shape. In recent

years it has become apparent that there are least two extrastriate cortical visual systems. Both start at the striate cortex (V1), the primary projection area in the occipital cortex, which receives direct inputs from the retina, by way of the lateral geniculate body. One system, termed the "where" system, extends dorsally from V1, to the posterior parietal (PP) cortex, and has been implicated in memory for location. The other, termed the "what" system, extends ventrally, from V1 to V2, V4, and then to the inferior temporal (IT) cortex and appears to determine object recognition.

Why should these particular visual systems have evolved separately? A possible clue lies in the realization that "where" may be too narrow a characterization of the function of the dorsal system. Instead, there is ample evidence that the dorsal system mediates motor interaction, in general. To be sure, location ("where") is a critical component of successful motor interaction: To pick up a coffee cup requires that one reach for the cup in a given location. If the cup is on the left side, one cannot reach for it on the right. Similarly, the metrics and dynamics of the grasp are closely tuned to the actual size and orientation of the handle. The motor interaction is not limited to reaching and grasping but also includes other motor functions such as navigation toward some location and avoiding obstacles along the way.

We can now appreciate a possible computational basis for why two separate visual systems might have evolved. The one for recognition must be able to activate the same representation despite variation in stimulus parameters that are critical for motor interaction, viz., position, size, and orientation in depth. Similarly, motor interaction does not require knowing what the object is. That is, we can reach for or navigate to or around an object without identifying it. It is our conjecture that the motor interaction system employs metric information in a coordinate space but the recognition system employs qualitative (viewpoint-invariant) differences in a structural description.

4.3 Parts

We thus have ample evidence for invariances in recognition, but how can we describe the representation itself? Should it be the particular edges and vertices presented in the image? Or a specific object model, such as a grand piano? Or of the object's parts? Or all three? Somewhat surprisingly, there is a single answer to this question. The magnitude of the perceptual priming is completely determined by the capacity to activate representations of the parts of an object; there is no contribution from the features (vertices and edges) actually present in the image or the global shape or an object model.

Nature of the representation: Priming Contour-Deleted Images. To assess what information is affected by priming, Biederman and Cooper (1991a) measured naming speed and accuracy with briefly presented stimuli by deleting every other image feature (edge and vertex) from each geon to create two complementary images of each object, as shown in Figure 2. That is, the two images for each object, when superimposed, would form an intact picture with no overlap in contour. The complementary images were created in such a way that each part (or geon) of the object could be recovered (or fail to be recovered) from each of the images. Although complementary images shared no edges and vertices, they presumably

would activate the same components. Because the amount of contour deleted from each image was substantial and included vertices, it is unlikely that a local process of filling-in could have completed the contour of these images (see Biederman & Cooper, 1991a for a more complete discussion).

These results are supportive of an earlier demonstration by Biederman (1987) who showed that pictures of common objects were unrecognizable when the contour was deleted in such a manner that the geons could not be recovered. When the same amount of contour deletion allowed recovery of the geons, recognition could be perfect.

On a first block of trials, subjects viewed a number of brief presentations of one member from each complementary pair which they named as quickly and as accurately as possible. On the second block, they would see either the identical image, its complement, or a same name-different

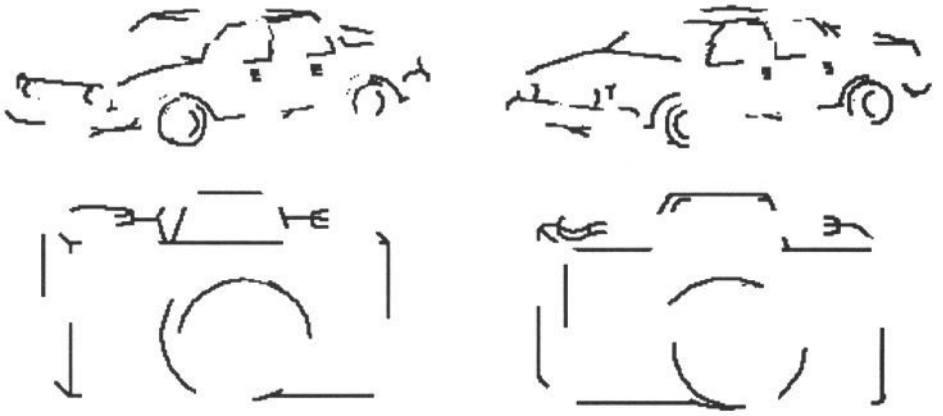

Figure 2: Sample complementary images produced by deleting alternate vertices and edges from each geon. From Cooper, Biederman, & Hummel (1991).

exemplar image (also contour-deleted) from a category with the same name and basic level concept but with a different shape. Mean correct naming reaction times and error rates were markedly lower to the identical image than the different exemplars, indicating that a portion of the priming was indeed visual. The critical comparison, however, concerned the relative performance of the complementary condition. If priming was a function of repetition of the specific vertices and edges in the image, then the complementary condition would have been equivalent to the different exemplar condition, as neither shared any features with the original image. Remarkably, there was no difference in performance in naming complementary and identical images, indicating that none of the priming could be attributed to the specific vertices and lines actually present in the image.

What then caused the priming? One possibility is that it was the parts (or geons), as they were common in the two conditions. Another possibility is that a semantic model of a subordinate category, e.g., a grand piano, rather than the basic level category, e.g., piano, was activated in the initial presentation. To test this possibility, an experiment was run in which complementary images were cre-

ated by deleting half the parts of the objects. With these stimuli, presumably, the same subordinate category would be activated from either members of a complementary pair, but through different parts. (This experiment required the use of objects that would require at least six parts to look complete.) The design was otherwise identical to that of the previous study. As with the first experiment, performance with the identical images was better than with the different exemplars. Now, however, performance with the complements was equivalent to that with the different exemplars, indicating that none of the priming could be attributed to a subordinate semantic model. By elimination, the two experiments, taken together, suggest that all of the priming can be attributed to a representation of the parts (and their interrelations) of the object.

4.4 Viewpoint-Invariant vs. Metric Properties?

But are these parts geons? Cooper & Biederman (1993) recently reported a series of experiments in which they compared the relative importance of differences in aspect ratio of a part (a metric property) with a difference in a viewpoint invariant characterization of the part. For example, a lamp that had a cylinder as its base could have been changed to one of a different aspect ratio or the cylinder could have been changed to a brick of the same aspect ratio. The results clearly supported the greater importance of the geon changes. Similarly, Biederman & Gerhardstein (1993) found that there was no effect of rotation in depth on recognition speed in a priming task as long as the object could be readily described as an arrangement of distinctive geons and the original geons remained in view. The rotation, of course, altered the aspect ratio and degree of curvature of the objects.

5 JIM: A Neural Net Implementation of Geon Theory

A neural net implementation of geon theory (Hummel & Biederman, 1992), termed JIM (for John and Irv's Model), is a seven layer network whose architecture is shown in figure 3 that takes as input a line drawing representing the orientation and depth discontinuities of an object and activates units representing a viewpoint-invariant structural description of the object specifying its geons and their relations. This description is activated regardless of whether the model has previously been exposed to the object.

The model's capacity for structural description derives from its solution to the dynamic binding problem of neural networks (specifying what goes with what): Independent units representing an object's parts (in terms of their shape attributes and relations) are bound temporarily when those attributes occur in conjunction in the system's input. Binding is initially achieved through "fast enabling links" (FELs) that phase lock the oscillatory activity of cells that are tuned to oriented image edges (in the first layer, which is a toy V1). In particular, the FELs cause synchrony in the firing of activated units that are collinear, parallel, or coterminate.

In the third layer (L3) of figure 3, for example, the units marked by filled circles represent the attributes of a brick. These all fire together and out of phase with

the units marked by the open circles, representing the cone. (s = straight; c = curved for the Axis and Cross-Section. p = parallel; n = nonparallel for Sides. v,d, and h refer to vertical, diagonal, and horizontal for the Orientation of the axis [fine and coarse]. b = bottom; t = top for Vertical Position.) Only 36 cells in that layer are required to specify the information for each part. Because the binding is temporary, these same cells can be used to code the other parts of the object as well as the parts of other objects, no matter where they are in the visual field. The binding is thus achieved without positing additional units for "anding."

L4 and L5 derive invariant relations (so that the same "above" cell fires in phase to the cone independent of where it is next located).The outputs of L3 that represent the distributed values of a geon and its orientation and aspect ratio, along with the outputs of L5 representing the relations, provide an input vector that self-organizes a unit in L6, termed a geon feature assembly. Units in L7 are object cells that self organize to an integration over successive outputs from L6. These operations produce a parts-based structural description that is subsequently used directly as a basis for viewpoint-invariant recognition. The model's recognition performance conforms well to the results from the shape priming experiments. Moreover, the manner in which the model's performance degrades due to accidental synchrony produced by an excess of phase sets suggests a basis for a theory of visual attention.

5.1 Binding via Fast Enabling Links (FELs)

A major contribution of the model is its proposal of a solution to the binding problem–determining what goes with what. Each cell has two kinds of connections to other cells: a) the standard connections that excite or inhibit the firing of a target cell, and b) fast enabling links (FELs) that cause (enable) cells that are simultaneously active to fire together if their receptive fields are: a) cocircular (or collinear), b) closely parallel, or c) coterminate. If a cell fires, it passes activation (i.e., excitation and inhibition) in a standard manner and an enabling signal over its FELs. In general, the activation and inhibition will not be to the same units that share FELs. The FELs produce synchronous firing of all the cells that are activated by a given geon while allowing cells activated by different geons to fire out of phase with each other. By not having FELs between the segments comprising a T vertex, as where the sides of the cone in the image in Figure 3 occlude the back edge of the brick on which it rests, the model causes all the features activated by each geon to fire in phase, but out of phase with the features activated by other geons.

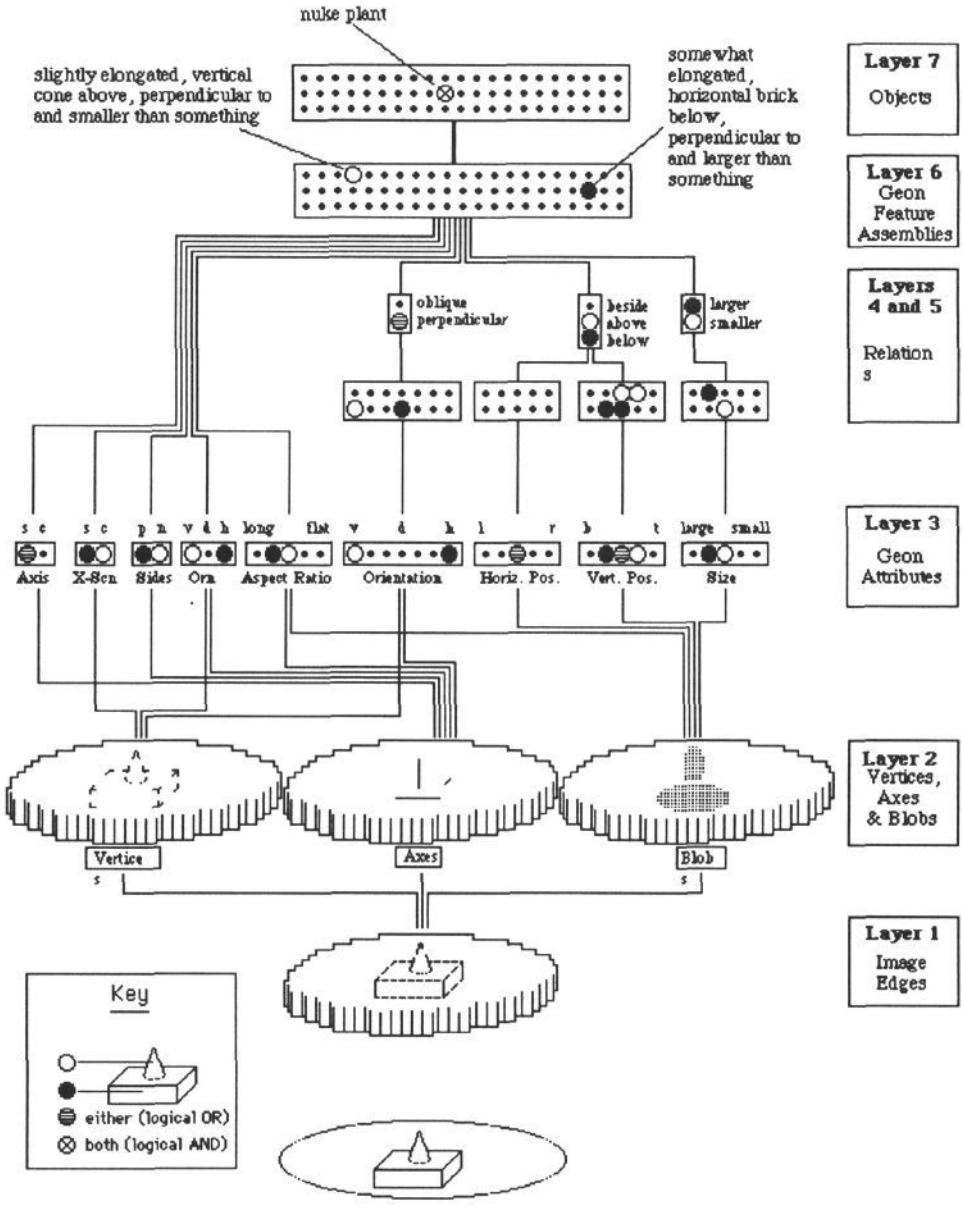

Figure 3: Neural net model for object recognition. From Hummel & Biederman (1992). With permission of the American Psychological Association.

6 Nongeonic Objects

We need to consider those classes of objects that ultimately achieve recognition, such as those that do not readily decompose into parts or are highly irregular. RBC predicts that these will, in general, be recognized more slowly than their more regular counterparts. Moreover, where the differences are not geonic, viewpoint invariance over depth rotation is lost (Biederman & Gerhardstein, 1993) and it is unlikely that differences in entry level classes in any culture would distinguish among such entities. The difficulty that people have with such objects serve to confirm geon theory.

In some cases, however, recognition does proceed quickly, although any currently implemented geon finder would have a terribly difficult time. Often these are with objects that have a great deal of decoration or detail. Our best guess is that in such instances activation of geons at a coarse scale has occurred, despite the detail. But it remains a challenge as to how to achieve such recognition in an implemented model.

Faces may represent a special case, in which the two-layer networks are more appropriate than an invariant-parts model. A face-recognition system of Buhmann et al. (1991) does an excellent job in recognizing faces, even with changes in expression and modest changes in orientation. JIM would fail at this–it would know that something was a face but not whose face it was. But Fiser, Biederman, & Cooper (1993) showed that the Buhmann et al system was insensitive to the effects of potent psychophysical variables in entry-level object recognition. For example, it recognized nonrecoverable as well as recoverable images but recognized a stored member of a complementary pair (as in figure 2) better than the complement. People cannot recognize nonrecoverable images and their recognition of a complement is as good as the originally presented image of a complementary pair.

Evidence that face recognition may require a different solution can be seen in the near impossibility of recognition in the presence contrast reversal (as with a photographic negative) or rotation in the plane. Entry-level object recognition is hardly affected by the former and only modestly by the latter. Perhaps it is not surprising that a neurological condition exists, prosopagnosia, which is results in impairment of face recognition but not entry-level object recognition.

References

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

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

[3] Biederman, I., & Cooper, E. E. (1991b). Evidence for complete translational and reflectional invariance in visual object priming. Perception, in press.

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

[5] Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence for 3D viewpoint invariance. Journal of Experimental Psychology: Human Perception and Performance, In press (December).

[6] Buhmann, J., Lange, J., von der Malsburg, C., Vorbrggen, J. C., & Wrtz, R. P. (1991). Object recognition in the dynamic link architecture: Parallel implementation of a transputer network. In B. Kosko (Eds.), Neural Networks for signal processing. (Pp 121-159). Englewood Cliffs, NJ: Prentice-Hall.

[7] Cooper, E. E., & Biederman, I. (1993). Metric versus viewpoint-invariant shape differences in visual object recognition. Poster presented at the Annual Meeting of The Association for Research in Vision and Ophthalmology, Sarasota, Fl. May.

[8] 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.

[9] Dickinson, S. J., Pentland, A. P., & Rosenfeld, A. (1992). From volumes to views: An approach to 3-D object recognition. Computer Vision, Graphics, and Image Processing: Image Understanding, 55, 130-154.

[10] Fiser, J., Biederman, I., & Cooper, E. E. (1993). To what extent can matching algorithms based on direct outputs of spatial filters account for human shape recognition? Submitted for publication.

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

[12] Koenderink, J. J., & van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. Biological Cybernetics, 32, 211-216.

[13] Lowe, D. G. (1987). The viewpoint consistency constraint. International Journal of Computer Vision, 1, 57-72.

[14] Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. Cognition, 32, 193-254.

[15] von der Malsburg, C. (1987). Synaptic plasticity as a basis of brain organization. In J. P. Chaneaux & M. Konishi (Eds.), The Neural and Molecular Bases of Learning (pp. 4111-432). John Wiley & Sons Limited

This research was supported by AFOSR Research Grant 90-0274 and McDonnell-Pew Foundation Program in Cognitive Neuroscience grant T89-01245-029 to IB, and an NSF Graduate Fellowship to E. E. C. The authors thank Toshio Uchiyama for his help with the ms. IB's Email: ib@rana.usc.edu.