Psychological Review
1992, Vol. 99, No. 3, 480-517

0033-295X/92/$3.00

Dynamic Binding in a Neural Network for Shape Recognition

John E. Hummel and Irving Biederman
University of Minnesota, Twin Cities

Given a single view of an object, humans can readily recognize that object from other views that preserve the parts in the original view. Empirical evidence suggests that this capacity reflects the activation of a viewpoint-invariant structural description specifying the object's parts and the relations among them. This article presents a neural network that generates such a description. Structural description is made possible through a solution to the *dynamic binding problem:* Temporary conjunctions of attributes (parts and relations) are represented by synchronized oscillatory activity among independent units representing those attributes. Specifically, the model uses synchrony (a) to parse images into their constituent parts, (b) to bind together the attributes of a part, and (c) to bind the relations to the parts to which they apply. Because it conjoins independent units temporarily, dynamic binding allows tremendous economy of representation and permits the representation to reflect the attribute structure of the shapes represented.

A brief glance at Figure 1 is sufficient to determine that it depicts three views of the same object. The perceived equivalence of the object across its views evidences the fundamental capacity of human visual recognition: Object recognition is invariant with viewpoint. That is, the perceived shape of an object does not vary with position in the visual field, size, or, in general, orientation in depth. The object in Figure 1 is unfamiliar, so the ability to activate a viewpoint invariant representation of its shape cannot depend on prior experience with it. This capacity for viewpoint invariance independent of object familiarity is so fundamental to visual shape classification that modeling visual recognition is largely a problem of accounting for it.[1]

This article presents a neural network model of viewpoint invariant visual recognition. In contrast with previous models based primarily on template or feature list matching, we argue that human recognition performance reflects the activation of a viewpoint invariant structural description specifying both the visual attributes of an object (e.g., edges, vertices, or parts) and the relations among them. For example, the simple object in Figure 1 might be represented as a vertical cone on top of a horizontal brick. Such a representation must bind (i.e., conjoin) the shape attribute *cone shaped* with the relational attribute *on top of* and the attribute *brick shaped* with the attribute *below;* otherwise, it would not specify which volume was on top of which. It is also necessary to use the same representation for *cone shaped* whether the cone is on top of a brick, below a cylinder, or by itself in the image otherwise the representation would not specify that it was the same shape each time. Traditional connectionist/neural net architectures cannot represent structural descriptions because they bind attributes by positing separate units for each conjunction (cf. Fodor & Pylyshyn, 1988). The units used to represent *cone shaped* would differ depending on whether the cone was on top of a brick, below a cylinder, or instantiated in some other relation. Representing structural descriptions in a connectionist architecture requires a mechanism for binding attributes dynamically; that is, the binding of attributes must be temporary so that the same units can be used in multiple conjunctions. A primary theoretical goal of this article is to describe how this dynamic binding can be achieved. The remainder of this section motivates the model by presenting the aforementioned arguments in greater detail.

Approaches to Visual Recognition: Why Structural Description?

Prima facie, it is not obvious that successful shape classification requires explicit and independent representation of shape attributes and relations. Indeed, most models of visual recognition are based either on template matching or feature list matching, and neither of these approaches explicitly represents relations. We critically evaluate each in turn. Some of the shortcomings of template matching and feature list matching models were described a quarter of a century ago (Neisser, 1967) but still have relevance to recent modeling efforts. We conclude from this critical review that template and feature models suffer from the same shortcoming: They trade off the capacity to

This research was supported by National Science Foundation graduate and Air Force Office of Scientific Research (AFOSR) postdoctoral fellowships to John E. Hummel and by AFOSR research grants (88-0231 and 90-0274) to Irving Biederman. We are grateful to Dan Kersten, Randy Fletcher, Gordon Legge, and the reviewers for their helpful comments on an earlier draft of this article and to Peter C. Gerhardstein for his help on early simulations.

Irving Biederman is now at the Department of Psychology, Hedco Neuroscience Build., University of Southern California.

Correspondence concerning this article should be addressed to John E. Hummel, who is now at the Department of Psychology, University of California, Franz Hall, 405 Hilgard Avenue, Los Angeles, California 90024-1563.

[1] Though the perceived shape is invariant, we are of course aware of the differences in size, orientation, and position among the entities in Figure 1. Those attributes may be processed by a system subserving motor interaction rather than recognition (Biederman & Cooper, 1992).

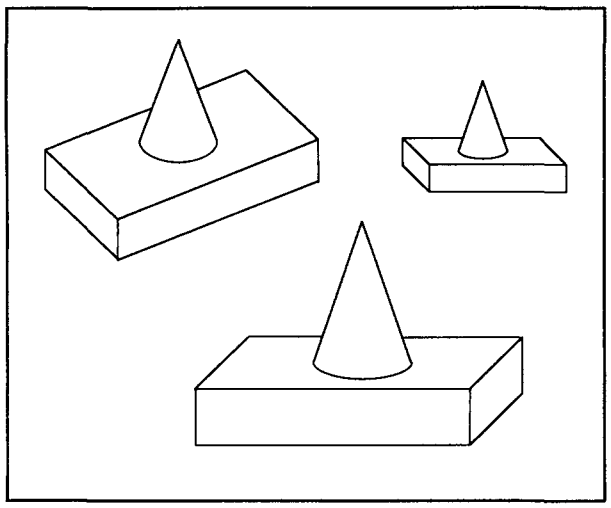


Figure 1. This object is readily detectable as constant across the three views despite its being unfamiliar.

represent attribute structures with the capacity to represent relations.

Template Matching

Template matching models operate by comparing an incoming image (perhaps after filtering, noise reduction, etc.) against a template, a representation of a specific view of an object. To achieve viewpoint invariant recognition, a template model must either (a) store a very large number of views (templates) for each known object or (b) store a small number of views (or 3-D models) for each object and match them against incoming images by means of transformations such as translation, scaling, and rotation. Recognition succeeds when a template is found that fits the image within a tolerable range of error. Template models have three main properties: (a) The fit between a stimulus and a template is based on the extent to which active and inactive points[2] in the image spatially correspond to active and inactive points in the template; (b) a template is used to represent an entire view of an object, spatial relations among points in the object coded implicitly as spatial relations among points in the template; and (c) viewpoint invariance is a function of object classification. Different views of an object are seen as equivalent when they access the same object label (i.e., either by matching the same template or by matching different templates with pointers to the same object label).

Template matching currently enjoys considerable popularity in computer vision (e.g., Edelman & Poggio, 1990; Lowe, 1987; Ullman, 1989). These models can recognize objects under different viewpoints, and some can even find objects in cluttered scenes. However, they evidence severe shortcomings as models of human recognition. First, recall that template matching succeeds or fails according to the number of points that mismatch between an image and a stored template and therefore predicts little effect of where the mismatch occurs. (An exception to this generalization can be found in alignment models, such as those of Ullman, 1989, and Lowe, 1987. If enough critical alignment points are deleted from an image—in Ullman's model, three are required to compute an alignment—the alignment process will fail. However, provided alignment can be achieved, the degree of match is proportional to the number of corresponding points.) Counter to this prediction, Biederman (1987b) showed that human recognition performance varied greatly depending on the locus of contour deletion. Similarly, Biederman and Cooper (1991b) showed that an image formed by removing half the vertices and edges from each of an object's parts (Figure 2a) visually primed[3] its complement (the image formed from the deleted contour, Figure 2b) as much as it primed itself, but an image formed by deleting half of the parts (Figure 2c) did not visually prime its complement at all (Figure 2d). Thus, visual priming was predicted by the number of parts shared by two images, not by the amount of contour.

Second, template models handle image transformations such as translation, rotation, expansion, and compression by applying an equivalent transformation to the template. Presumably, such transformations would require time. However, Biederman and Cooper (1992) showed no loss in the magnitude of visual priming for naming object images that differed in position, size, or mirror image reflection from their original presentation. Similarly, Gerhardstein and Biederman (1991) showed no loss in visual priming for objects rotated in depth (up to parts occlusion).

A more critical difficulty arises from the template's representing an object in terms of a complete view. Global image transformations, albeit time consuming, are not catastrophic, but any transformation applied only to a part of the object, as when an object is missing a part or has an irrelevant part added, could readily result in complete classification failure. Yet only modest decrements, if any, are observed in recognition performance with such images (Biederman, 1987a). Biederman and Hilton (1991) found that visual priming was only modestly affected by changes of 20% to 40% in the aspect ratio of one part of two-part objects. However, qualitative changes (e.g., from a cylinder to a brick) that preserved aspect ratio resulted in priming decrements that were at least as great as those produced by the aspect ratio changes. In essence, the core difficulty for template models is that they do not make explicit the information that is critical to the representation of shape by humans. For example, which of the two objects in the bottom of Figure 3 is more similar to the standard in the top of the figure? Most people would judge Object A to be more similar because the cone on Object B is rounded rather than pointed. However, a template model would select B as more similar because the brick on the bottom of A is slightly flatter than the bricks of the other two objects. As a result, the best fit between A and the standard mismatches along the entire bottom edge. The number of mismatching pixels between A and the standard is 304, compared with only 70 pixels' mismatch for B.

[2] *Point* is used here to refer to any simple, localized image element, including a raw image intensity value, zero-crossing, or local contour element.

[3] *Visual* (rather than name or concept) *priming* was defined as the advantage enjoyed by an identical image over an object with the same name but a different shape. The task required that subjects rapidly name briefly presented masked images.

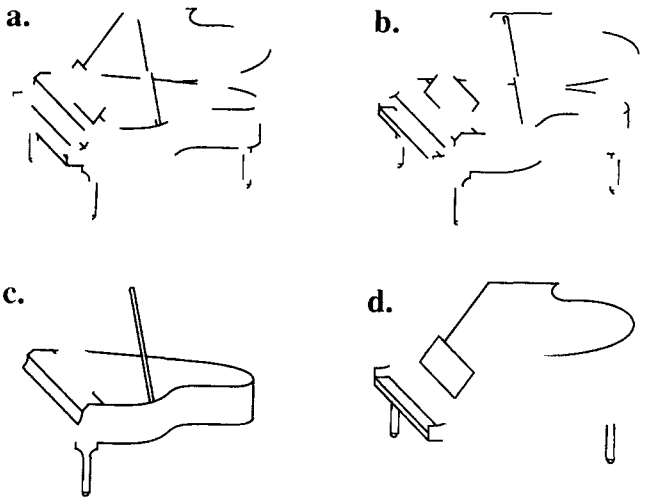


Figure 2. Examples of contour-deleted images used in the Biederman and Cooper (1991b) priming experiments. (a: An image of a piano as it might appear in the first block of the experiment in which half the image contour was removed from each of the piano's parts by removing every other edge and vertex. Long edges were treated as two segments. b: The complement to Panel a formed from the contours removed from Panel a. c: An image of a piano as it might appear in the first block of trials in the experiment in which half the parts were deleted from the piano. d: The complement to Panel c formed from the parts deleted from Panel c. From "Priming Contour-Deleted Images: Evidence for Intermediate Representations in Visual Object Recognition" by I. Biederman & E. E. Cooper, 1991, *Cognitive Psychology, 23,* Figures 1 and 4, pp. 397, 403. Copyright 1991 by Academic Press. Adapted by permission.)

Another difficulty with templates—one that has been particularly influential in the computer vision community—is their capacity to scale with large numbers of objects. In particular, the critical alignment features used to select candidate templates in Ullman's (1989) and Lowe's (1987) models may be difficult to obtain when large numbers of templates must be stored and differentiated. Finally, a fundamental difficulty with template matching as a theory of human recognition is that a template model must have a stored representation of an object before it can classify two views of it as equivalent. As such, template models cannot account for viewpoint invariance in unfamiliar objects. Pinker (1986) presented several other problems with template models.

The difficulties with template models result largely from their insensitivity to attribute structures. Because the template's unit of analysis is a complete view of an object, it fails to explicitly represent the individual attributes that define the object's shape. The visual similarity between different images is lost in the classification process. An alternative approach to recognition, motivated largely by this shortcoming, is feature matching.

Feature Matching

Feature matching models extract diagnostic features from an image and use those as the basis for recognition (e.g., Hinton, 1981; Hummel, Biederman, Gerhardstein, & Hilton, 1988; Lindsay & Norman, 1977; Selfridge, 1959; Selfridge & Neisser, 1960). These models differ from template models on each of the three main properties listed previously: (a) Rather than operating on pixels, feature matching operates on higher order image attributes such as parts, surfaces, contours, or vertices; (b) rather than classifying images all at once, feature models describe images in terms of multiple independent attributes; and (c) visual equivalence consists of deriving an equivalent set of visual features, not necessarily accessing the same final object representation. Feature matching is not subject to the same shortcomings as template matching. Feature matching models can differentially weight different classes of image features, for example, by giving more weight to vertices than contour midsegments; they can deal with transformations at the level of parts of an image; and they can respond in an equivalent manner to different views of novel objects.

However, viewpoint-invariant feature matching is prone to serious difficulties. Consider a feature matching model in which shapes are represented as collections of vertices. To achieve just simple translational invariance (thus ignoring the additional complications associated with scale and orientation invariance), each vertex must, by definition, be represented independently of its location in the visual field. However, this location independence makes the representation insensitive to the spatial configuration of the features. Consequently, any configuration of the appropriate features will produce recognition for the object, as illustrated in Figure 4. This problem is so obvious that it seems as if it must be a straw man: With the right set of features, the problem will surely just go away. It is tempting to think that the relations could be captured by extracting more complex features such as pairs of vertices in particular relations or perhaps triads of vertices (e.g., Figure 5). Indeed, as the features extracted become more complex, it becomes more difficult to find images that will fool the system. However, until the features become templates for whole objects, there will

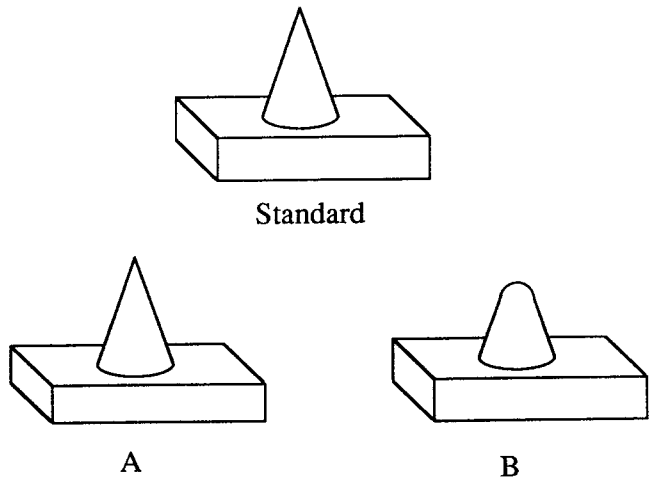


Figure 3. Which object in the lower part of the figure looks more like the standard object in the top of the figure? (A template model would likely judge Panel b as more similar to the standard than Panel a. People do not.)

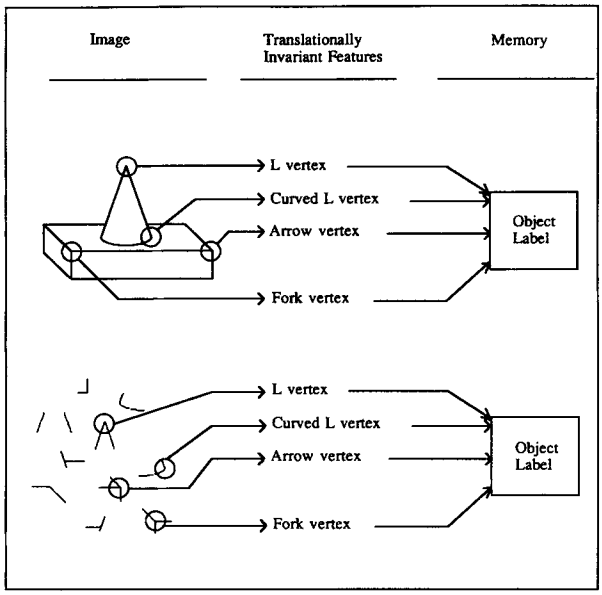


Figure 4. Illustration of illusory recognition with translationally invariant feature list matching on the basis of vertices. (Both images contain the same vertices and will therefore produce recognition for the object.)

always be some disordered arrangement of them that will produce illusory recognition. Furthermore, this difficulty is not limited to models based on two-dimensional (2-D) features such as vertices. As depicted in Figure 6, 3-D features can also be reconfigured to produce illusory recognition.

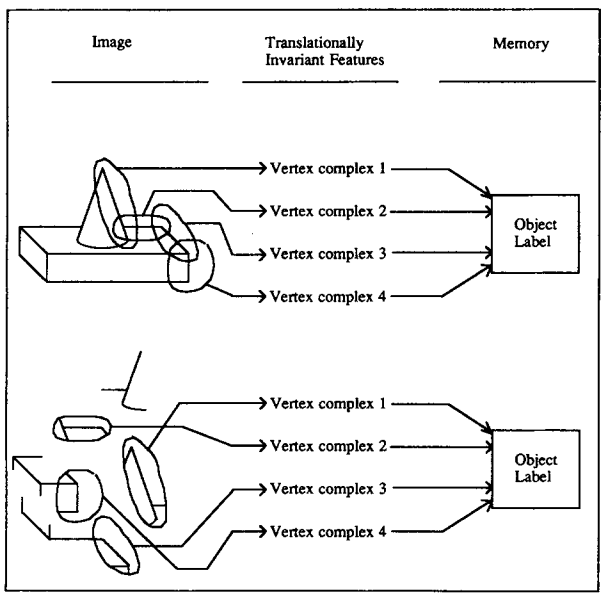


Figure 5. Illustration of illusory recognition, with translationally invariant feature list matching on the basis of feature complexes composed of two or more vertices in particular relations to one another. (Both images contain the critical vertex complexes and will produce recognition for the object.)

Feature and template models both belong to a class of models that treat spatial relations implicitly by coding for precise metric relationships among the points in an object's image. These models trade off the capacity to represent relations with the capacity to represent the individual attributes belonging to those relations. Models that code for small, simple features (feature models) ignore most of the relations among the attributes of an object. Such models will evidence sensitivity to attribute structures but will also be prone to illusory recognition (i.e., they will recognize objects given disordered arrangements of their features). Those that code for whole objects implicitly capture the relations among attributes as relations among points in the features (templates). Such models will avoid illusory recognition but will be insensitive to attribute structures. Models in between these extremes will simply suffer a combination of these difficulties. To escape this continuum, it is necessary to represent the relations among an object's attributes explicitly.

Structural Description

A *structural description* (Winston, 1975) explicitly represents objects as configurations of attributes (typically parts) in specified relations to one another. For example, a structural description of the object in Figure 1 might consist of *vertical cone on top of horizontal brick*. Structural description models avoid the pitfalls of the implicit relations continuum: They are sensitive to attribute structures, but because they represent the relations among those attributes, they are not prone to illusory recogni-

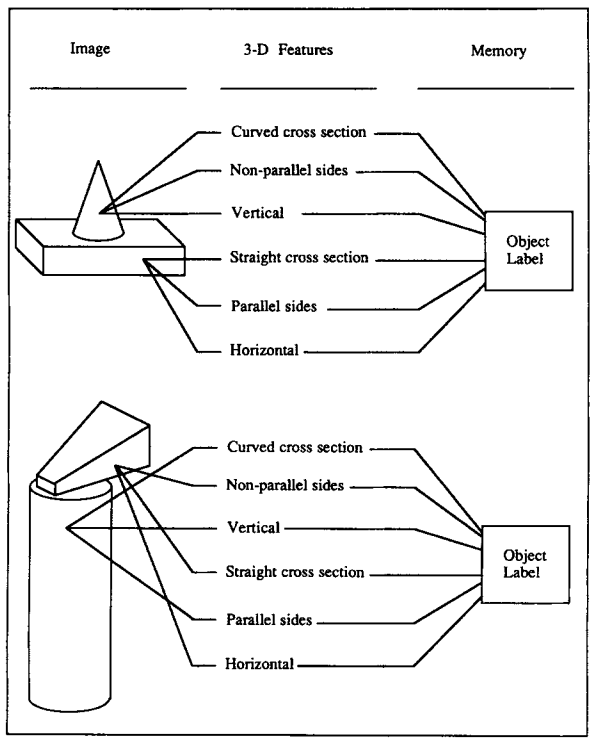


Figure 6. Illustration of illusory recognition with feature list matching on the basis of 3-D features. (Both images contain the critical 3-D features and will produce recognition for the object.)

tion. These models can also provide a natural account of the fundamental characteristics of human recognition. Provided the elements of the structural description are invariant with viewpoint, recognition on the basis of that description will also be invariant with viewpoint. Provided the description can be activated bottom up on the basis of information in the 2-D image, the luxury of viewpoint invariance will be enjoyed for unfamiliar objects as well as familiar ones.

Proposing that object recognition operates on the basis of viewpoint-invariant structural descriptions allows us to understand phenomena that extant template and feature theories cannot handle. However, it is necessary to address how the visual system could possibly derive a structural description in the first place. This article presents an explicit theory, implemented as a connectionist network, of the processes and representations implicated in the derivation of viewpoint-invariant structural descriptions for object recognition. The point of departure for this effort is Biederman's (1987b) theory of recognition by components (RBC). RBC states that objects are recognized as configurations of simple, primitive volumes called *geons* in specified relations with one another. Geons are recognized from 2-D viewpoint-invariant[4] properties of image contours such as whether a contour is straight or curved, whether a pair of contours is parallel or nonparallel, and what type of vertex is produced where contours coterminate. The geons derived from these 2-D contrasts are defined in terms of viewpoint-invariant 3-D contrasts such as whether the cross-section is straight (like that of a brick or a pyramid) or curved (like that of a cylinder or a cone) and whether the sides are parallel (like those of a brick) or nonparallel (like those of a wedge).

Whereas a template representation necessarily commits the theorist to a coordinate space, 2-D or 3-D, the structural description theory proposed here does not. This is important to note because structural descriptions have been criticized for assuming a metrically specified representation in a 3-D object-centered coordinate space. Failing to find evidence for 3-D invariant recognition for some unfamiliar objects, such as crumpled paper (Rock & DiVita, 1987), bent wires (Edelman & Weinshall, 1991; Rock & DiVita, 1987), and rectilinear arrangements of bricks (Tarr & Pinker, 1989), some researchers have rejected structural description theories in favor of representations based on multiple 2-D views. But the wholesale rejection of structural descriptions on the basis of such data is unwarranted. RBC (and the implementation proposed here) predict these results. Because a collection of similarly bent wires, for example, tend not to activate distinctive viewpoint-invariant structural descriptions that allow one wire to be distinguished from the other wires in the experiment, such objects should be difficult to recognize from an arbitrary viewpoint.

Connectionist Representation and Structural Description: The Binding Problem

Given a line drawing of an object, the goal of the current model is to activate a viewpoint-invariant structural description[5] of the object and then use that description as the basis for recognition. Structural descriptions are particularly challenging to represent in connectionist architectures because of the binding problem. *Binding* refers to the representation of attribute conjunctions (Feldman, 1982; Feldman & Ballard, 1982; Hinton, McClelland, & Rumelhart, 1986; Sejnowski, 1986; Smolensky, 1987; von der Malsburg, 1981). Recall that a structural description of the nonsense object in Figure 1 must specify that *vertical* and *on top of* are bound with *cone shaped*, whereas *horizontal* and *below* are bound with *brick shaped*. This problem is directly analogous to the problem encountered by the feature matching models: Given that some set of attributes is present in the system's input, how can it represent whether they are in the proper configuration to define a given object?

The predominant class of solutions to this problem is conjunctive coding (Hinton et al., 1986) and its relatives, such as tensor product coding (Smolensky, 1987), and the interunits of Feldman's (1982) dynamic connections network. A conjunctive representation anticipates attribute conjunctions and allocates a separate unit or pattern for each. For example, a conjunctive representation for the previously mentioned shape attributes would posit one pattern for *vertical cone on top of* and completely separate patterns for *horizontal cone on top of*, *horizontal cone below*, and so forth. A fully *local* conjunctive representation (i.e., one in which each conjunction is coded by a single unit) can provide an unambiguous solution to the binding problem: On the basis of which units are active, it is possible to know exactly how the attributes are conjoined in the system's input. *Vertical cone on top of horizontal brick* and *vertical brick on top of horizontal cone* would activate nonoverlapping sets of units. However, local conjunctive representations suffer a number of shortcomings.

The most apparent difficulty with local conjunctive representations is the number of units they require: The size of such a representation grows exponentially with the number of attribute dimensions to be represented (e.g., 10 dimensions, each with 5 values would require 5^{10} units). As such, the number of units required to represent the universe of possible conjunctions can be prohibitive for complex representations. Moreover, the units have to be specified in advance, and most of them will go unused most of the time. However, from the perspective of representing structural descriptions, the most serious problem with local conjunctive representations is their insensitivity to attribute structures. Like a template, a local conjunctive unit responds to a specific conjunction of attributes in an all-or-none fashion, with different conjunctions activating different units. The similarity in shape between, say, a cone that is on top of something and a cone that is beneath something is lost in such a representation.

These difficulties with local conjunctive representations will not come as a surprise to many readers. In defense of conjunctive coding, one may reasonably protest that a fully local conjunctive representation represents the worst case, both in terms

[4] The term *viewpoint invariant* as used here refers to the tendency of the image feature's classification to remain unchanged under changes in viewpoint. For instance, the degree of curvature associated with the image of the rim of a cup will change as the cup is rotated in depth relative to the viewer. However, the fact that edge is curved rather than straight will remain unchanged in all but a few accidental views of the cup.

[5] Henceforth, it is assumed that the elements of the structural description are geons and their relations.

of the number of units required and in terms of the loss of attribute structures. Both these problems can be attenuated considerably by using coarse coding and other techniques for distributing the representation of an entity over several units (Feldman, 1982; Hinton et al., 1986; Smolensky, 1987). However, in the absence of a technique for representing attribute bindings, distributed representations are subject to cross talk (i.e., mutual interference among independent patterns), the likelihood of which increases with the extent of distribution. Specifically, when multiple entities are represented simultaneously, the likelihood that the units representing one entity will be confused with the units representing another grows with the proportion of the network used to represent each entity. von der Malsburg (1987) referred to this familiar manifestation of the binding problem as the *superposition catastrophe.* Thus, the costs of a conjunctive representation can be reduced by using a distributed representation, but without alternative provisions for binding the benefits are also reduced.

This tradeoff between unambiguous binding and distributed representation is the connectionist's version of the implicit relations continuum. In this case, however, the relations that are coded implicitly (i.e., in the responses of conjunctive units) are binding relations. Escaping this continuum requires a way to represent bindings dynamically. That is, we need a way to temporarily and explicitly bind independent units when the attributes for which they code occur in conjunction.

Recently, there has been considerable interest in the use of temporal synchrony as a potential solution to this problem. The basic idea, proposed as early as 1974 by Peter Milner (Milner, 1974), is that conjunctions of attributes can be represented by synchronizing the outputs of the units (or cells) representing those attributes. To represent that Attribute A is bound to Attribute B and Attribute C to Attribute D, the cells for A and B fire in synchrony, the cells for C and D fire in synchrony, and the AB set fires out of synchrony with the CD set. This suggestion has since been presented more formally by von der Malsburg (1981, 1987) and many others (Abeles, 1982; Atiya & Baldi, 1989; Baldi & Meir, 1990; Crick, 1984; Crick & Koch, 1990; Eckhorn et al., 1988; Eckhorn, Reitboeck, Arndt, & Dicke, 1990; Gray et al., 1989; Gray & Singer, 1989; Grossberg & Somers, 1991; Hummel & Biederman, 1990a; Shastri & Ajjanagadde, 1990; Strong & Whitehead, 1989; Wang, Buhmann, & von der Malsburg, 1991).

Dynamic binding has particularly important implications for the task of structural description because it makes it possible to bind the elements of a distributed representation. What is critical about the use of synchrony in this capacity is that it provides a degree of freedom whereby multiple independent cells can specify that the attributes to which they respond are currently bound. In principle, any variable could serve this purpose (e.g., Lange & Dyer, 1989, used signature activations), so the use of synchrony, per se, is not theoretically critical. Temporal synchrony simply seems a natural choice because it is easy to establish, easy to exploit, and neurologically plausible.

This article presents an original proposal for exploiting synchrony to bind shape and relation attributes into structural descriptions. Specialized connections in the model's first two layers parse images into geons by synchronizing the oscillatory outputs of cells representing local image features (edges and vertices): Cells oscillate in phase if they represent features of the same geon and out of phase if they represent features of separate geons. These phase relations are preserved throughout the network and bind cells representing the attributes of geons and the relations among them. The bound attributes and relations constitute a simple viewpoint-invariant structural description of an object. The model's highest layers use this description as a basis for object recognition.

The model described here is broad in scope, and we have made only modest attempts to optimize any of its individual components. Rather, the primary theoretical statement concerns the general nature of the representations and processes implicated in the activation of viewpoint-invariant structural descriptions for real-time object recognition. We have designed these representations and processes to have a transparent neural analogy, but we make no claims as to their strict neural realism.

A Neural Net Model of Shape Recognition

Overview

The model (JIM; John and Irv's model) is a seven-layer connectionist network that takes as input a representation of a line drawing of an object (specifying discontinuities in surface orientation and depth) and, as output, activates a unit representing the identity of the object. The model achieves viewpoint invariance in that the same output unit will respond to an object regardless of where its image appears in the visual field, the size of the image, and the orientation in depth from which the object is depicted. An overview of the model's architecture is shown in Figure 7. JIM's first layer (L1) is a mosaic of orientation-tuned cells with overlapping receptive fields. The second layer (L2) is a mosaic of cells that respond to vertices, 2-D axes of symmetry, and oriented, elongated blobs of activity. Cells in L1 and L2 group themselves into sets describing geons by synchronizing oscillations in their outputs.

Cells in L3 respond to attributes of complete geons, each cell representing a single value on a single dimension over which the geons can vary. For example, the shape of a geon's major axis (straight or curved) is represented in one bank of cells, and the geon's location is represented in another, thereby allowing the representation of the geon's axis to remain unchanged when the geon is moved in the visual field. The fourth and fifth layers (L4 and L5) determine the relations among the geons in an image. The L4–L5 module receives input from L3 cells representing the metric properties of geons (location in the visual field, size, and orientation). Once active, units representing relations are bound to the geons they describe by the same phase locking that binds image features together for geon recognition. The output of L3 and L5 together constitute a structural description of an object in terms of its geons and their relations. This representation is invariant with scale, translation, and orientation in depth.

The model's sixth layer (L6) takes its input from both the third and fifth layers. On a single time slice (ts), the input to L6 is a pattern of activation describing one geon and its relations to the other geons in the image (a *geon feature assembly*). Each cell in L6 responds to a particular geon feature assembly. The cells

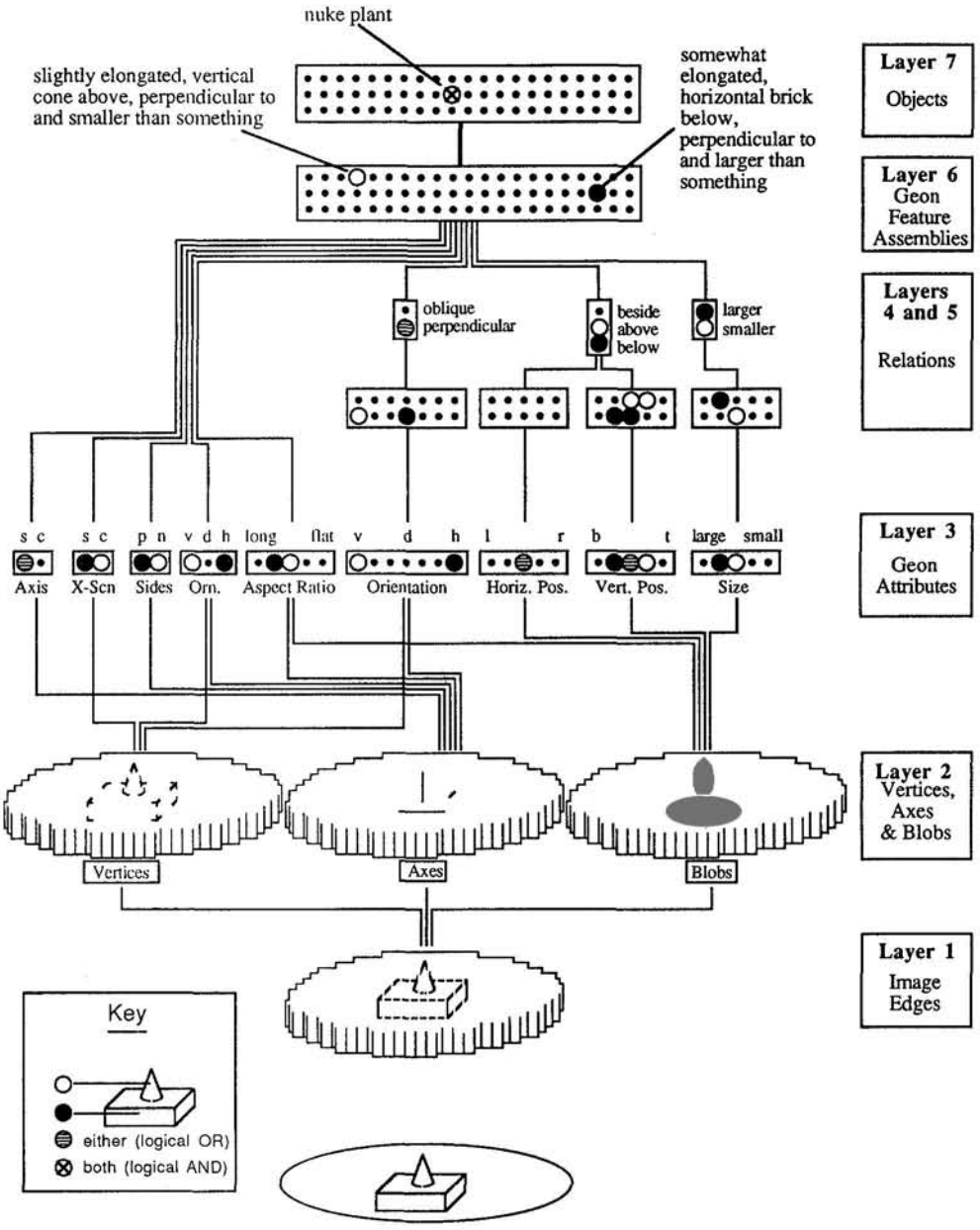


Figure 7. An overview of JIM's architecture indicating the representation activated at each layer by the image in the Key. (In Layers 3 and above, large circles indicate cells activated in response to the image, and dots indicate inactive cells. Cells in Layer 1 represent the edges (specifying discontinuities 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 (Axis), straight (s) or curved (c); Cross-section shape (X-Scn) straight or curved; whether the sides are parallel (p) or nonparallel (n); Coarse orientation (Orn.), vertical (v), diagonal (d), or horizontal (h); aspect ratio, elongated (long) to flattened (flat); Fine orientation (Orientation), vertical, two different diagonals, and four different horizontals; Horizontal position (Horiz. Pos.) in the visual field, left (l) to right (r); Vertical position in the visual field, 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, and cells in Layer 7 respond to complete objects, defined as conjunctions of cells in Layer 6.)

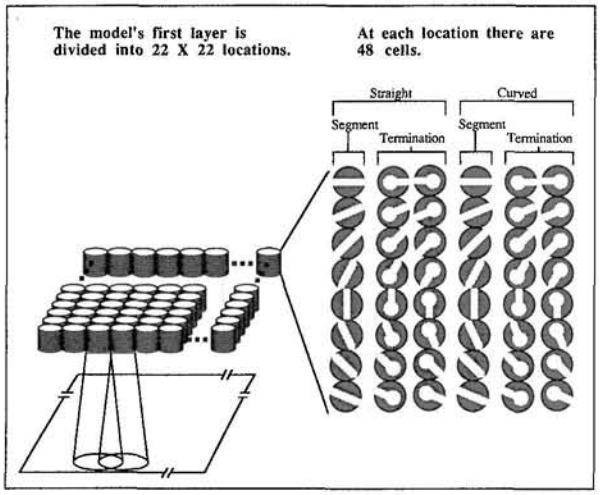


Figure 8. Detail of the model's first layer. (Image edges are represented in terms of their location in the visual field, orientation, curvature, and whether they terminate within the cell's receptive field or pass through it.)

in L7 sum the outputs of the L6 cells over time, combining two or more assemblies into a representation of a complete object.

Layer 1: Representation of a Line Drawing

The model's input layer (L1) is a mosaic of orientation-tuned cells with overlapping receptive fields (Figures 8 and 9). At each of 484 (22 × 22) locations,[6] there are 48 cells that respond to image edges in terms of their orientation, curvature (straight vs. curved), and whether the edge terminates within the cell's receptive field (termination cells) or passes through (segment cells). The receptive field of an L1 cell is thus defined over five dimensions: x, y, orientation, straight versus curved, and termination versus segment. The net input to L1 cell i (N_i) is calculated as the sum (over all image edges j) of products (over all dimensions k):

$$N_i = \sum_j \prod_k (1 - |E_{jk} - C_{ik}|/W_k)^+, \quad (1)$$

where E_{jk} is the value of edge j on dimension k (e.g., the value 1.67 radians on the dimension orientation), C_{ik} is cell i's preferred value on dimension k (e.g., 1.85 radians), and W_k is a parameter specifying the width of a cell's tuning function in dimension k. The superscripted + in Equation 1 indicates truncation below zero; L1 cells receive no negative inputs. The value of edge j on the dimensions x and y (location) is determined separately for each cell i as the point on j closest to C_{ix}, C_{iy}. Image segments are coded coarsely with respect to location: W_x and W_y are equal to the distance between adjacent clusters for segment cells (i.e., if and only if cells i and j are in adjacent clusters, then $|C_{ix} - C_{jx}| = W_x$). Within a cluster, segment cells are inhibited by termination cells of the same orientation. To reduce the calculations and data structures required by the computer simulation, edge orientation is coded discretely (i.e., one cell per cluster codes the orientation of a given image edge), and for terminations, location is also coded discretely (a given termination represented by activity in only one L1 cell). The activation of cell i (A_i) is computed as the Weber function of its net input:

$$A_i = N_i/(1 + N_i). \quad (2)$$

Layer 2: Vertices, Two-Dimensional Axes, and Blobs

The model's second layer (L2) is a mosaic of cells that respond to vertices, 2-D axes of parallel and nonparallel symmetry, and elongated, oriented blobs at all locations in the visual field.

Vertices

At each location in the visual field, there is one cell for every possible two- and three-pronged vertex. These include *L*s, arrows, forks, and tangent *Y*s (see Biederman, 1987b; Malik, 1987) at all possible orientations (Figure 9a). In addition, there are cells that respond to oriented *lone terminations,* endpoints of edges that do not coterminate with other edges, such as the stem of a T vertex. Vertex cells at a given location receive input only from cells in the corresponding location of the first layer. They receive excitatory input from consistent L1 termination cells (i.e., cells representing terminations with the same orientation and curvature as any of the vertex's legs) and strong inhibition from segment cells and inconsistent termination cells (Figure 9b). Each L2 lone termination cell receives excitation from the corresponding L1 termination cell, strong inhibition from all other L1 terminations at the same location, and neither excitation nor inhibition from segment cells. The strong inhibition from L1 cells to inconsistent L2 cells ensures that (a) only one vertex cell will ever become active at a given location and (b) no vertex cells will become active in response to vertices with more than three prongs.

Axes and Blobs

The model also posits arrays of axis-sensitive cells and blob-sensitive cells in L2. The axis cells represent 2-D axes of parallelism (straight and curved) and non-parallel symmetry (straight and curved). However, the connections between these cells and the edge cells of L1 have not been implemented. Computing axes of symmetry is a difficult problem (cf. Brady, 1983; Brady & Asada, 1984; Mohan, 1989) the solution of which we are admittedly assuming. Currently, the model is given, as part of its input, a representation of the 2-D axes in an image. Similarly, cells sensitive to elongated, oriented regions of activity (blobs) are posited in the model's second layer but have not been implemented. Instead, blobs are computed directly by a simple re-

[6] To reduce the computer resources required by the implementation, a square lattice (rather than a more realistic hexagonal lattice) was used in the simulations reported here. However, the use of a square lattice is not critical to the model's performance.

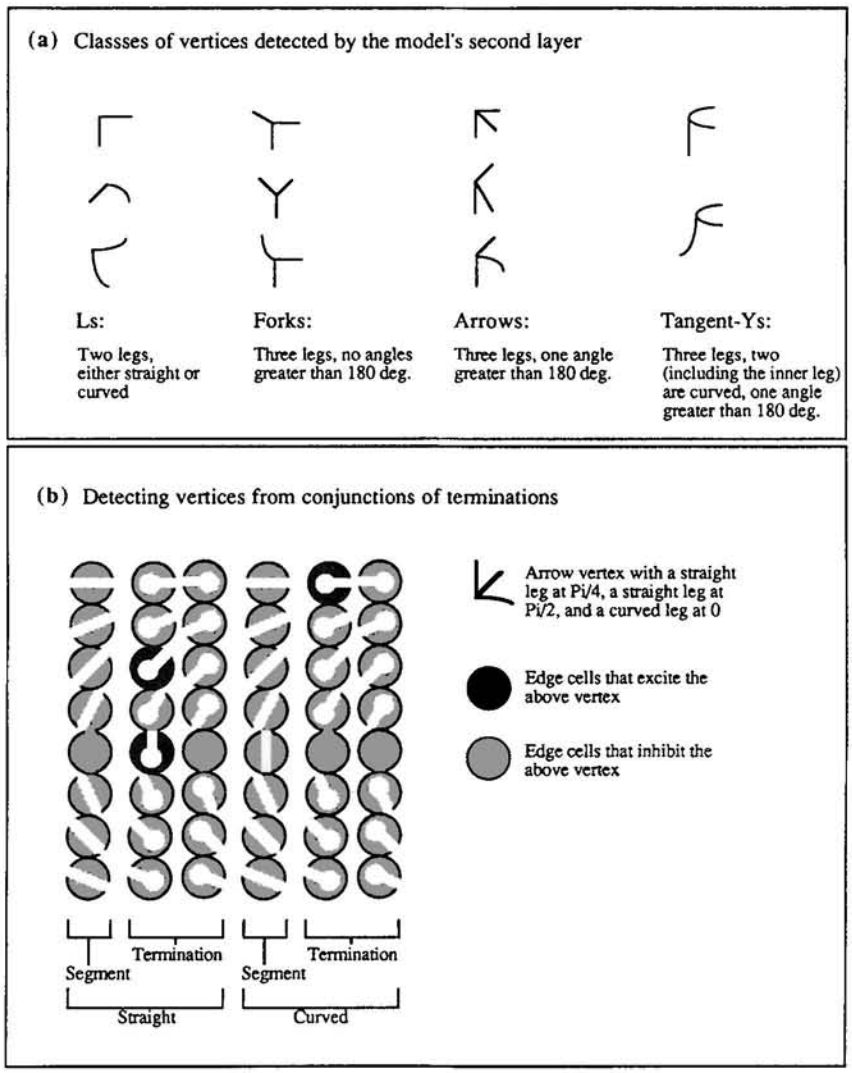


Figure 9. a: An illustration of the types of vertices detected in the model's second layer. b: An illustration of the mapping from the edge cells in a given location in Layer 1 to a vertex cell in the corresponding location of Layer 2. deg. = degree.

gion-filling algorithm.[7] These computations yield information about a geon's location, computed as the blob's central point, size, and elongation. Elongation is computed as the square of the blob's perimeter divided by its area (elongation is an inverse function of Ullman's, 1989, compactness measure).

Axis and blob cells are assumed to be sensitive to the phase relations established in LI and therefore operate on parsed images (image parsing is described in the next section). Because this assumption restricts the computation of blobs and axes to operate on one geon at a time, it allows JIM to ignore axes of symmetry that would be formed between geons.

Image Parsing: Grouping Local Image Features Into Geons

Image parsing is a special case of the binding problem in which the task is to group features at different locations in the visual field. For example, given that many local features (edges, vertices, axes, and blobs) are active, how can the model know which belong together as attributes of the same geon (Figure 10)? As solving this problem is a prerequisite to correct geon identification, an image-parsing mechanism must yield groups that are useful for this purpose. The most commonly proposed constraint for grouping—location or proximity (Crick, 1984; Treisman & Gelade, 1980)—is insufficient in this respect. Even

[7] Regions containing geons are filled by an iterative algorithm that activates a point in the visual field if (a) there is an active edge cell at that location or (b) it is surrounded by active points in X or Y. This nonconnectionist algorithm fills regions occupied by geons. These regions are then assumed to correspond to the receptive fields of blob cells in the second layer. The properties of the receptive field (such as area, perimeter, and central point) are calculated directly from the region by counting active points.

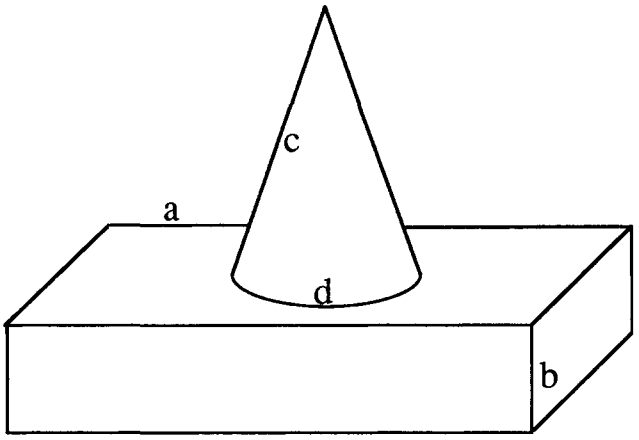


Figure 10. How does the brain determine that Segments a and b belong together as features of the brick, whereas c and d belong together as features of the cone? (Note that proximity alone is not a reliable cue: a is closer to both c and d than to b.)

if there could be an a priori definition of a location (what criterion do we use to decide whether two visual entities are at the same location?), such a scheme would fail when appropriate grouping is inconsistent with proximity, as with Segments a and b in Figure 10. JIM parses images into geons by synchronizing cells in L1 and L2 in a pair-wise fashion according to three simple constraints on shape perception. The mechanism for synchronizing a pair of cells is described first, followed by a discussion of the constraints exploited for grouping.

Synchronizing a Single Pair of Cells

Each edge and vertex cell i is described by four continuous state variables, activation (A_i), output (O_i), output refractory (R_i), and refractory threshold (Θ_i), that vary from zero to one. A cell will generate an output if and only if it is active and its output refractory is below threshold:

$$\text{if and only if } A_i > 0 \quad \text{and} \quad R_i \leq \Theta_i, \quad \text{then} \quad O_i = A_i, \quad \text{otherwise } O_i = 0. \tag{3}$$

When cell i is initially activated, R_i is set to 1.0 and Θ_i is set to a random value between 0 and 1. R_i decays linearly over time:

$$R_i(t) = R_i(t-1) - k, \qquad k \ll 1.0, \tag{4}$$

where t refers to the current time slice. (A time slice is a discrete interval within which the state of the network is updated.) When its refractory reaches threshold ($R_i \leq \Theta_i$), the cell fires ($O_i = A_i$), resets its refractory ($R_i = 1.0$), and re-randomizes its refractory threshold. An active cell in isolation will fire with a mean period of $0.5/k$ time slices[8] and an amplitude[9] of A_i.

Two cells can synchronize their outputs by exchanging an enabling signal over a *fast enabling link* (FEL). FELs are a class of fast, binary links completely independent of the standard connections that propagate excitation and inhibition: Two cells can share a FEL without sharing a standard connection and vice versa. FELs induce synchrony in the following manner: When cell j fires, it propagates not only an output along its connections but also an enabling signal along its FELs. An enabling signal is assumed to traverse a FEL within a small fraction of a time slice. When the enabling signal arrives at an active cell i, it causes i to fire immediately by pushing its refractory below threshold:

$$R_i(t_n) = R_i(t_o) - \Sigma_j FEL_{ij} E_j \tag{5}$$

where $R_i(t_o)$ is the refractory of cell i at the beginning of t, $R_i(t_n)$ is the refractory at some later period within t, FEL_{ij} is the value of the FEL (0 or 1) from j to i, and E_j (the enabling signal from cell j) is 1 if $O_j > 0$, and 0 otherwise. Note that the refractory state of a cell will go below threshold[10]—causing the cell to fire—whenever at least one cell with which it shares a FEL fires. When a cell fires because of an enabling signal, it behaves just as if it had fired on its own: It sends an output down its connections, an enabling signal down its FELs, sets its refractory to 1.0, and randomizes its refractory threshold.

An enabling signal induces firing on the same time slice in which it is generated; that is, its recipient will fire in synchrony with its sender. Furthermore, the synchrony is assumed to be transitive. If Cell A shares a FEL with Cell B and B shares one with C, then C will fire when A fires provided both B and C are active. It is important to note that enabling signals have no effect on, and do not pass through, inactive cells: if B is inactive, then A's firing will have no effect on either B or C. In the current model, the FELs are assumed to have functionally infinite propagation speed, allowing two cells to fire in synchrony regardless of the number of intervening FELs and active cells. Although this assumption is clearly incorrect, it is also much stronger than the computational task of image parsing requires. In the Discussion section, we explore the implications of the temporal parameters of cells and FELs.

The property of transitivity, the absence of propagation through inactive cells, and functionally infinite propagation speed allow us to define a *FEL chain:* Two cells, A and B, are said to lie on the same FEL chain if at least one path can be found from A to B by traversing FELs and active cells. All cells on the same FEL chain will necessarily fire in synchrony. Cells on separate FEL chains will not necessarily fire in synchrony, but they may fire in synchrony by accident (i.e., if their respective output refractories happen to go below threshold at the same time). The possibility of accidental synchrony has important implications for the use of synchrony to perform binding. These implications are addressed in the Discussion section. However, in this section, it is assumed that if two cells lie on different FEL chains, they will not fire in synchrony.

The Set of Fast Enabling Links

Active cells sharing FELs will fire in synchrony. Because our goal is to synchronize local features of the same geon while

[8] The mean refractory threshold for a cell (Θ_i) will be 0.5. Assuming that $\Theta_i = 0.5$, it will take $0.5/k$ time slices for R_i to decay from 1.0 to Θ_i.

[9] Real neurons spike with an approximately constant amplitude. A cell's firing can be thought of as a burst of spikes, with amplitude of firing proportional to the number of spikes in the burst.

[10] $\Sigma_j FEL_{ij} \geq 1.0 \geq R_i(t_o)$. Therefore $[R_i(t_o) - \Sigma_j FEL_{ij}] = R_i(t_n) \leq 0 \leq \Theta_i$.

keeping separate geons out of synchrony, a FEL should only connect two cells if the features they represent are likely to belong to the same geon. A pair of cells will share a FEL if and only if that pair satisfies one of the following three conditions:

Condition I: Local coarse coding of image contours. This condition is satisfied if both cells represent image edges of the same curvature and approximately the same orientation and have overlapping receptive fields. As depicted in Figure 11, a single image contour (or edge) will tend to activate several cells in L1. All the cells activated by a single contour will typically belong to the same geon[11] and should therefore be grouped. Locally, such cells will tend to satisfy Condition I. The model groups the local pieces of a contour (i.e., groups edge cells responding to the same contour) using FELs between all pairs of L1 cells with similar orientation and curvature preferences and overlapping receptive fields (Figure 12). Note that not all cells responding to a given contour will necessarily satisfy all these criteria; for example, the receptive fields of two cells at opposite ends of a long contour might not overlap. However, by virtue of the intervening units, they will lie on the same FEL chain. Indirect phase locking using long FEL chains does not pose a problem except insofar as the propagation time for an enabling signal from A to B increases with the number of FELs to be traversed. An important issue for exploration concerns how the synchrony for such distant units will generalize with more realistic assumptions about propagation speeds for FELs.

The FELs corresponding to Condition I can be derived strictly from the statistical properties of coarsely coded image edges. If two cells A and B satisfy Condition I, the conditional probability that A will be active given that B is active [p(A|B)] should (a) be much greater than the base probability that A is active [P(A|B) > p(A)] and (b) be approximately equal to P(B|A). In the case of the present model, the only cells that satisfy both criteria are L1 edge cells with overlapping receptive fields, identical curvature preferences, and similar orientation preferences. In the general case, two cells will tend to satisfy both criteria whenever they code overlapping regions of the same attribute space. For example, if our representation of edges coded *degree* of curvature (rather than simply coding an edge discretely as either *straight* or *curved*), then Condition I would have to be modified by adding "and have similar curvature preferences." Because the FELs corresponding to Condition I connect cells whose activity should covary, they should be capable of self-organizing with a simple Hebbian learning rule and a stimulus set consisting of contours longer than the diameter of the cells' receptive fields.

We ran a small simulation that corroborated this conjecture. A 12-row × 10-column hexagonal lattice of segment- and termination-sensitive cells was exposed to random triangles (Figure 13), and FELs were updated using a modified Hebbian learning rule. Edges were coded coarsely, with respect to both location and orientation. Termination cells in this model did not inhibit consistent segment cells. Because of the small size of this simulation, it was not necessary to make all the simplifying assumptions required by JIM. Specifically, this model differed from JIM in its more realistic hexagonal lattice of edge cells, its coarse coding of orientation, and the fact that terminations did not inhibit same-orientation segments within a cluster. Initially, all cells were connected by FELs of strength zero to all other

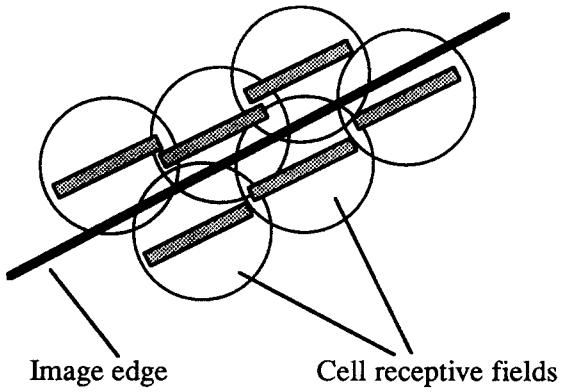


Figure 11. A single image edge will tend to activate several cells in Layer 1. (Locally, these cells will tend to have similar orientation preferences and overlapping receptive fields.)

cells in the same or adjacent locations (see Figure 13). The simulation was run for 200 iterations. On each iteration, a randomly generated triangle was presented to the model. Cell activations were determined by Equations 1 and 2. FELs were updated by the Hebbian rule:

$$\text{if } A_i + A_j > 0, \text{ and } \begin{cases} \text{if } A_iA_j > 0, & \text{then } \Delta FEL_{ij} = \nu A_iA_j(1 - |FEL_{ij}|) \\ \text{if } A_iA_j = 0, & \text{then } \Delta FEL_{ij} = -\nu\tau(A_i + A_j)(1 - |FEL_{ij}|), \end{cases}$$

$$\text{if } A_i + A_j = 0, \text{ then } \Delta FEL_{ij} = 0, \qquad (6)$$

where ν is a learning rate parameter, and τ determines the rate of growth toward negative values relative to the growth toward positive values. With this learning rule, FELs from active cells to other active cells grow more positive, FELs from active cells to inactive cells grow more negative, and FELs between pairs of inactive cells do not change. By the end of the 200 iterations, strong positive FELs had developed between cells with overlapping receptive fields and identical or adjacent orientation preferences. All other potential FELs were negative or zero.

Condition II: Cotermination in an intrageon vertex. This condition is satisfied if one cell represents a termination and the other represents a consistent vertex or lone termination at the same location. Image contours that coterminate in a two- or three-pronged vertex likely belong to the same geon. The model groups contours into geons by positing FELs between termination cells in L1 and cells representing consistent two- and three-pronged vertices in L2. Recall that by Condition I, an L1 termination cell will fire in phase with the rest of the cells representing the same contour. If at least one—but not more than two—other termination cells are active at the same location (reflecting the cotermination of two or three contours in an

[11] There are important exceptions to this generalization as evident in Figure 19.

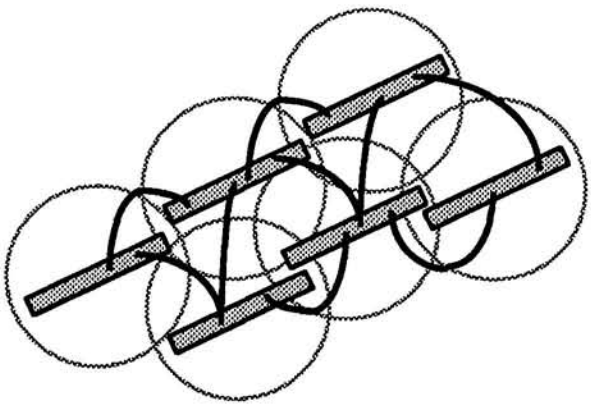


Figure 12. Fast enabling links (FELs) corresponding to Condition I (local coarse coding of image contours) connect cells in Layer 1 that have similar orientation preferences and overlapping receptive fields. (FELs are indicated by arcs in the figure.)

intrageon vertex), then a vertex cell will be activated in L2 (Figure 14). By virtue of the termination-to-vertex FELs, each L1 termination cell will fire in phase with the corresponding L2 vertex cell, and by transitivity, all the terminations and their respective contours will fire in phase with one another. In this manner, FELs corresponding to Condition II group separate contours into geons (Figure 14).

In contrast with vertices produced by cotermination, T vertices are produced where one surface occludes another, the top belonging to the occluding surface and the stem to the occluded surface. Therefore, the parts of a T vertex should not be grouped. In JIM's first layer, a T vertex is represented by an active termination cell of one orientation (the stem) and an active segment cell of a different orientation (the top). In L2, a T vertex is represented only as an active lone termination with the orientation and curvature of the *T*'s stem (recall that L1 edge cells inhibit all L2 vertex cells except lone terminations). Lone terminations share FELs with L1 termination cells, but not with L1 segment cells. Therefore, the contours forming the top and stem of a T vertex are not synchronized, allowing geons that meet at T vertices to remain out of synchrony with one another.

Condition III: Distant collinearity through lone terminations. Both cells represent lone terminations, their orientations are complementary, and they are collinear. Although the separate parts of a T vertex (viz., stem and top) will not belong to the same geon (except in some accidental views), the stems of separate *T*s may. When an edge in the world projects as two separate image contours because of occlusion, the endpoints of those contours will activate lone termination cells at the points of occlusion (Figure 15a). If the edge is straight, the lone terminations will be collinear and have complementary orientations. Here, *complementary* has the specific meaning that (a) the orientations differ by 180° (± some error, e) and (b) the orientation of

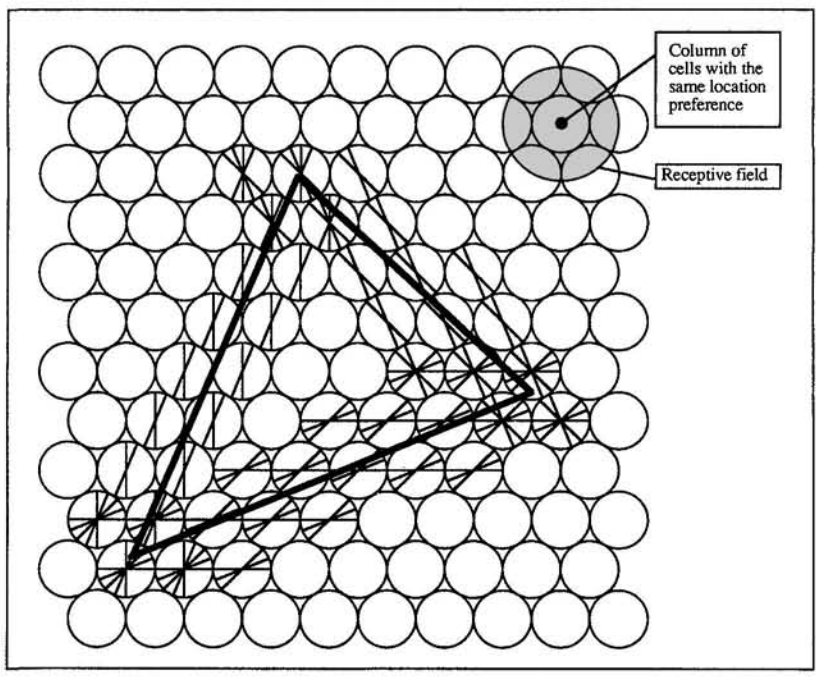


Figure 13. FELs were allowed to self organize in response to images of random triangles as illustrated in the figure. (In this simulation, cells were arrayed in a hexagonal lattice with overlapping receptive fields and orientation preferences. Thick lines in the figure indicate image edges, thin lines indicate active cells [the location and orientation of a line corresponding to the preferred location and orientation of the cell, respectively], and circles indicate locations over which cell receptive fields were centered. The degree of overlap between cell receptive fields is shown in the upper right corner of the figure. L = layer.)

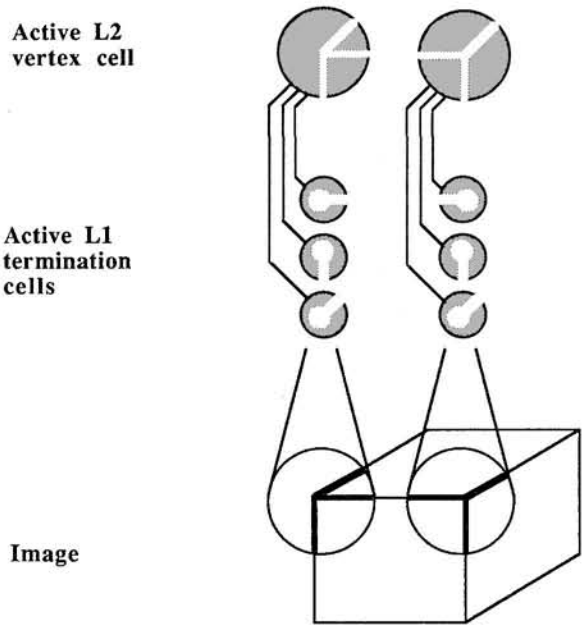


Figure 14. If two or more image contours coterminate at a point, multiple termination cells (one for the end of each contour) will be activated in the corresponding location in Layer (L) 1. (As long as fewer than three L1 termination cells are active, a vertex cell will become active at the same location in L2. L2 vertex cells share fast enabling links [FELs; indicated by arcs] with all corresponding L1 termination cells.)

a vector extending from either termination point to the other differs by 180° ± e from the direction in which the contour extends from that point (Figure 15b). The model groups the separate parts of an occluded edge using FELs between all pairs of distant lone termination cells with collinear receptive fields and complementary orientation preferences.

It is important to stress that only L2 lone termination cells share these distant FELs. Because lone terminations are inhibited by inconsistent L1 termination cells, multiple contours coterminating at a point will prevent distant collinear grouping. For example, Contours 1 and 2 in Figure 16a will group because lone terminations are activated at the occlusion points, but the L vertices in Figure 16b will prevent the same contours from grouping because the extra terminations (1′ and 2′) inhibit the lone termination cells. This property of distant collinear grouping only through lone terminations has the important consequence that separate geons will not group just because they happen to have collinear edges.

Fast Enabling Links: Summary and Extensions

Recall that cells on the same FEL chain will fire in synchrony, and barring accidental synchrony, cells on separate chains will fire out of synchrony. The FELs posited for this model will parse an image into its constituent geons provided (1) within geons, features lie on the same FEL chain and (2) features of separate geons lie on separate FEL chains. Provision 1 will be true for all geons composed of two- or three-pronged vertices (Figure 17) unless (a) enough vertices are occluded or deleted to leave the remaining vertices on separate FEL chains (Figure 18a) or (b) contour is deleted at midsegment, and additional contour is added that inhibits the resulting lone terminations (Figure 18b). Provision 2 will be true for all pairs of geons unless (a) they share a two- or three-pronged vertex (Figure 19a), (b) they share a contour (Figure 19b), or (c) they share a pair of complementary, collinear lone terminations (Figure 19c).

Conditions I, II, and III constitute a simplified set of constraints that can easily be generalized. Currently, curved edges are grouped on the basis of the same constraints that group straight edges (Conditions I and III), except that a greater difference in orientation is tolerated in grouping curved edge cells than in grouping straight edge cells. These conditions could be generalized by replacing the term *collinear* with the more general term *cocircular.* Two edge segments are *cocircular* if and only if both can be tangents to the same circle (collinearity is a special case of cocircularity with circles of infinite radius). If direction and degree of curvature were made explicit in the representation of edges, Condition I could be generalized to use this information. This generalization would also allow us to exploit an additional constraint on grouping: Matched pairs of coterminating edges with curvatures of opposite signs ("cusps") are formed when two convex volumes interpenetrate. This constraint, termed the *transversality regularity,* was introduced by Hoffman and Richards (1985) as an explicit cue to parsing. Like the top and stem of a T vertex, terminations forming cusps would not be linked by FELs, thereby implicitly implementing the transversality regularity.[12]

Finally, we should note that the use of collinearity and vertices to group image edges into parts is not unique to the current proposal. A number of computer vision models, for example, Guzman (1971), Waltz (1975), and Malik (1987), group and label edges in line drawings. Line-labeling models operate by the propagation of symbolic constraints among data structures representing the local features (i.e., edges and vertices) in a line drawing. As output, these models produce not an object classification, but a representation in which the 3-D nature of each local feature is specified (e.g., as convex, concave, occluding, or shadow). For example, these models can detect when line drawings represent impossible objects. By contrast, human observers are slow at detecting the impossibility of an object. The current proposal for grouping differs from line labeling in that (a) it is concerned only with the grouping of local image features, not their labeling; (b) it explicitly seeks to derive the grouping in a neurally plausible manner; and (c) once grouped, the features are used to derive a parts-based structural description that is subsequently used for object classification. It remains to be seen what role line labeling might play in object recognition versus, say, depth interpretation of surfaces.

Layer 3: Geons

The model's first two layers represent the local features of an image and parse those features into temporal packages corre-

[12] Parsing would only occur at matched cusps because a figure with only one cusp—such as a kidney—would allow enabling signals to pass, not through the cusp but around the back, along the smoothly curved edge.

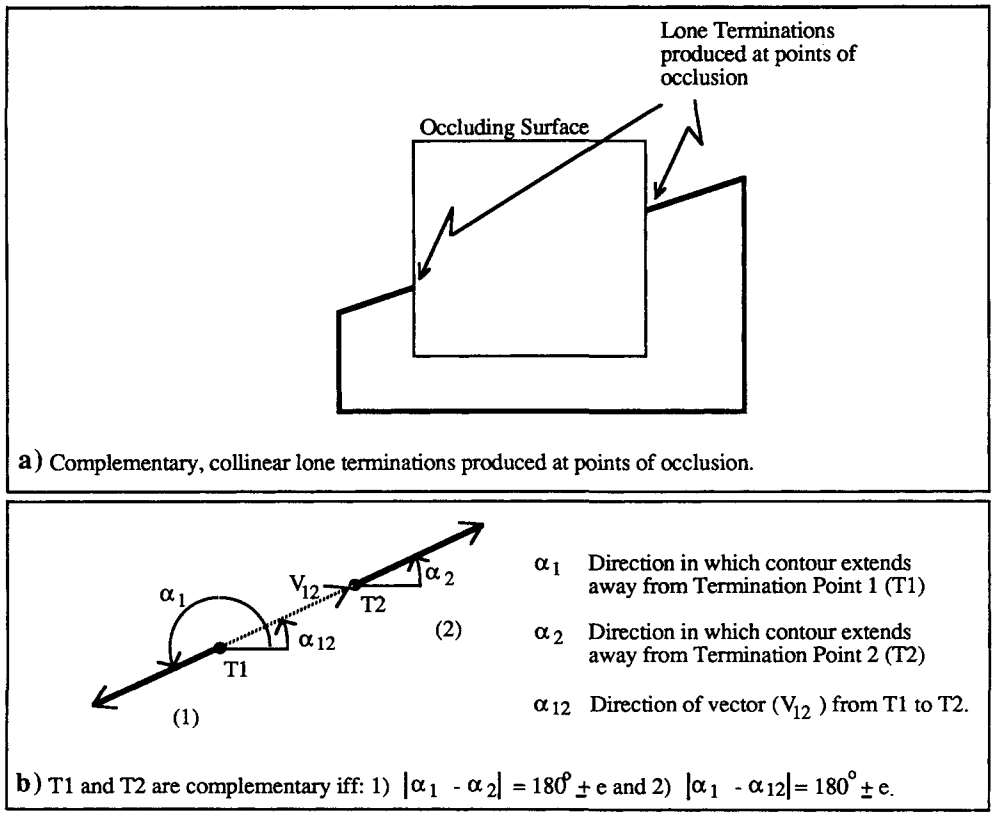


Figure 15. a: Complementary, collinear lone terminations are produced when a single straight edge in the world is occluded by a surface. b: Illustration of the definition of *complementary lone terminations.*

sponding to geons. The third layer uses these grouped features to determine the attributes of the geons in the image. Each cell in L3 responds to a single geon attribute and will respond to any geon that possesses that attribute, regardless of the geon's other properties. For example, the cell that responds to geons with curved axes will respond to a large, curved brick in the upper left of the visual field, a small, curved cone in the lower right, and so forth. Because the geons' attributes are represented independently, an extraordinarily small number of units (58) is sufficient to represent the model's universe of geons and relations. The binding of attributes into geons is achieved by the temporal synchrony established in the first and second layers. It is at this level of the model that viewpoint invariance is achieved, and the first elements of a structural description are generated.

Representation of Geons

Cells in the model's third layer receive their inputs from the vertex, axis, and blob cells of L2. The details of the L2 to L3 mapping are described after the representation of L3 is described. Layer 3 consists of 51 cells that represent geons in terms of the following eight attributes (Figure 7):

Shape of the major axis. A geon's major axis is classified either as straight (as the axis of a cone) or curved (like a horn). This classification is *contrastive* in that degrees of curvature are not discriminated. One L3 cell codes for straight axes and one for curved.

Shape of the cross section. The shape of a geon's cross section is also classified as either straight (as the cross section of a brick or wedge) or curved (like a cylinder or cone). This attribute is contrastive, and the model does not discriminate different shapes within these categories: A geon with a triangular cross section would be classified as equivalent to a geon with a square or hexagonal cross section. One L3 cell codes for straight cross sections and one for curved.

Parallel versus nonparallel sides. Geons are classified according to whether they have parallel sides, such as cylinders and bricks, or nonparallel sides, like cones and wedges. This attribute is also contrastively coded in two cells: one for parallel and one for nonparallel.

Together, these three attributes constitute a distributed representation capable of specifying eight classes of geons: *Brick* (straight cross section, straight axis, and parallel sides), *Cylinder* (curved cross section, straight axis, and parallel sides), *Wedge* (straight cross section, straight axis, and nonparallel sides), *Cone* (curved cross section, straight axis, and nonparallel sides), and their curved-axis counterparts. Contrasts included in Biederman's (1987b) RBC theory that are not discriminated by JIM include (a) whether a geon with nonparallel sides contracts to a point (as a cone) or is truncated (as a lamp shade), (b) whether the cross section of a geon with nonparallel sides both

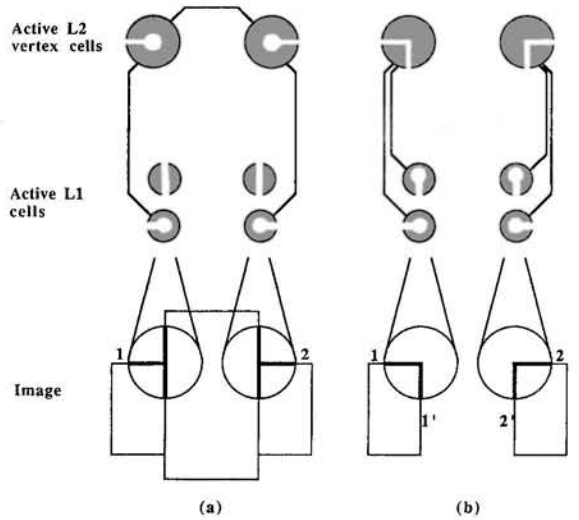


Figure 16. a: Contours 1 and 2 will group because lone terminations are activated at the occlusion points. b: Contours 1′ and 2′ produce L vertices that prevent the same contours (1 and 2) from grouping. The extra terminations (1′ and 2′) inhibit the lone termination cells. L2 and L1 = Layer 2 and Layer 1.

expands and contracts as it is propagated along the geon's axis (as the cross section of a football) or only expands (as the cross section of a cone), and (c) whether the cross section is symmetrical or asymmetrical.

Aspect ratio. A geon's aspect ratio is the ratio of the diameter of its cross section to the length of its major axis. The model codes five categories of aspect ratio: approximately 3 or more to 1 (3+:1), 2:1, 1:1, 1:2, and 1:3+. These categories are coded coarsely in one cell per category: Each aspect ratio cell responds to a range of aspect ratios surrounding its preferred value, and cells with adjacent preferred values respond to overlapping ranges.

Coarse orientation. A geon's orientation is represented in two separate banks of cells in L3: a *coarse* bank, used directly for recognition (i.e., the outputs go to L6), and a *fine* bank, described later, used to determine the orientation of one geon relative to another (the outputs go to L4). The coarse bank consists of three cells, one each for horizontal, diagonal, and vertical orientations (Figure 7). The coarse orientation cells pass activation to L6 because the orientation classes for which they are selective are diagnostic for object classification, as with the difference between the vertical cylinder in a coffee mug and the horizontal cylinder in a klieg light. However, finer distinctions, such as left-pointing horizontal versus right-pointing horizontal, typically do not distinguish among basic level classes. A klieg light is a klieg light regardless of whether it is pointing left or right.

Fine orientation. The coarse representation of orientation is not precise enough to serve as a basis for determining the relative orientation of two geons. For example, two geons could be perpendicular to one another and be categorized with the same orientation in the coarse representation (e.g., both legs of a T square lying on a horizontal surface would activate the coarse horizontal cell). The more precise fine representation is used to determine relative orientation. The fine cells code seven orientation classes (Figure 20): two diagonal orientations (left end up and right end up), four horizontal orientations (perpendicular to the line of sight, left end closer to viewer, right end closer, and end toward viewer), and one vertical orientation. Each orientation is represented by one cell.

Size. A geon's size is coded coarsely in 10 cells according to the proportion of the visual field it occupies. The activation (A_i) of a size cell in response to a geon is given by

$$A_i = (1 - |C_i - G|/W_i)^+, \tag{7}$$

where C_i is the preferred size of cell i, W_i is the width of the cell's receptive field (0.1 for size cells), and G is the proportion of the visual field occupied by the geon. The preferred sizes of the L3 size cells start at 0.0 and advance in increments of 0.1 up to 0.9.

Location in the visual field. A geon's location in the visual field is defined as the position of its centroid (the mean x- and y-coordinates for the set of all points inside the geon). The horizontal and vertical components of a geon's position are coded independently and coarsely in 10 cells each. The activation of a location cell is given by Equation 7, where C_i corresponds to the cell's preferred position. Location cells are ordered by positions starting at the left and bottom edges of the visual field and are incremented in equal intervals to the right and top edges, respectively. For example, the cell for $x = 1$ (far left) responds when a geon's centroid is close to the left edge of the visual field, $y = 1$ responds to centroids near the bottom edge, and $x = 5$ responds to centroids just to the left of the vertical midline of the visual field.

Activating the Geon Attribute Cells

Cells in the model's third layer receive their inputs from the vertex, axis, and blob cells of L2, but the second and third layers

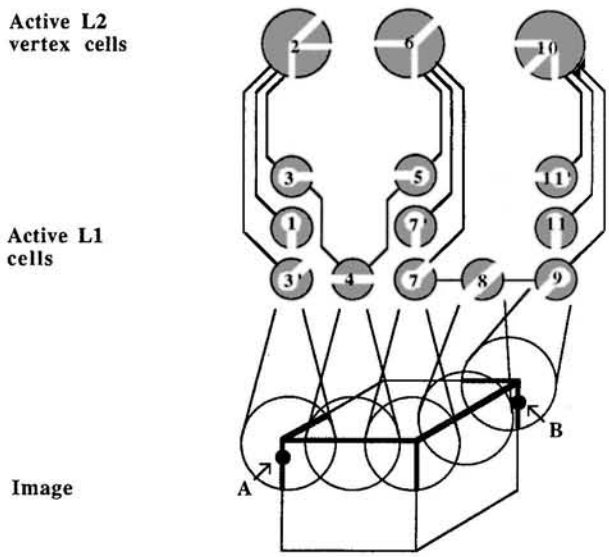


Figure 17. Illustration of grouping features into geons using fast enabling links (FEL) chains. (Consider Points a and b on the brick. When the cell representing Point a [Cell 1] fires, it will cause Cell 2 to fire, which will cause Cell 3 to fire, and so on, until this process has reached Point b [Cell 11]. L = layer.)

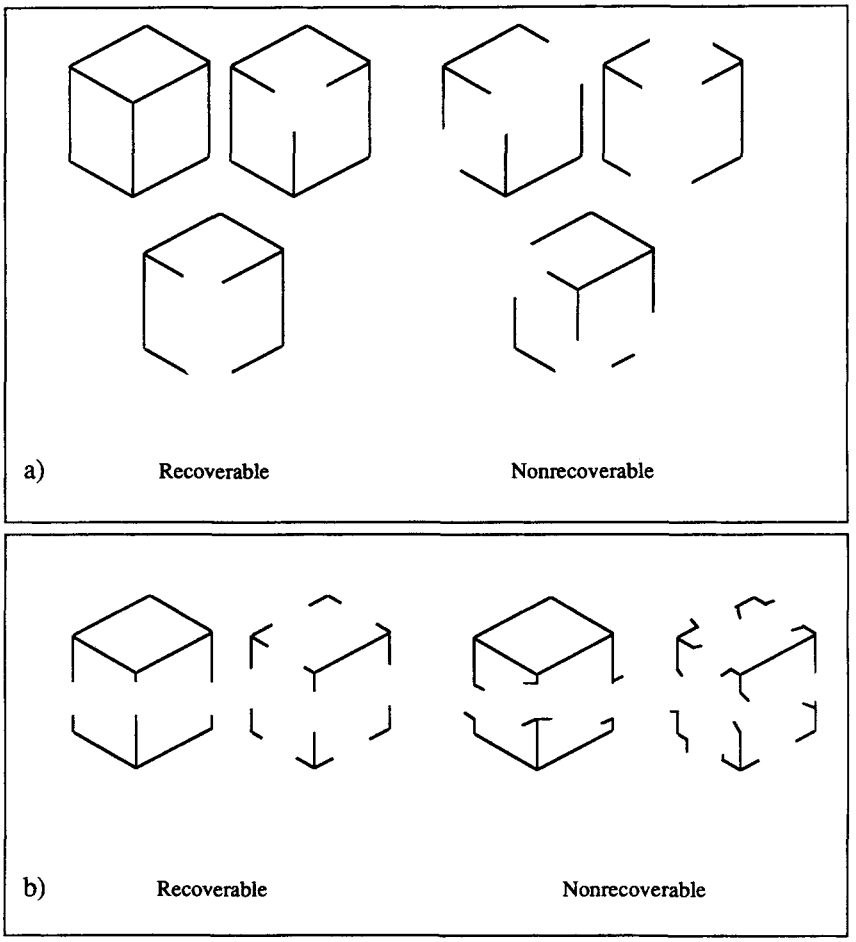


Figure 18. Conditions under which the proposed FELs will group the local features of a geon (recoverable) or fail to group local features into geons (nonrecoverable). (a: Features of the nonrecoverable geons will lie on separate FEL chains because of deletion of vertices. b: Features of the nonrecoverable geons will lie on separate FEL chains because of extraneous vertices introduced where the original contour has been deleted.)

of the model are not fully interconnected. Rather, L3 cells receive bottom-up excitation from consistent L2 cells only, and L3 cells representing inconsistent hypotheses (e.g., curved axis and straight axis) are mutually inhibitory. For example, L2 vertex cells send their outputs to the L3 cells for straight and curved cross section, but neither excite nor inhibit the L3 size cells. With the exception of cells for size, location, and aspect ratio, L3 cells compute their activations and outputs by Equation 2. The lateral inhibition among inconsistent L3 cells is implemented by normalizing the net inputs (N_i) to the L3 cells by the equation

$$N_i = N_i^k/(\Sigma_j N_j^k + N_i^k), \qquad k > 1, \tag{8}$$

for all j such that cell j inhibits cell i.

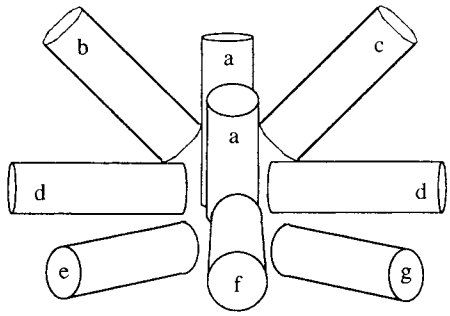
The remainder of this section describes the pattern of connectivity from L2 to L3. In the discussion that follows, two conventions are adopted. First, if a class of L2 cells sends no output to a class of L3 cells, they are not mentioned in the discussion of that L3 attribute (e.g., vertex cells are not mentioned in the discussion of the L3 size cells). The second convention stems from those cases where a given L3 attribute receives input from a class of L2 cells, but is insensitive to some of the dimensions for which those cells code. For example, the L3 cells for size receive input from the L2 blob cells but are not sensitive to the locations, orientations, or elongations of those blobs. In such cases, the irrelevant dimensions are not mentioned, and the L3 cells compute their inputs by summing the outputs of L2 cells over the irrelevant dimensions (e.g., the L3 size cells sum the outputs of L2 blob cells over all orientations, locations, and elongations).

Axis shape and parallel versus nonparallel sides. Both the shape of a geon's major axis and whether its sides are parallel or nonparallel are determined on the basis of the 2-D axes in the image. L2 cells representing axes of nonparallel symmetry excite the L3 cell for nonparallel sides, and cells for axes of parallelism excite the L3 cell for parallel sides. L2 cells representing curved axes excite the L3 cell for curved axis, and cells repre-

senting straight axes excite L3's straight axis cell. The parallel-sides and non-parallel-sides cells are mutually inhibitory, as are the straight-axis and curved-axis cells.

Cross section. Whether a geon's cross section is straight or curved is determined on the basis of the vertices active in L2. Tangent-Y vertices and L vertices with at least one curved leg excite the L3 curved-cross-section cell. Forks and arrows excite the straight cross section cell. Straight cross section and curved cross section cells are mutually inhibitory. This mapping is summarized in Figure 21.

Size and location. L3 cells representing geon size and location take their input from the blobs derived in L2. Recall that a blob represents a region of the visual field with a particular location, size, orientation, and elongation. It is assumed that each blob cell excites a set of L3 size and location cells consistent with its receptive field properties. The receptive field properties of the size and location cells were described earlier; their inputs (G in Equation 7) come from the blob computations. It is assumed that the term $(1 - |C_i - G|/W_i)^+$ is equal to the connection weight to an L3 size or location cell from an active blob cell. For example, as shown in Figure 22, a given blob might respond to a region of activity just left and above the middle of the visual field, elongated slightly, oriented at or around $\Pi/4$ (45°) and occupying between 10% and 15% of the visual field. In the L3 location bank, this blob will strongly excite $x = 5$ and more weakly excite $x = 4$, and it will strongly excite $y = 5$ and weakly excite $y = 6$. In the L3 size bank, it will excite the cells for 10% and 15% of the visual field. Figure 23 shows the response of L3 size and location cells as a function of a geon's size and location.



Fine Orientations	Coarse Orientations
a. Vertical	Vertical
b. Diagonal, left end up c. Diagonal, right end up	Diagonal
d. Horizontal e. Horizontal, left end closer to viewer f. Horizontal, end toward viewer g. Horizontal, right end closer to viewer	Horizontal

Figure 20. The fine orientation cells code seven geon orientations. (The coarse orientation cells code only three. Each fine orientation corresponds to exactly one coarse orientation.)

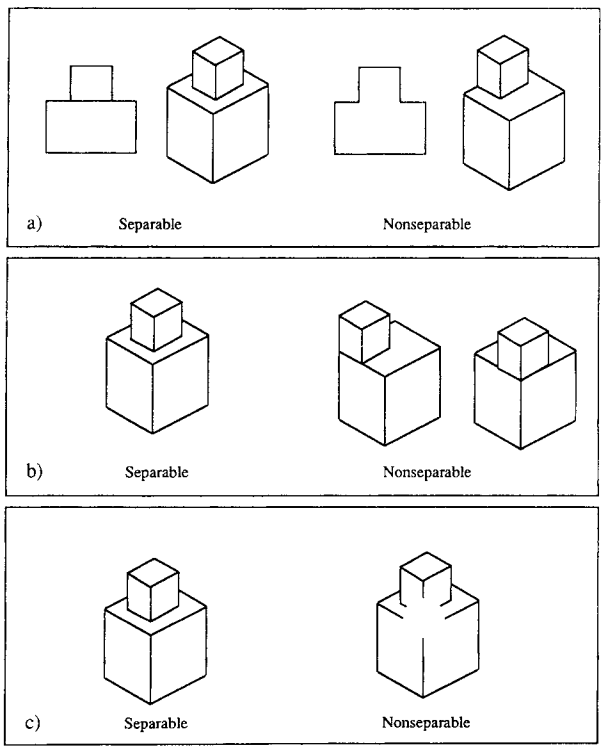


Figure 19. Conditions under which the proposed fast enabling links (FELs) will separate the parts of an object (separable) or fail to separate the parts (nonseparable). (a: The nonseparable geons share a two- or three-pronged vertex. b: The nonseparable geons share a contour. c: The nonseparable geons share a pair of complementary, collinear lone terminations [along the central vertical edge].)

Aspect ratio. L3 aspect ratio cells calculate their activation by Equation 7, where C_i is the cell's preferred aspect ratio, and W_i is the width of the cell's receptive field. The receptive field properties of these cells are illustrated in Figure 23c. The elongation of the L2 blobs supplies the value of a geon's aspect ratio (G in Equation 7). However, as illustrated in Figure 24a, a blob with a given elongation and orientation (e.g., minor axis = 1 and major axis = 2 at orientation = 0 radian) could arise either from an elongated geon at the same orientation (cross-section width = 1 and axis length = 2 at 0 radian) or from a flattened geon at the perpendicular orientation (2:1 at $\Pi/2$ radians). The model resolves this ambiguity by comparing the blob's orientation to the orientation of the geon's longest 2-D axis. If the orientations are approximately equal, then the longer of the two aspect ratios is chosen (e.g., 1:2); if they are approximately orthogonal, then the shorter aspect ratio is chosen (2:1).

In the simulations reported here, these comparisons are made directly on the basis of the data structures describing the axes and blobs in an object's image. However, the comparisons could easily be performed by an intermediate layer of cells as follows (see Figure 24b): Each cell in the intermediate layer (labeled *Layer 2.5 Orientation × Aspect Ratio Cells*) represents a conjunction of aspect ratio and 2-D orientation[13] and receives

[13] Only one L2.5 unit need exist for Aspect Ratio 1:1 because such blobs, being round, have no 2-D orientation.

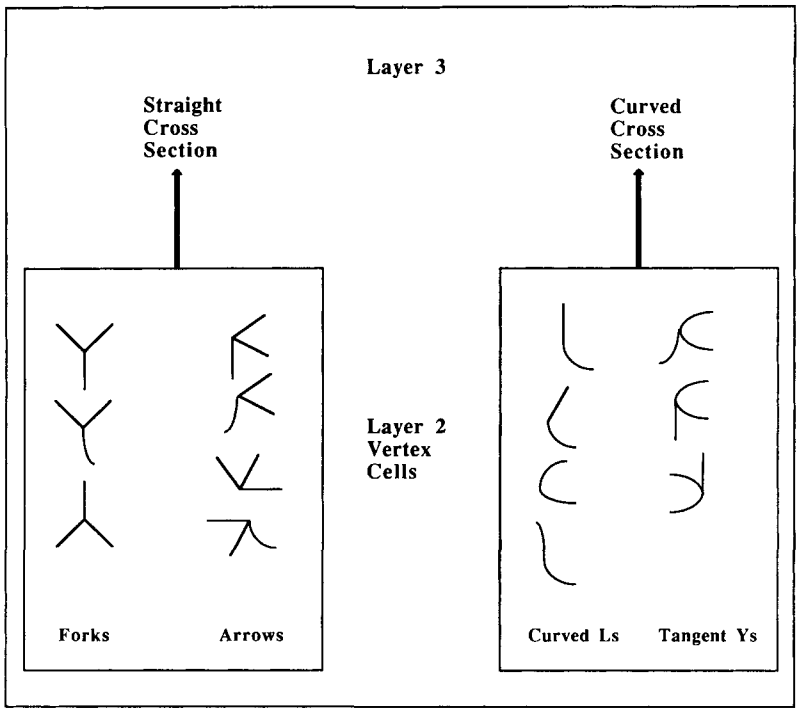


Figure 21. Forks and arrows excite the straight cross section cell. (Tangent-Y vertices and L vertices with at least one curved leg excite the Layer 3 curved-cross-section cell. Straight cross section and curved cross section cells are mutually inhibitory.)

inputs from both axis and blob cells. Each blob cell excites two L2.5 units, one with the same orientation and aspect ratio and one with the perpendicular orientation and reciprocal aspect ratio. Each axis cell excites all L2.5 cells consistent with its orientation. The correct aspect ratio can be chosen by computing the net input to each L2.5 cell as the product of its axis and blob inputs. The L2.5 cell outputs can then be summed over orientation to determine the inputs to the L3 aspect ratio cells. For example, assume that the elongated horizontal cylinder on the right of Figure 24a was presented to the network. It would activate the blob and axis cells shown in Figure 24 (a and b, respectively). The blob would excite two cells in L2.5 (vertically hatched in Figure 24b): 2:1 at Orientation 4 and 1:2 at Orientation 0. The axis cells would excite all L2.5 cells for Orientation 0 (horizontally hatched cells). The only L2.5 cell receiving inputs from both an axis and a blob would be the one for 1:2 at Orientation 0. This cell would excite the L3 aspect ratio cell for 1:2, which is the correct aspect ratio for that geon.

Orientation (coarse and fine). L3 fine orientation cells receive input from the vertex and axis cells in L2. In practice, the coarse orientation cells receive their input from the corresponding fine cells (Figure 20), but in principle, they could just as easily determine their inputs on the basis of the L2 axis and vertex cells. Both axes and vertices are used to determine orientation because neither is sufficient alone. Consider a geon with a straight axis (of parallelism or symmetry) that is vertical in the visual plane (Figure 25). On the basis of the orientation of the axis alone, it is impossible to determine whether the geon is oriented vertically or horizontally, with its end toward the viewer. Cells representing vertical axes therefore excite both the L3 vertical and horizontal-end-on cells. The resulting ambiguity is resolved by the geon's vertices. If the geon is standing vertically, the straight legs of its tangent-Y vertices will extend away from the points where they terminate with an angle greater than Π and less than 2Π. Therefore, all cells representing tangent-Y vertices with straight legs extending between Π and 2Π excite the L3 cell for vertical. All those with straight legs extending between 0 and Π excite horizontal-end-on. This mapping is summarized in Figure 26. The orientation with the most bottom-up support is selected by inhibitory interactions (Equation 8) among the fine orientation cells.

Summary of Layers 1–3

The model's first three layers parse an image into its constituent geons and activate independent representations of each of the geon's attributes. For example, given the image in Figure 1, the model will parse the local features of the cone and the brick into separate groups, features within a group firing in synchrony. Each group then activates the cells in L3 that describe the geon they comprise. L3 cells are temporally yoked to their inputs: They fire only on time slices in which they receive inputs. When the L2 cells representing the cone fire, the L3 cells for straight axis, curved cross section, nonparallel sides, vertical, and the cells for its aspect ratio, size, and location become active and fire. Likewise, when the L2 cells representing the brick fire, the L3 cells for straight axis, straight cross section, parallel sides, horizontal, and the cells for its aspect ratio, size, and location will fire.

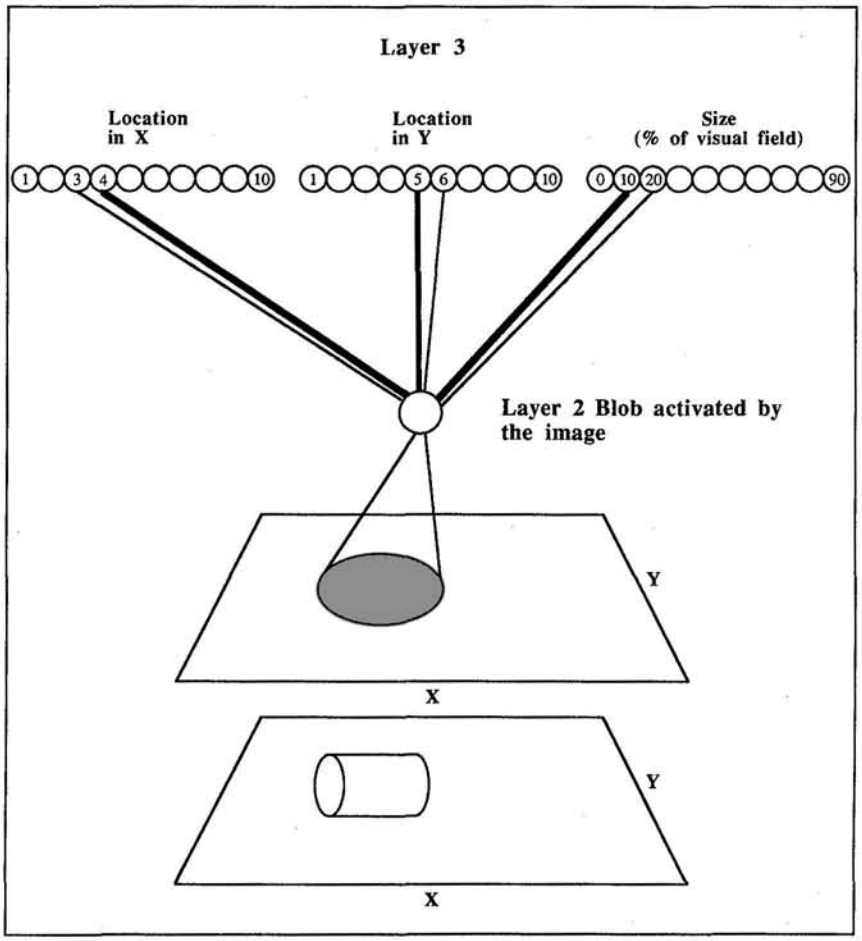


Figure 22. Layer 2 blobs activate Layer 3 cells representing a geon's size and location in the visual field.

Layers 4 and 5: Relations Among Geons

Of the eight attributes represented in L3, five—axis shape, cross-section shape, parallelism of sides, coarse orientation, and aspect ratio—pass activation directly to the model's sixth layer (Figure 7). The remaining attributes—size, location, and fine orientation—pass activation to Layer 4, which, in conjunction with Layer 5, derives relative size, relative location, and relative orientation. The computational goals of Layers 4 and 5 are threefold. First, the relations among geons must be made explicit. For example, rather than representing that one geon is below another implicitly by representing each of their locations, the *below* relation is made explicit by activating a unit that corresponds uniquely to it. Second, the relations must be bound to the geons they describe. If one geon is below another in the visual field, the unit for below must be synchronized with the other units describing that geon. Finally, the relations must be invariant with geon identity and viewpoint so that, for example, the below unit will fire whenever one geon is below another, regardless of the shape of the geons, their locations in the visual field, orientation in depth, and size.

These goals are satisfied in two steps. In the first, L4 cells act as AND gates, responding when conjunctions of L3 cells fire on different (but nearby) time slices. In the second step, L5 cells OR together the outputs of multiple L4 cells, responding to the relations in a viewpoint-invariant manner. As illustrated in Figure 27, each L4 cell responds to a specific relation (e.g., below) in conjunction with a specific value on the dimension over which that relation is defined (e.g., $y = 1$). L4 cells respond to the following types of conjunctions: above–below conjoined with position in y, right–left with position in x, larger–smaller with size, and perpendicular–oblique with orientation. L5 contains only one cell for each relation: above, below, beside (which replaces the left–right distinction), larger, smaller, perpendicular, and oblique.

L4 cells receive both excitatory inputs and enabling signals from L3 (Figure 28). An L3 cell will excite an L4 cell if the L4 cell's value satisfies its relation with respect to the L3 cell. For example, the L3 $y = 3$ cell excites the L4 cell for below$|y = 1$, because $y = 1$ is below $y = 3$ ($y = 1$ satisfies below with respect to $y = 3$). $y = 3$ also excites above$|y = 5$, above$|y = 6$, and so forth. Excitatory connections from L3 to L4 are unit strength, and there are no inhibitory connections. L4 cells sum their inputs over time by the equation

$$\Delta A_i = \gamma E_i(1 - A_i) - \delta A_i, \tag{9}$$

where A_i is the activation of L4 cell i, E_i is its excitatory input, and γ and δ are growth and decay parameters, respectively. L3 cells send enabling signals to L4 cells that respond to the same

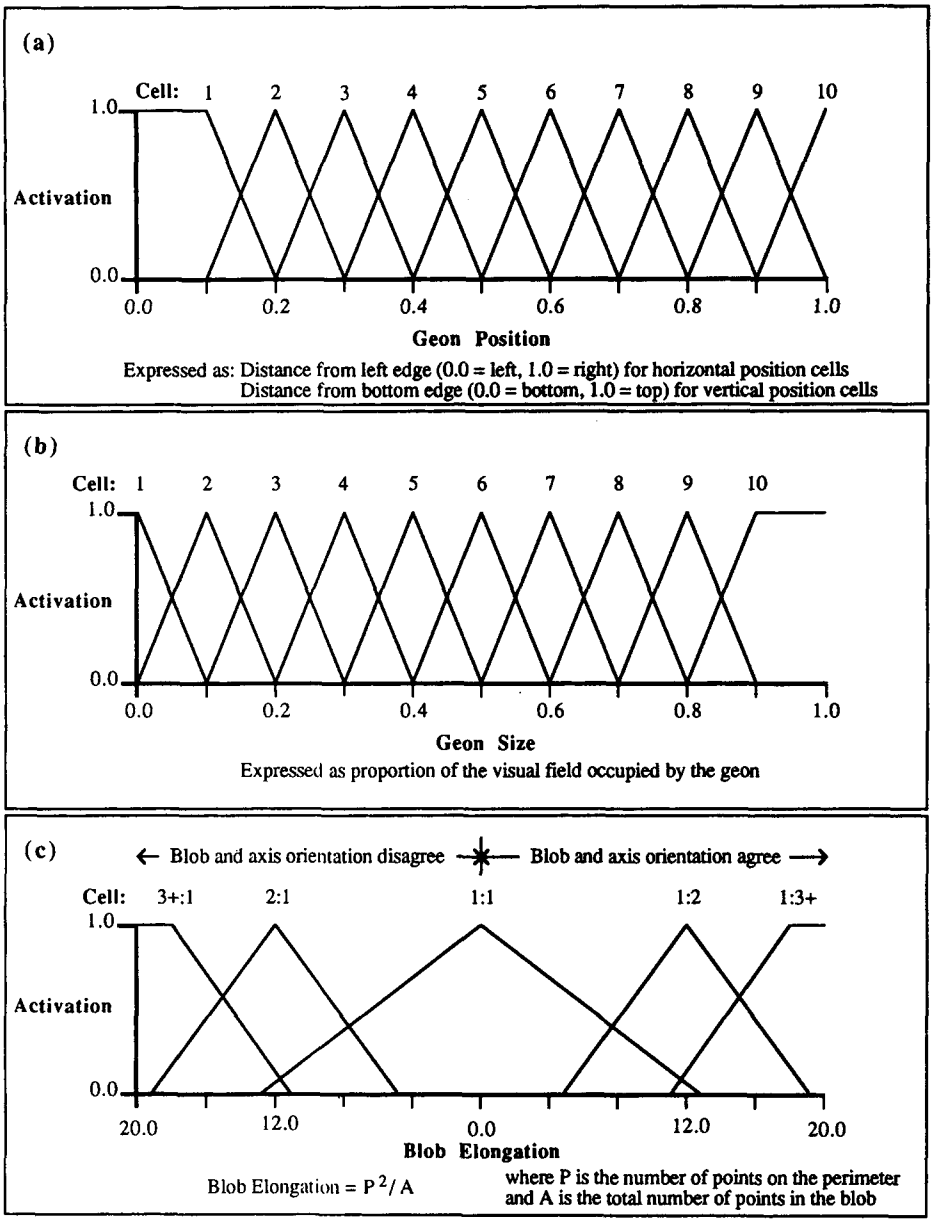


Figure 23. Receptive field properties of Layer 3 (a) location, (b) size, and (c) aspect ratio cells.

value; for example, $y = 1$ sends enabling signals to below|$y = 1$. L4 cells ensure proper geon-relation binding by firing only when they receive enabling signals. The invariant L5 relation cells sum the outputs of the corresponding L4 cells. For example, the L5 below cell receives excitation from below|$y = 1$, below|$y = 2$, and so on.

The relation *below* is used to illustrate how this architecture activates and binds invariant relations to geons, but the mechanism works in exactly the same way for all relations. Suppose, as shown in Figure 28, there is a geon near the bottom of the visual field (Geon A) and another nearer the middle (Geon B). $y = 1$ and $y = 2$ will fire in synchrony with the other L3 units describing Geon A, say, on time slices 1, 5, and 9. Similarly, $y = 2$ and $y = 3$ will fire in synchrony with the other properties of Geon B, say, on time slices 3, 7, and 11. Recall that below|$y = 1$ receives an excitatory input from $y = 3$. Therefore, when $y = 3$ fires (Time Slice 3), below|$y = 1$ will receive an excitatory input, and its activation will go above zero. On Time Slice 5, $y = 1$ will fire and send an enabling signal to below|$y = 1$, causing it to fire (i.e., in synchrony with $y = 1$ and, by transitivity, Geon A's other properties). Then below|$y = 1$ sends an excitatory signal to the L5 below cell, causing it to fire with geon A's other properties. In a directly analogous manner, *above* will come to fire in synchrony with the other properties of Geon B.

One problem with this architecture is a potential to "hallucinate" relations between a geon and itself. Such hallucinations can result if a geon's metric properties are coded coarsely, as they are in this model. For example, a given geon's vertical

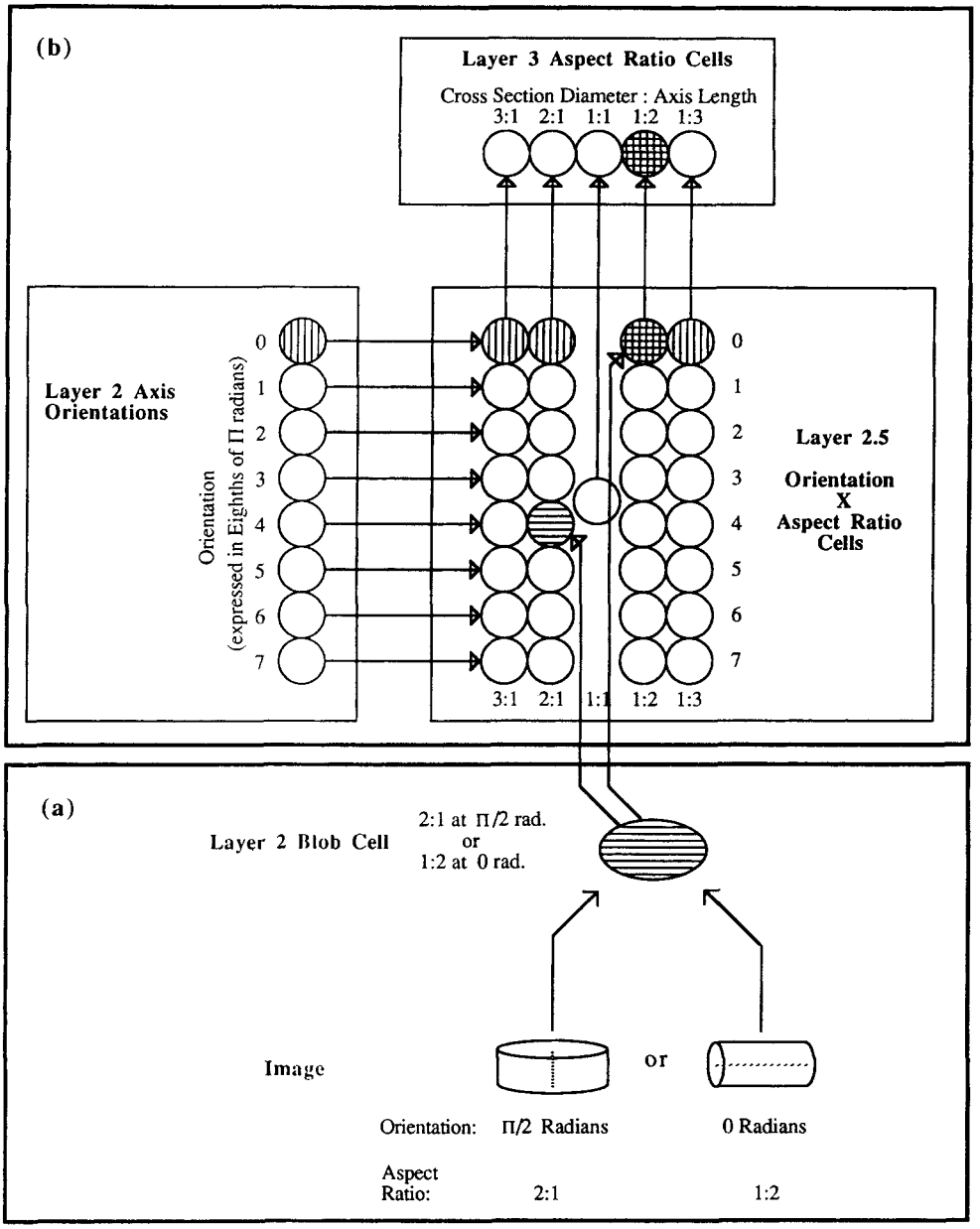


Figure 24. Determining a geon's aspect ratio. (a: A given blob is consistent with two aspect ratios, an elongated geon with the same orientation as the blob or a flattened geon whose orientation is perpendicular to that of the blob. b: A layer of units that could determine aspect ratio from the axis and blob cells activated by a geon's image. rad. = radian.)

position in the visual field might be represented by simultaneous activity in both $y = 2$ and $y = 3$. Because $y = 2$ is below $y = 3$, the L4 below|$y = 2$ cell could become active and fire in response to the presence of that geon even if there are no others in the visual field. This problem is overcome by giving each L4 cell a blind spot for the L3 value directly flanking its preferred value. For example, below|$y = 2$ receives excitation from $y = 4$, $y = 5, \ldots y = 10$ but not from $y = 3$. The L4 blind spot prevents hallucinations, but it has the negative side effect that, for small enough differences in metric values, relations may be undetectable. For example, two very flat geons, one on top of the other, may activate the same L3 vertical position cells, only with slightly different ratios (say, one excites $y = 2$ strongly and $y = 3$ weakly; the other excites $y = 2$ weakly and $y = 3$ strongly). Because of the blind spot, the model would be insensitive to the above–below relation between such geons.

Layers 6 and 7: Geon Feature Assemblies and Objects

Together, Layers 3 and 5 produce a pattern of activation, termed a *geon feature assembly* (GFA), describing a geon in

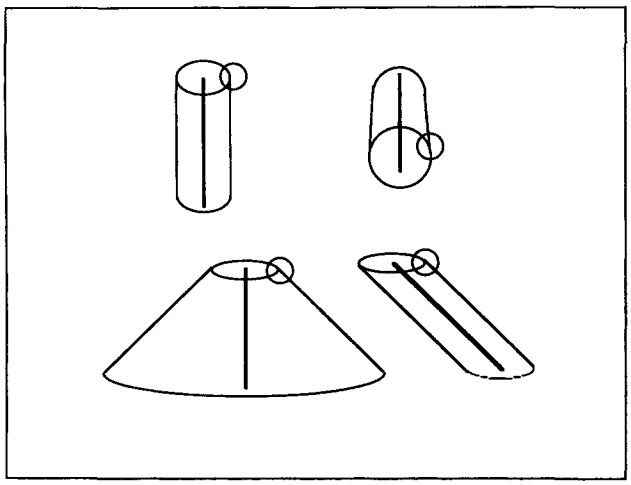


Figure 25. Alone, the orientation of a geon's major axis is not sufficient to determine the 3-D orientation of that geon. (As shown on the left, a vertical axis could arise from a horizontal geon with its end toward the viewer or from a vertical geon. Similarly, as shown on the right, the orientation of a vertex is also insufficient to determine the orientation of the geon.)

terms of its shape and general orientation as well as its location, size, and orientation relative to the other geons in the image. The collection of GFAs (over different time slices) produced in response to an object constitutes a simple structural description that is invariant with translation, scale, and orientation in depth. Furthermore, the model will produce the same description just as quickly in response to an object whether it is "viewing" that object for the first time or the twentieth. In this sense, the model's first five layers accomplish the primary theoretical goals of this effort. However, to assess the sufficiency of this representation for viewpoint-invariant recognition, we have included two additional layers (6 and 7) that use the L3–L5 output to perform object classification.

The connections and FELs in JIM's first five layers are fixed, reflecting our assumption that the perceptual routines governing the activation of structural descriptions remain essentially unchanged past infancy. The acquisition of new objects is assumed to consist entirely in the recruitment of units (using modification of connections to existing cells) in Layers 6 and 7. There is an important and difficult question here as to exactly how and when new object classes are acquired. For example, when we see an object that is similar in shape to some familiar class of objects, how do we decide whether it is similar enough to belong to that class or whether it should become the first member of a new object class? ("I think it's a hot dog stand, but it could be some kind of device for sweeping the sidewalk.") Proper treatment of this question requires a theory of categorization, and the emphasis of this effort is on viewpoint invariance. Although it is possible that JIM's GFAs could serve as the representation in such a theory, a complete theory also requires processes to operate on its representations. Rather than attempt to propose such a theory, we have chosen instead to use a theoretically neutral procedure for providing the model with its vocabulary of objects. This simplified "familiarization" procedure is described shortly. First, let us consider how the model's last two layers activate a representation of a familiar object class given a structural description as input.

Given that almost any upright view of an object will produce the same GFA pattern in L3–L5 (within a small range of error), the task of classifying an object from its GFA pattern is straightforward. It is accomplished by allowing cells in L6 to be recruited by specific GFAs and cells in L7 to be recruited by conjunctions of L6 cells. If an object is in JIM's vocabulary,

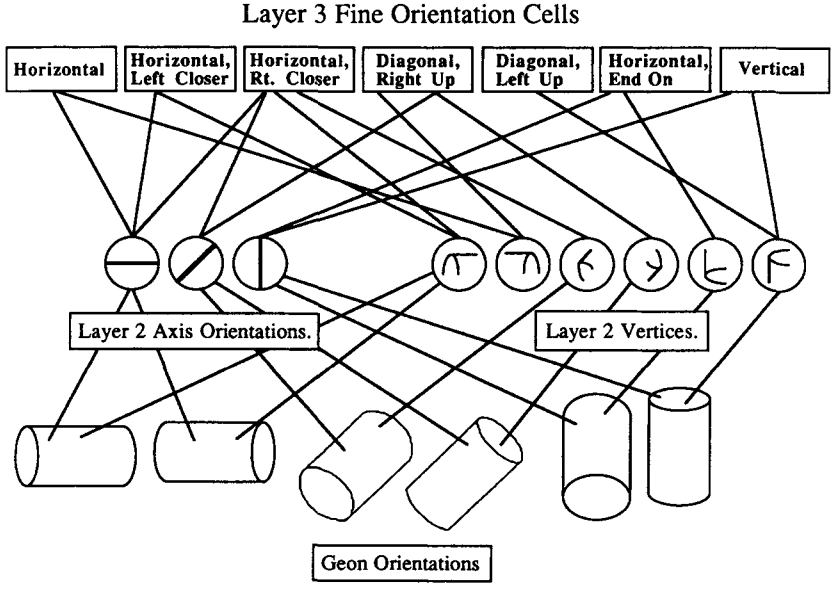


Figure 26. Illustration of the mapping from Layer 2 axis and vertex cells to Layer 3 orientation cells. (Rt. = right.)

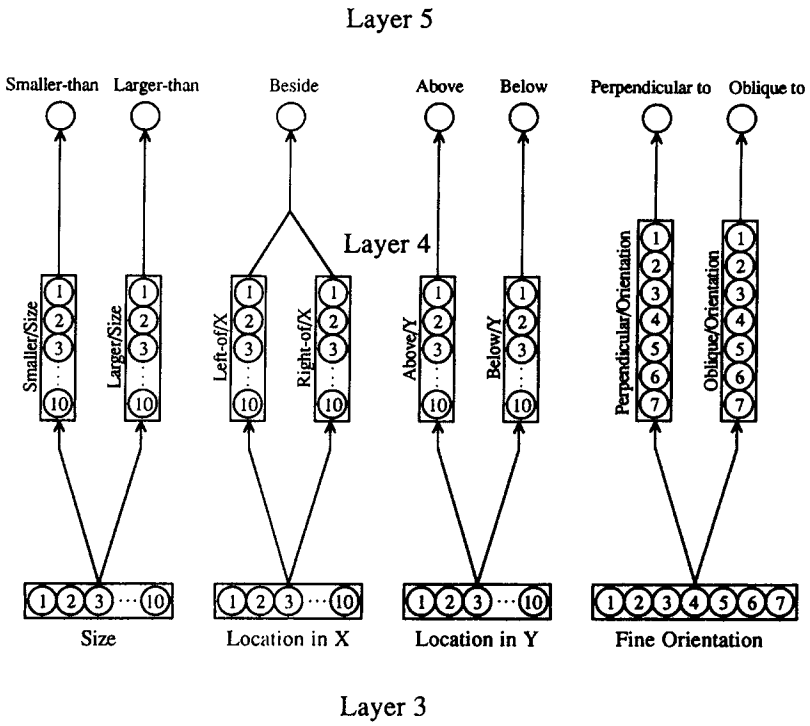


Figure 27. Detail of Layers 4 and 5. (Each Layer 4 cell responds to a specific relation [e.g., below] in conjunction with a specific value on the dimension over which that relation is defined [e.g., $Y = 1$]. Layer 5 contains only one cell for each relation.)

then each of its GFAs will activate a different cell in L6 (henceforth, GFA cells). The outputs of the GFA cells are summed over separate time slices to activate an L7 unit representing the class of the object.

The cells of L6 are fully interconnected to L5 and to the subset of L3 that passes activation to L6 (i.e., all L3 cells except those for size, position, and fine orientation). One L6 cell is a dummy cell, with unit strength excitatory connections from all L3–L5 cells. This cell is included to mimic the effect of GFA cells that are still free to be recruited in response to novel object classes. The remaining cells are selective for specific conjunctions of geon and relation attributes (for the current vocabulary, there are 20). For example, one GFA cell receives strong excitation from the L3 cells for curved cross section, straight axis, nonparallel sides, vertical, aspect ratio 1:1, and aspect ratio 1:2, and from the L5 cells for above, smaller than, and perpendicular to; the remaining connections to this cell are negligibly small. Thus, this GFA cell is selective for slightly elongated vertical cones that are above, smaller than, and perpendicular to other geons. Other GFA cells are selective for different patterns, and all GFA cells (including the dummy cell) are mutually inhibitory.

The most difficult problem confronting a GFA cell is that the pattern for which it is selective is likely to overlap considerably with the patterns selected by its competitors. For example, many objects contain geons with curved cross sections or geons that are below other geons, and so forth. Furthermore, some GFAs may be subsets of others: One GFA cell might respond to vertical cylinders below other geons, and another might respond to vertical cylinders below and beside other geons. To allow the GFA cells to discriminate such similar patterns, we adopted an excitatory input rule described by Marshall (1990), and others:

$$E_i = \Sigma_j O_j w_{ij} / (\alpha + \Sigma_j w_{ij}), \tag{10}$$

where E_i is the excitatory input to L6 cell i, O_j is the output of cell j (in L3 or L5), w_{ij} is the weight of the connection from j to i, and α is a constant. This equation normalizes a cell's excitatory input to a Weber function of the sum of its excitatory connections, making it possible for the cell to select for patterns that overlap with—or are even embedded within—the patterns selected by other cells. To illustrate, consider a simple network with two output cells (corresponding to L6 cells in JIM) and three input cells (corresponding to the L3–L5 cells), as shown in Figure 29. For simplicity, assume outputs and connection weights of 0 or 1, and let $\alpha = 1$. Output Cell 1 is selective for input Pattern ABC; that is, $w_{1A} = w_{1B} = w_{1C} = 1$, and Output Cell 2 is selective for Pattern AB. If ABC is presented to this network (i.e., $O_A = O_B = O_C = 1$), then, by Equation 10, E_1 will be 0.75, and E_2 will be 0.67. In response to ABC, Cell 1 receives a greater net input and therefore inhibits Cell 2. By contrast, if Pattern AB is presented to the network, E_1 will be 0.50, E_2 will be 0.67, and Cell 2 will inhibit Cell 1.

By allowing the GFA cells to select for overlapping and embedded patterns, this input rule allows JIM to discriminate

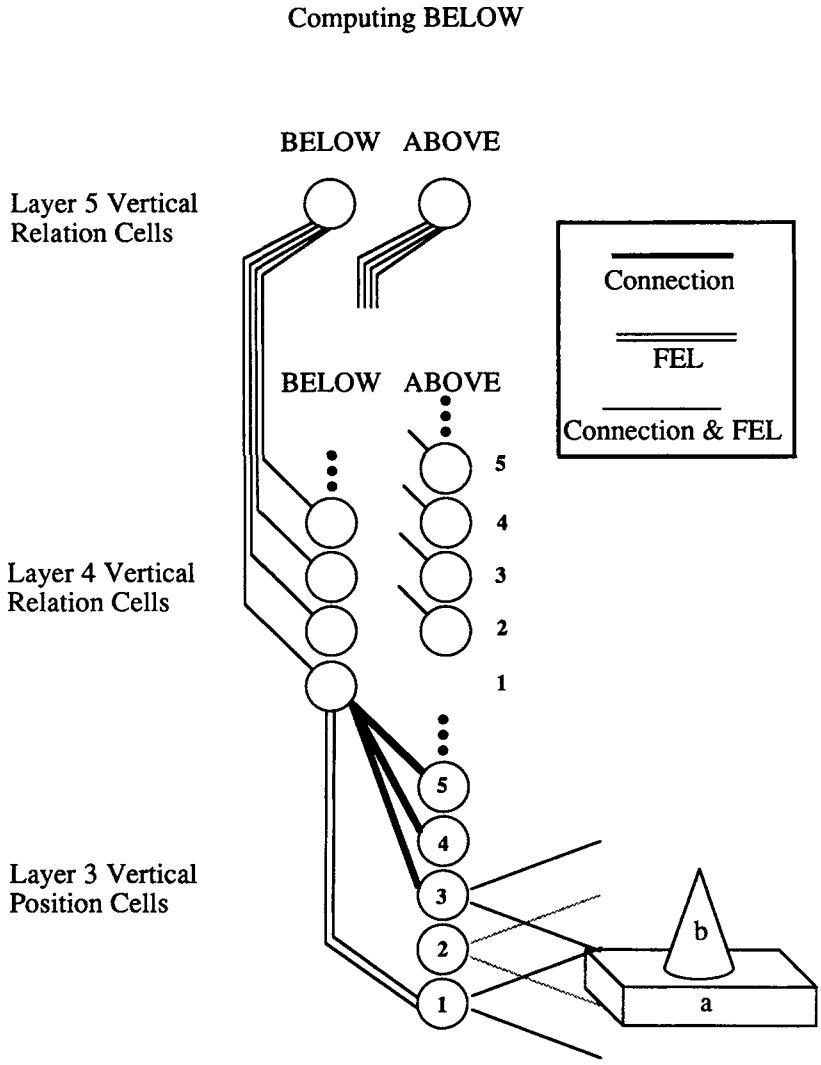


Figure 28. Operation of Layers 4 and 5 illustrated with the relation below. (FEL = fast enabling link.)

objects with very similar structural descriptions: Each possible pattern of activation over L3 and L5 cells could potentially recruit a different GFA cell in L6. Given 21 inputs to L6, the number of GFA cells could in principle be as great as 2^{21} (or larger, if more than two degrees of activity were discriminated). However, because GFA cells are recruited only as the model adds objects to its vocabulary, the number of such cells that would realistically be required (even to represent the approximately 150,000 object models familiar to an adult [Biederman, 1988] is considerably smaller).

Simulations

JIM was implemented and run on an IBM PSII, Model 70. Simulations tested JIM's capacity for invariance with translation, scale changes, left–right image reflection, and rotations in depth and in the visual plane. These simulations were conducted in two phases, a familiarization phase and a test phase. During familiarization, the model was presented with one view of each of 10 simple objects and allowed to modify the connection weights to L6 and L7 in response to the patterns of activity produced in L3 and L5. After familiarization, the connections were not allowed to change. During the test phase, JIM was presented with each object in 10 views: the original (baseline) view that was used for familiarization and 9 novel views. Its performance was evaluated by comparing its L7 responses to the test images with its baseline responses at that layer.

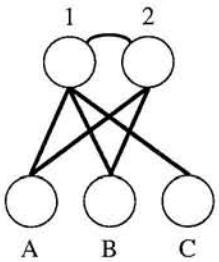
Familiarization: Creating the Object Vocabulary

Let us refer to the vector of bottom-up excitatory connection weights to a GFA cell as its *receptive field.* Before familiarization, all GFA cells (except for the dummy cell) were initialized: their receptive fields were set to vectors of zeros, and their output connections to all L7 object cells were set to −0.5.

JIM's vocabulary currently consists of 2 three-geon objects and 8 two-geon objects. The baseline view of each object is depicted in Figure 30. JIM was familiarized with one object at a

time, each by the following procedure: The model was given the baseline image of an object as input and run for 20 time slices (ts). Each time the L1–L2 features of a geon fired, L3 and L5 produced a GFA as output. The GFA from the latest ts associated with each geon was selected as the familiarization GFA (denoted GFA*) for that geon. For example, the GFA*s for the telephone in Figure 30 were selected by presenting the baseline image to the model and running the model for 20 ts. Assume that the L2 features of the brick fired on ts 2, 7, 11, and 15, and those of the wedge fired on 4, 8, 13, and 18. The L3/L5 outputs for ts 15 would be selected as GFA* for the brick, and those for ts 18 as GFA* for the wedge.

Once an object's GFA*s were generated, the object was added to JIM's vocabulary by recruiting one L6 GFA cell for each GFA* and one L7 object cell for the object as a whole. For each GFA*, *i*, a GFA cell was recruited by computing the Euclidean distance[14] (D_{ij}) between *i* and the receptive fields of all previously recruited GFA cells, *j*. If, for all *j*, D_{ij} was greater than 0.5, a new GFA cell was recruited for GFA* *i* by setting the receptive field of an unrecruited GFA cell (i.e., one still in its initialized state) equal to GFA* *i* and setting the connection from that GFA cell to the associated object cell to 1.0. If a previously recruited cell, *j*, was found such that $D_{ij} \leq 0.5$, then the receptive field of that cell was set to the mean of itself and GFA* *i*, and its connection to the associated object cell was set to 1.0. This procedure recruited 20 GFA cells for the complete set of 22 GFA*s in the training set.



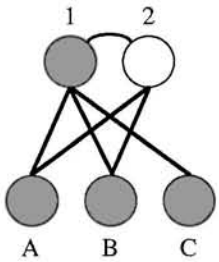
Excitatory Weights:

$w_{1A} = w_{1B} = w_{1C} = 1$

$w_{2A} = w_{2B} = 1, \; w_{2C} = 0$

Net Excitatory Input:

$$E_i = \sum_{j=A}^{C} O_j w_{ij} / (\alpha + \sum_{j=A}^{C} w_{ij}), \; \alpha = 1$$

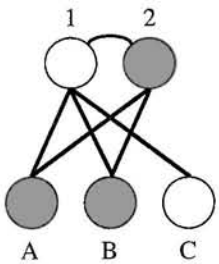


Input Pattern [ABC]:

$O_A = O_B = O_C = 1$

$E_1 = (3)/(1 + 3) = 0.75$

$E_2 = (2)/(1 + 2) = 0.67$



Input Pattern [AB]:

$O_A = O_B = 1, \; O_C = 0$

$E_1 = (2)/(1 + 3) = 0.50$

$E_2 = (2)/(1 + 2) = 0.67$

Figure 29. The Weber fraction excitatory input rule allows Output Cell 2 to select for Input Pattern AB and Output Cell 1 to select for Input Pattern ABC.

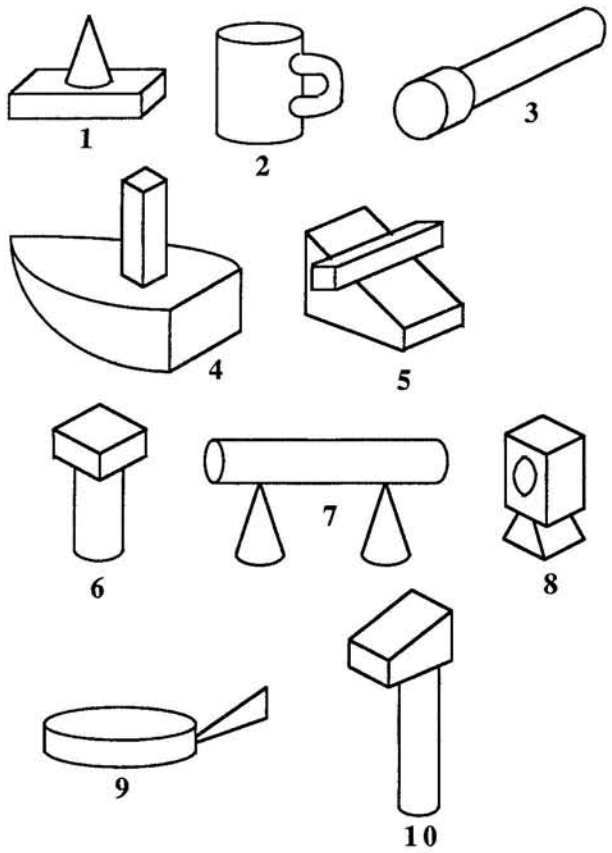


Figure 30. The baseline view of each object in JIM's vocabulary. (Objects 7 and 8 contain three geons, all others contain two. Objects 8 and 9 contain ambiguous geons: the central ellipse in Object 8 contains no axis information, and the "handle" on the "frying pan" [Object 9] contains no cross-section information. Objects 1 and 10 have the same geon and relation attributes in different combinations [see Figure 7].)

It is particularly important to note that this procedure establishes the model's vocabulary on the basis of only one view of each object. As such, exposure to many different views of each object cannot account for any viewpoint invariance demonstrated in the model's recognition performance. Also, this procedure is tantamount to showing the model a line drawing and instructing it that "this is an *X*." In an earlier version of the model (Hummel & Biederman, 1990b), the object vocabulary was developed by allowing the sixth and seventh layers to self-organize in response to GFA*s selected as described earlier. With a five-object training set, that procedure settled on a

[14] Recall that both the GFAs and the L6 receptive fields are 21-dimensional vectors. The Euclidean distance (D_{ij}) between two vectors, *i* and *j*, is calculated as the square root of the sum over vector elements of the squared differences of corresponding vector elements: $D_{ij} = (\Sigma_k (i_k - j_k)^2)^{0.5}$.

stable pattern of connectivity in L6 and L7, and the resulting connections produced recognition results very similar to those produced by the current version of the model. However, where the current familiarization procedure is capable of acquiring objects one at a time, the self-organizing system required all the objects to be presented at the same time. Where the current procedure establishes a pattern of connectivity in one exposure to an object, the self-organizing algorithm required 3,000 presentations of the training set to settle into a stable configuration. The self-organization was also highly sensitive to its parameters and to the ratio of the number of GFA*s to the number of L6 cells (e.g., it would not settle unless the number of L6 cells exactly equaled the number of GFA*s). We decided to use the current familiarization procedure instead because we want the model's behavior to reflect the performance of its lst five layers rather than the idiosyncrasies of any particular learning algorithm.

Test Simulations: Conditions and Procedure

After familiarization, JIM's capacity for viewpoint-invariant recognition was evaluated by running two blocks of simulations. Each block consisted of 100 runs of the model, 10 objects presented in each of 10 conditions: Condition 1, Baseline, in which the original (familiarization) image was presented; Condition 2, Translation, in which the original image was moved to a new position in the visual field; Condition 3, Size, in which the original image was reduced to between 70% and 90% of its original size; Condition 4, Mirror Reversal of the original image; Condition 5, Depth rotation of 45° to 70° of the original object; Conditions 6 to 10, in which five images were created by rotating the original image in the visual plane (22.5°, 45°, 90°, 135°, and 180°). Blocks 1 and 2 differed only in the number, N, of ts for which the model was allowed to run on each image: In Block 1, $N = 20$, and in Block 2, $N = 40$.

Simulations were conducted by activating the set of L1 edge cells and L2 axis cells corresponding to the image of an object and allowing the model to run for N ts. Cells in all layers of the model updated their activations and outputs as described in the previous sections. On each of the n ts in which L1–L2 outputs were generated[15], the activity of the target cell (the L7 cell corresponding to the correct identity of the object) was monitored. No data were collected on those $(N - n)$ ts in which no output was generated.

Four response metrics were calculated for each simulation because, alone, any one metric has the potential to yield a misleading characterization of performance. The metrics calculated were maximum activation of the target cell activated during the simulation (Max), mean activation of the target cell over the n ts in which data were collected, proportion (P) of all object cell mean activations attributable to the target cell, and the mean activation multiplied by proportion (MP). Max and mean provide raw measures of the target cell's response to an image. P and MP reflect the strength of the target cell's response relative to the responses of all other object cells. The Appendix discusses how these response metrics were calculated and their relative merits and disadvantages.

As is evident in Figure 31, all the response metrics yielded the same qualitative picture of the model's performance, so most simulations are reported only in terms of max. For each block of simulations, JIM's performance in each condition (e.g., baseline, translation, and so forth) is reported in terms of the mean (over objects) of max, but the ordinate of each graph is labeled with the individual response metric (e.g., if the graph shows mean max's over objects in each condition, the ordinate will be labeled *max*). Error bars indicate the standard error of the mean. Because each metric is proportional to the strength and correctness of the model's response to an image, high values of the response metrics are assumed to correspond to low reaction times and error rates in human subjects.

Test Simulations: Results

There is a stochastic component to the refractory thresholds of the cells in L1 and L2, so the output of the target cell in response to an image is subject to random variation. Specifically, the cell's output will reflect the number of times the features of each geon in the image fires, the number of ts between different geons' firing, and the order in which they fire. To derive an estimate of the amount of variation that can be expected for a given image of an object, the baseline view of Object 1 was run 20 times for 20 ts per simulation and 20 times for 40 ts per simulation. The means and standard deviations (unbiased estimate) of the four response metrics obtained in the 20-ts runs were max = 0.499 (SD = 0.016), mean = 0.304 (SD = 0.022), P = 1.0 (SD = 0.0), and MP = 0.304 (SD = 0.022). The values obtained in the 40-ts runs were max = 0.605 (SD = 0.005), mean = 0.436 (SD = 0.012), P = 1.0 (SD = 0.0), and MP = 0.436 (SD = 0.012). These figures are reported only to provide an estimate of the amount of random variation that can be expected in JIM's performance.

Translation, Size, Mirror Reversal, and Rotation in Depth

Recall that humans evidence no perceptual cost for image translations, scale changes, and mirror-image reversals (Biederman & Cooper, 1991a, 1992), and only a very modest cost for rotations in depth, even with nonsense objects (Gerhardstein & Biederman, 1991). Similarly, JIM's performance reveals complete invariance with translation, size, and mirror-image reversals. Although every test condition entailed translating the original image (it is impossible to scale, rotate, or mirror reflect an image without affecting where its constituent edges fall in the visual field), JIM was tested on one image that underwent only a translation from the baseline image. JIM was also tested with one scaled image (reduced to between 70% and 90% of its original size), one left–right mirror-reflected image, and one depth-rotated image of each object. Limitations of the stimulus creation program made precise rotations in depth impossible (the stimulus creation program represents objects as 2-D line drawings rather than 3-D models). Therefore, each depth rotated image was created by rotating the object approximately 45° to 70° from its baseline orientation; these included rotations both

[15] Because there is a stochastic component to the L1 and L2 cells' firing, a subset of the time slices will pass during any given run without any L1 or L2 cells firing (and, therefore, no other cells will fire). On these time slices, no data were gathered.

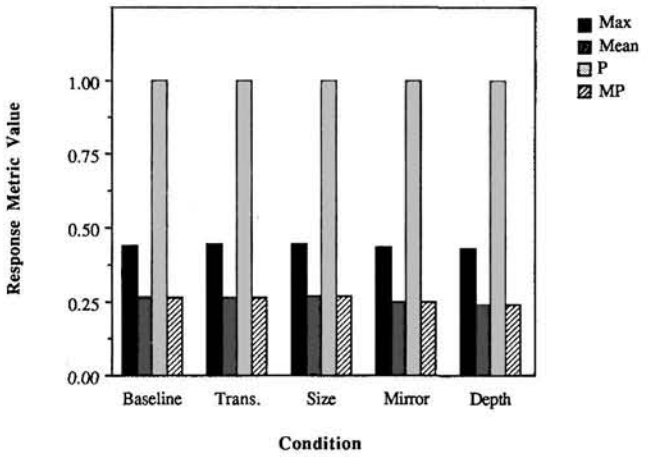


Figure 31. JIM's performance in the baseline, translation (Trans.) only, size, mirror-image reversed (Mirror), and depth rotated (Depth) conditions expressed in terms of the average-maximum (Max), mean proportion (P), and mean activation multiplied by proportion (MP) response metrics over objects in each condition. (These data were gathered in simulations lasting 20 time slices.)

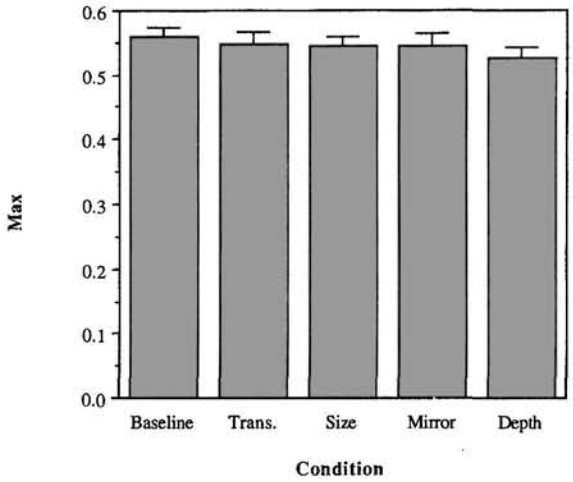


Figure 32. JIM's performance in the baseline, translation (Trans.) only, size, mirror-image reversed (Mirror), and depth rotated (Depth) conditions expressed in terms of the average maximum (Max) response metric over objects in each condition. (These data were gathered in simulations lasting 40 time slices.)

about the object's vertical axis and about a horizontal axis perpendicular to the line of sight.

Figure 31 depicts JIM's performance in each of these conditions in terms of all four response metrics in Block 1 (simulations lasting 20 ts). By each response metric, performance in these conditions was indistinguishable from baseline performance. Figure 32 shows the max for these conditions in Block 2 (40 ts). These figures also reveal complete invariance in each condition, with the exception of a very modest cost for rotation in depth. This cost most likely reflects changes in the geons' aspect ratios resulting from the depth rotations. Note that mean P over objects was 1.0 for all conditions in Block 1, indicating that in each simulation only the target cell achieved a mean activation greater than zero. Although it is not shown, the mean P over objects was also 1.0 for all these conditions in Block 2.

Rotation in the Visual Plane

By contrast to rotations in depth, humans evidence a perceptual cost for recognizing images that have been rotated in the visual plane (Jolicoeur, 1985). Typically, subjects show a monotonic increase in response time and error rate, with the number of degrees rotated from upright to approximately 135°. Subjects respond faster at 180° (when the object is upside down) than they do at 135°, producing a W-shaped rotation function over 360° (Jolicoeur, 1985).

JIM's performance under rotations in the visual plane was tested with stimuli rotated 22.5°, 45°, 90°, 135°, and 180° from the baseline images. Figures 33 and 34 show max as a function of degrees of rotation for Blocks 1 and 2, respectively. Again, JIM's performance revealed a trend very similar to that of human subjects. In terms of JIM's operation, the cusp in the rotation function at 180° reflects two effects. First, for objects with a primarily vertical structure (i.e., the spatial relations among the geons include *above* and *below,* but not *beside*), rotations between upright (0°) and 180° create spurious *beside* relations that are absent in the upright and 180° views. For example, rotate a lamp 45° in the visual plane and the lampshade, which is above the base in the upright view, will be above and beside the base. Continue to rotate the lamp until it is upside down and this spurious *beside* relation disappears. Second, a 180° rotation in the visual plane preserves the original coarse orientations of the object's component geons (*horizontal, diagonal,* or *vertical*) more than rotations greater or less than 180°.

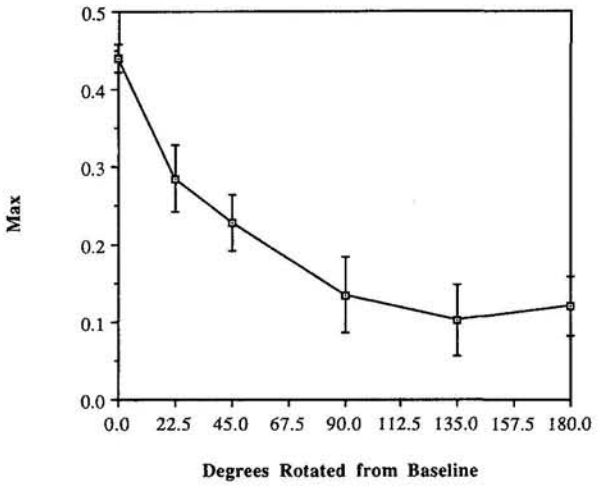


Figure 33. JIM's recognition performance as a function of degrees rotated in the visual plane from baseline (0.0°). (Performance is expressed in terms of the average maximum [Max] response metric over objects in each condition. These data were gathered in simulations lasting 20 time slices.)

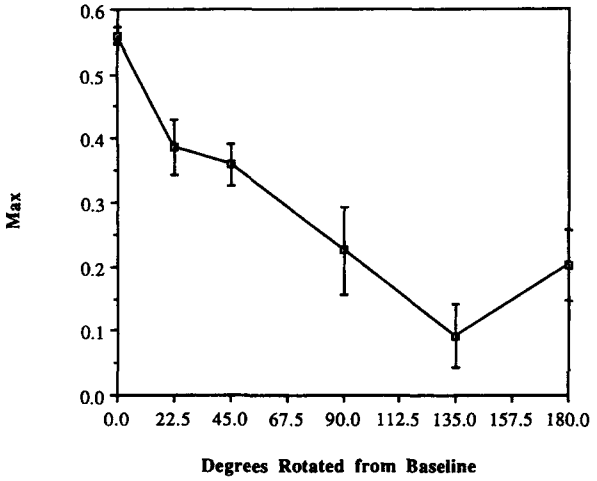


Figure 34. JIM's recognition performance as a function of degrees rotated in the visual plane from baseline (0.0°). (Performance is expressed in terms of the average maximum [Max] response metric over objects in each condition. These data were gathered in simulations lasting 40 time slices.)

Discussion of Test Simulations

The simulations reported earlier reveal a high degree of translation, scale, and mirror-reflection invariance in JIM's recognition of objects. Performance with objects rotated in depth and in the visual plane also resembles the performance of human subjects under similar circumstances. Comparison of results from Block 1 and Block 2 suggests that the length of time for which a simulation is run has little or no qualitative impact on the model's performance (mean and max were higher on average for the 40 ts simulations, but the relative values across conditions were unaffected). Similarly, comparison of the various response metrics suggests that the observed results are not dependent on the particular metric used. However, because the results are based on simulations with only 10 objects, it could be argued that they reflect the model's use of some sort of diagnostic feature for each object, rather than the objects' structural descriptions.

This diagnostic feature hypothesis is challenged by the results of two additional test conditions conducted with images of the objects in JIM's vocabulary. In the first condition, a scrambled image of each object was created by randomly rearranging blocks of the original image such that the edges' orientations were unchanged and the vertices remained intact. An example of a scrambled image is given in Figure 35. If JIM's performance reflects the use of a diagnostic list of 2-D features for each object, it would be expected to achieve at least moderate activation in the target cell corresponding to each scrambled image (although the axes of parallelism and symmetry are destroyed by the scrambling, the vertices are preserved), but if it is sensitive to the relations among these features, it would be expected to treat these images as unfamiliar. In the second condition, the intact baseline images were presented to the model, but the binding mechanism was disabled, forcing the separate geons to fire in synchrony. This condition demonstrated the effect of accidental synchrony on JIM's performance; if the model's capacity for image parsing is truly critical to its performance, recognition would be expected to fail in this condition as well. Simulations in both conditions lasted for 40 ts. JIM's performance on the scrambled image and forced accidental synchrony conditions is shown in Figure 36. In both conditions, performance was reduced to noise levels.

Discussion

JIM is capable of activating a representation of an object given a line drawing of that object as input. Moreover, that representation is invariant with translation, scale, and left–right mirror reversal even when the model has previously been exposed to only one view of the object. This section discusses the important aspects of JIM's design and explores their implications. Specifically, the implications of the independent attribute representation used in L3 and L5 are reviewed shortly. The model's use of dynamic binding plays a pivotal role in this capacity for independent attribute representations. The theoretical and empirical implications of using temporal synchrony for dynamic binding is discussed in some detail. Also addressed are additional findings for which JIM provides an account, some novel predictions for human recognition performance, and some limitations of the current architecture.

The Nature of Temporal Binding and Its Empirical Consequences

What Temporal Binding Is Not

The notion of temporal binding is only starting to become familiar to most psychologists, and its empirical consequences are not obvious. Specifically, binding through temporal synchrony as described here should not be confused with the grouping of stimulus elements that are presented in close temporal contiguity. Indeed, JIM produces temporal asynchrony for the different geons in an object even though they are presented simultaneously. The confusion between binding through synchrony and grouping features presented in close temporal

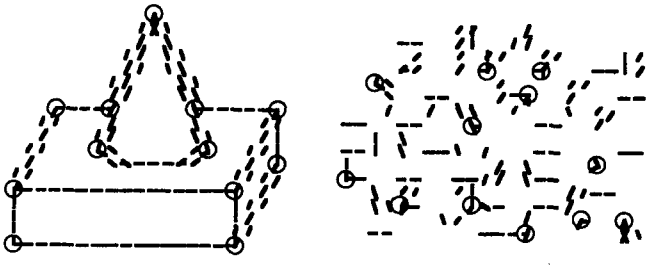


Figure 35. Left: The representation of the baseline image of Object 1 in JIM's first layer. (Line segments correspond to active cells in Layer 1, the location and orientation of a segment corresponding to the preferred location and orientation of the cell, respectively. Circles indicate locations in Layer 2 containing active vertex (or lone termination) cells. Right: The scrambled image version of the baseline image for Object 1. Vertices and edge orientations are preserved, but the positions of 2 × 2 blocks of the 22 × 22 cluster input layer are randomly rearranged.)

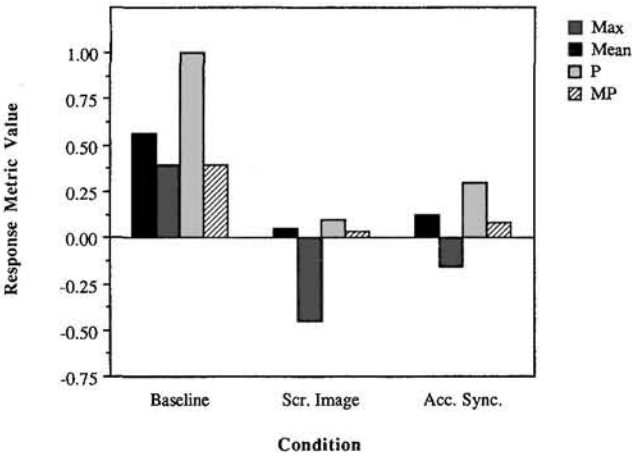


Figure 36. JIM's recognition performance in the baseline, scrambled image (Scr. Image), and forced accidental synchrony (Acc. Sync.) conditions. (Performance is expressed in terms of all response metrics [averaged over objects in each condition]. These data were gathered in simulations lasting 40 time slices. Max = maximum activation of the target cell activated during the simulation; P = proportion; MP = mean activation multiplied by proportion.)

contiguity is not hypothetical. A report by Keele, Cohen, Ivry, Liotti, and Yee (1988) is critical of temporal binding on the grounds that temporal contiguity in stimulus presentation did not predict feature binding.

Keele et al. (1988) conducted three experiments testing whether common location or common time of occurrence is a stronger cue to feature binding in rapidly presented images. In each experiment, illusory conjunctions were better predicted by common location in the visual field than by co-occurrence in time. Failing to find support for the primacy of stimulus contiguity in feature binding, these researchers rejected temporal binding theory. However, what their data show is that location is a more critical cue than temporal coincidence in determining how features should be conjoined. Such data do not and cannot falsify the hypothesis that synchronous activity is the manner in which a feature conjunction—even one established on the basis of common location—is represented. How a binding is represented and what cues are used to determine that binding are different issues entirely. Although it is possible to evaluate a given set of proposed binding cues with behavioral measures, we suggest that behavioral tests are inadequate in principle to falsify the temporal binding hypothesis. Rather, questions about the neural mechanism of binding, by definition, reside in the domain of neuroscience.

Physiological Evidence for Temporal Binding

Is there any evidence for binding through temporal synchrony that would be compatible with our proposal? Recently, Gray et al. (1989) reported the existence of a 50 Hz oscillatory potential in Area 17 of the anesthetized cat. The potentials were not of individual spikes (which were filtered out) but the summed dendritic activity of a large number of neurons in a cortical column. With multiple recordings, Gray et al. computed the cross correlation of this activity at different sites whose receptive fields had been determined. Moderately high cross correlations, suggesting phase locking of oscillatory activity,[16] were observed for placements at adjacent sites, whatever the cell's orientation preference and nearby sites that had similar orientation preferences. In JIM's terms, these could reflect the FELs corresponding to Condition 1, local coarse coding of image contours, among units with overlapping receptive fields and similar orientation preferences.

The most provocative results were obtained from recordings made at widely separated sites that had nonoverlapping receptive fields. The cross-correlation values for these sites were essentially zero unless the orientation preferences were collinear, in which case the values were positive, but modest. If bars were translated separately through the receptive fields but in opposite directions, as shown in Figure 37a, then the correlations were low. Translating the separate bars in a correlated motion, as shown in Figure 37b, increased the cross correlation. However, joining the two bars into one bar, so that the intervening portion of the visual field was bridged, as shown in Figure 37c, dramatically increased the value of the cross correlation.

JIM's Single-Unit Phase-Locking Predictions

We suggest two additional tests of the proposed solution to the binding problem using the Gray et al. (1989) experimental paradigm. In particular, we describe experiments to assess (a) whether the phase locking can turn corners and (b) whether collinear phase locking can be broken through intervening vertices.

Can phase locking turn corners? As a test of whether phase locking can turn corners, consider three sites in visual cortex with noncollinear receptive fields, as shown in Figure 38a. These experiments might be best performed with awake primates trained to maintain fixation, in the manner of Moran and Desimone (1985). Assume that in the absence of stimulation, or with independent stimulation, the cross-correlation values between the sites are low. However, if Sites 1 and 2 were grouped into a single shape (Figure 38b), with Site 3 grouped into another shape, would the cross correlation be high only between Sites 1 and 2? If Sites 1 and 3 were grouped into a single shape, with Site 2 the odd unit out, Sites 1 and 3 would be expected to have a high cross correlation, with Site 2 uncorrelated (Figure 38c). Similarly, Sites 2 and 3 could be grouped, with Site 1, the odd unit out. An attractive aspect of this test is that it can be performed on any three sites, because the shapes can be designed after the tuning preferences and receptive fields for the units are determined.[17]

[16] The cross correlations are the domain of frequency rather than phase (C. M. Gray, personal communication, April 1991). That is, high cross correlations result when neurons' firing rates increase and decrease together. It is the phases of the curves describing firing rate that are observed to be locked in these investigations.

[17] It is possible that within-shape correlations could be produced by an attentional spotlight (Crick, 1984). If a fourth site was added to the shape with the odd unit out in Figures 38b or 38c, spotlight theory would predict that there would be correlated firing within only one shape at a time. The theory presented here predicts within-shape correlated firing for both shapes simultaneously. (This analysis was proposed by G. E. Hinton, personal communication, February 1992.)

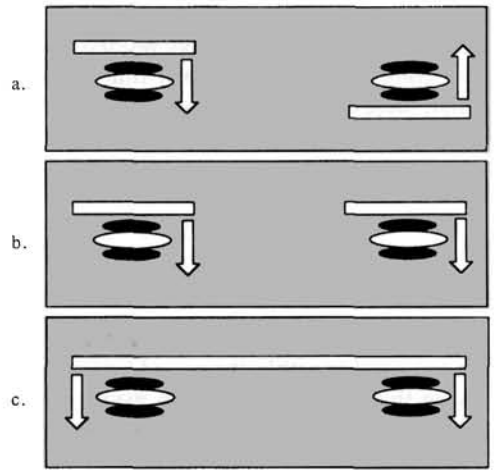


Figure 37. Summary of the stimulus conditions in the experiments by Gray, Konig, Engel, and Singer (1989). (Recordings were made from sites with distant, collinear receptive fields in Area 17 of the anesthetized cat. a: When the receptive fields were stimulated with separate bars of light moving in opposite directions, the cross correlation of the activity in the separate sites was low. b: When the bars were translated across the receptive fields in phase, cross correlations were moderate. c: Cross correlations were highest when the receptive fields were stimulated with a single long bar of light.)

Do vertices prevent collinear phase locking? Can collinear phase locking be broken by intervening vertices that group the collinear segments into separate shapes? For this problem, units with collinear receptive fields would first have to be recruited, as shown in Figure 39a. Assume that collinear bars of light, as shown in Figure 39b, would result in a high correlation. The critical test would be whether the phase locking would be broken if the bars' collinear endpoints were replaced by vertices, as shown in Figure 39c. Here we would expect that the phase locking would be greatly reduced.

Open Questions About Synchrony for Binding and Structural Description

The data reported by Gray et al. (1989), suggest the existence of neural activity that might provide a basis for perceptual organization through temporal binding in a manner compatible with the FELs developed for JIM. However, they should be interpreted with caution. First, phase locking is inferred from cross correlations indicating an average phase lag of zero, but it is unclear whether this statistic reflects the phase locking in individual cells or simply a distribution of phase relations whose mean is a zero phase lag. A related question concerns whether the phase locking reflects lateral interactions among cortical neurons or simply a statistical epiphenomenon (e.g., perhaps the phase relations reflect the statistical or temporal properties of the inputs to cortex). This statistical epiphenomenon explanation was challenged by a recent report by Engel, Konig, Kreiter, and Singer (1991). Using a paradigm similar to that of Gray et al., Engel et al. showed that synchrony could extend across the corpus callosum. Section of the callosum resulted in a loss of synchrony across the hemispheres but left the synchrony unaffected within the hemispheres.

Second, there is some debate as to whether phase locking plays a functional role in binding visual attributes or subserves some different function. For example, Bower (1991) has argued that similar (but global) phase locking in olfactory cortex plays a role in the modification of synaptic strengths but serves no purpose for binding attributes. Indeed, global phase locking (in which all active units fire approximately in phase) could not serve to differentiate bound subsets of active units. Third, phase locking in visual cortex has only been observed with moving stimuli and is notoriously difficult to measure with stationary stimuli (C. M. Gray, personal communication, April 1991). Finally, even if we assume that the observed phase locking in Area 17 of the cat were serving to bind visual attributes, it is unclear whether a 50 Hz oscillation is rapid enough to implement binding for real-time visual recognition.

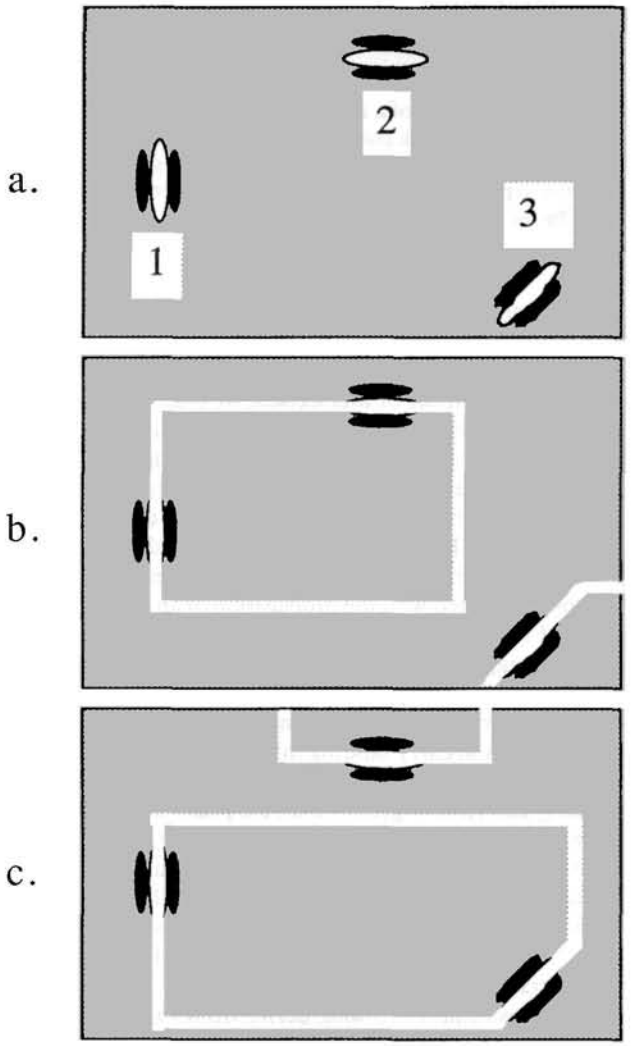


Figure 38. Illustration of a multiple-electrode test of whether phase locking can turn corners. (a: Three receptor sites, 1, 2, and 3, with noncollinear receptive fields. b: Will Sites 1 and 2 fire in phase, with 3 out of phase? c: Will Sites 1 and 3 fire in phase, with 2 out of phase?)

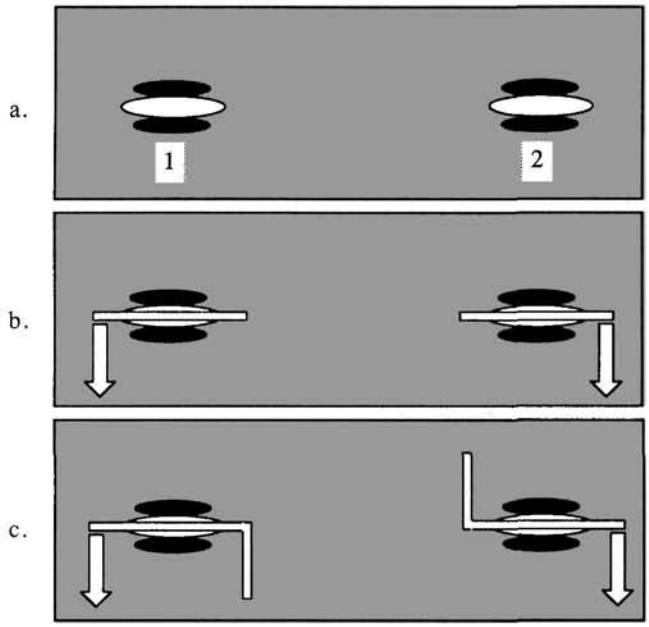


Figure 39. Illustration of a multiple-electrode test of whether intervening vertices can break phase locking in collinear sites. (a: Two receptor sites, 1 and 2, with collinear receptive fields. b: Sites 1 and 2 should fire in phase. c: Will the phase locking between Sites 1 and 2 be broken by the intervening vertices?)

These problems raise the issue of what would be required for temporal synchrony to be a plausible solution to the binding problem for structural description and real-time shape recognition. In the remainder of this section, we offer some considerations and speculations with regard to this issue.

Timing considerations. A primary issue concerns the minimum timing requirements for binding and their compatibility with known neurophysiology. For the purposes of JIM, we have assumed that FELs operate with infinite speed, allowing two cells to synchronize instantly. Realistically, there will be some time cost associated with the propagation of an enabling signal across a FEL (unless FELs are implemented as electrical gap junction synapses). How will this time cost manifest itself in the imposition of synchrony on cells? Two considerations bear on this question. First, strict synchrony is not required for phase relations to carry binding information. Rather, it is only necessary that the phase lag between separate groups of cells be detectably and reliably greater than the phase lag between cells within a group. Second, Konig and Schillen (1991) present simulations demonstrating that some classes of coupled oscillators can phase lock (with zero phase lag) provided the transmission time of the coupling signal is one third or less of the oscillation period. Thus, it is possible to achieve tight synchrony even with nonzero transmission times.

Other constraints on timing derive from the task of shape recognition. For our proposal to be viable, synchrony must be exploitable within the approximately 100 ms that it is estimated (Biederman, 1987b; Thorpe, 1990) the brain requires to activate a representation of an object from the first appearance of a stimulus. Thorpe estimated that 10 synapses must be traversed from the retina to inferior temporal cortex (or IT, where the final object representation presumably resides), leaving 10 ms per layer of computation, just enough time to generate one spike. At first, these estimates seem to preclude the use of temporal binding for recognition. However, because cells in Layer L do not need to wait for a spike to arrive at Layer L + 1 to spike a second time, additional cycles (spikes) would take approximately 10 ms each. (If the first spike from a given neuron in Layer L arrives at IT at time t, then a second spike from that neuron would arrive at time $t + 10$ ms.) Thus, two cycles would take not 200 ms but 110 ms. To exploit temporal binding would likely require multiple assemblies (such as GFAs) firing out of phase with one another within the period between t and $t + 10$ ms. Satisfying this temporal constraint would require a very rapid mechanism for establishing synchrony among the cells in an assembly.

Are fast enabling links necessary? Synchronized oscillations can easily be produced in neural networks without positing specialized connections for that purpose (e.g., see Atiya & Baldi, 1989; Grossberg, 1973). Indeed, most neural network models that exploit phase locking establish the phase locking using standard excitatory–inhibitory interactions among the cells. However, the independence of FELs and standard excitatory–inhibitory connections in JIM has important computational consequences. Specifically, this independence allows JIM to treat the constraints on feature linking (by synchrony) separately from the constraints on property inference (by excitation and inhibition). That is, cells can phase lock without influencing one another's level of activity and vice versa. Although it remains an open question whether a neuroanatomical analog of FELs will be found to exist, we suggest that the distinction between feature linking and property inference is likely to remain an important one.

Compatibility with moving images. The temporal binding mechanism that we have described here was designed for use with stationary images. An important question about such a binding mechanism concerns its compatibility with moving images. The answer to this question will ultimately hinge, at least in part, on the speed of such a mechanism: At what image velocity would a given visual feature, say, an image edge, remain within the receptive field of a neuron in V1 or V2 of visual cortex (where we assume the proposed binding mechanism to be operating) for too short a time to allow the binding mechanism to work, and how does this figure compare with the velocity at which human recognition performance begins to fail? Cells in Area V1 have receptive field diameters between 0.5° and 2° of visual angle (Gross, 1973). Let us return to the estimate discussed earlier for the duration of a binding cycle (the time between successive spikes) and assume that an image property must remain in the receptive field of a cell for 10 ms for the binding mechanism to work reliably. To move in and out of the receptive field of a V1 cell with a receptive field of 0.5° in under a single cycle of this duration (thereby disrupting binding), an image would have to be translated at a velocity of 50°/s. Thus, even if we relax the constraints on the speed of binding to conform with the 50 Hz oscillation (20 ms per cycle) reported by Gray et al. (1989), the binding mechanism would be expected to be robust with translations up to 25°/s. Although shape from motion is a readily demonstrated phenomenon, we

know of no data on the effect of motion on recognition performance of images that can be readily recognized when stationary.

Accidental Synchrony

As noted previously, cells on separate FEL chains will fire in synchrony if their output refractories happen to go below threshold at the same time. When this type of accidental synchrony occurs between two cells, all the members of their respective FEL chains will also fire in synchrony, resulting in a binding error. As demonstrated in JIM's performance, such binding errors can have devastating effects on recognition. Currently, JIM is equipped with no provisions for preventing accidental synchrony. As such, it is at the mercy of stimulus complexity: The probability of JIM's suffering an accidental synchrony increases with the number of cells in L1 and L2 that are active. Were JIM required to deal with scenes of realistic complexity, it would be incapable of keeping the separate components of those scenes from accidentally synchronizing with one another. Although synchrony provides a solution to the dynamic binding problem, it is subject to an intrinsic capacity limitation. The value of this limit is proportional to the duration of the period between firings (e.g., spikes or bursts of spikes) divided by the duration of a firing. Note that this limit does not refer to the number of things that can be bound together. There is no a priori reason to expect a limit on the number of cells in an assembly or a FEL chain. Rather, it is a limit on the number of things that can be simultaneously active without being bound together.

To deal with the complexity of the natural visual environment, a biological system that used synchrony to perform dynamic binding would need mechanisms for reducing the likelihood of accidental synchrony. It is tempting to speculate that visual attention may serve this purpose (Hummel & Biederman, 1991). Although we shall defer detailed consideration of such an account of visual attention to a later article, two implications are worthy of note here. First, such an account posits a role for visual attention that is the opposite of that proposed in traditional spotlight models (e.g., Kahneman & Treisman, 1984; Treisman, 1982; Treisman & Gelade, 1980). The spotlight model holds that visual attention serves to actively bind visual attributes together. By contrast, the limitations of synchrony as a solution to dynamic binding suggest that attention may be required to keep independent visual attributes separate. Second, it would follow from such an account that attention should serve to inhibit activity in unattended regions rather than enhance activity in attended ones: A straightforward way to reduce the likelihood that irrelevant visual features will accidentally fire in synchrony with those currently of interest is to inhibit them. The suggestion that attention should inhibit unattended regions is supported by single unit recordings in the macaque by Moran and Desimone (1985). They showed that a stimulus within the receptive field of a V4 neuron would fail to elicit a response from that neuron if the animal was not attending (though maintaining fixation) to a region of the visual field within the neuron's receptive field.

Additional Implications of Fast Enabling Link Chains

JIM was developed to model the fundamental invariances of human visual recognition, and the simulations reported here were designed to evaluate its capacity to do so. However, there is a variety of other findings for which JIM also provides a natural account. Although we have run no formal simulations of them, this section briefly presents some of these findings. This section is included primarily to suggest additional implications of the model and to underscore the generality of its fundamental principles. More formal treatment of these findings is warranted but beyond the scope of this article.

One critical principle underlying JIM is that binding occurs automatically for features at different locations in the visual field, provided they meet the conditions embodied in the FELs. This principle, and the specific FELs that implement it, provide a natural account of data recently reported by Donnelly, Humphreys, and Riddoch (1991). These investigators recorded reaction times (RTs) for the detection of an inward pointing L vertex among a configuration of three, four, or five outward pointing distractor vertices that formed all but one of the corners of a quadrilateral, pentagon, or hexagon, respectively (Figure 40). When the target was absent it was replaced by another outward pointing vertex in such a position as to add the last vertex to the shape. Examples from target absent and target present trials are shown in Figure 40, left and right columns, respectively.

Because the endpoints of the distractor vertices were collinear, these vertices would be synchronized by the FELs between distant collinear lone terminations (Condition III), producing a single assembly containing all the distractors. If RT is assumed to be a positive function of the number of independent assemblies in a stimulus (rather than the number of vertices), then no increase in RTs would be expected with the increase in the number of distractors. This is precisely what Donnelly et al. (1991) found. The assumption that subjects were sensitive to the number of assemblies rather than the number of vertices is further evidenced by the shorter RTs for the target absent responses (there would be two assemblies in the target present condition, one for the target and one for the distractor, and only one in the target absent condition).

By contrast, consider Figure 41, which shows another condition studied by Donnelly et al. (1991); target absent trials are shown in the left column and target present in the right. These stimuli were similar to those described earlier, except that the distractor vertices were pointing inward and the target outward. Note that with these stimuli, the lone terminators of the separate vertices no longer satisfy the conditions for grouping defined in Condition III, as they would be collinear only by passing through a vertex. As such, JIM would not bind these vertices, and there would be one assembly per vertex. By the assumption that RT is a positive function of the number of assemblies, JIM would predict an increase in search times as a function of the number of distractors in this condition. Again, this is what Donnelly et al. (1991) reported: RTs for detecting a target vertex increased linearly with the number of distractor vertices.

The principle of not grouping collinear edges through vertices also provides an account of the well-known demonstra-

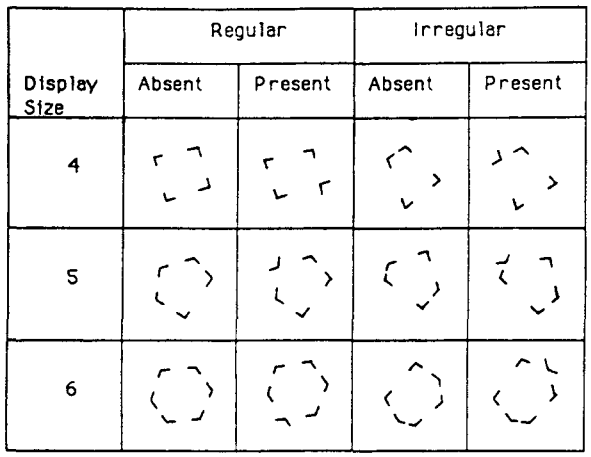


Figure 40. Displays for a search task in which the target was an inward-pointing vertex and the terminators of the distractors were collinear without passing through a vertex. Search times were unaffected by the number of vertices. (From "Parallel Computation of Primitive Shape Descriptions" by N. Donnelly, G. W. Humphreys, and M. J. Riddoch, 1991, *Journal of Experimental Psychology: Human Perception and Performance, 17,* Figure 1, p. 563. Copyright 1991 by the American Psychological Association. Reprinted by permission.)

tions of Bregman (1981) and Leeper (1935; Hummel & Biederman, 1991). Bregman's demonstration is shown in Figure 42 (Panels a and b). It is difficult to form a coherent percept of the image in Panel a. However, when the same elements are presented in the presence of what appears to be an occluding surface (Panel b), it is easy to see that the figure depicts an array of *B*s. Bregman interpreted this effect in terms of the cues for grouping provided by the occluding surface. Blickle (1989) proposed a nonocclusional account of this demonstration. He noted that accidental L vertices were produced where the occluder was removed in the nonoccluded image (Figure 42a) and hypothesized that they may have prevented the elements from grouping. When the L vertices are removed, the elements once again form a coherent percept even in the absence of an explicit occluder (Figure 42c). Blickle applied the same analysis toward understanding why Leeper's (1935) figures, such as the elephant in Figure 42d, were so difficult to identify. Blickle argued that the L vertices prevent the visual system's grouping the separate relevant contours of the fragments into a common object. Indeed, when the L vertices are removed, the elements become much easier to group (Figure 42e). The role played by the L vertices in these demonstrations is precisely what would be expected based upon JIM's account of grouping.

A similar analysis applies to Biederman's (1987b) study on

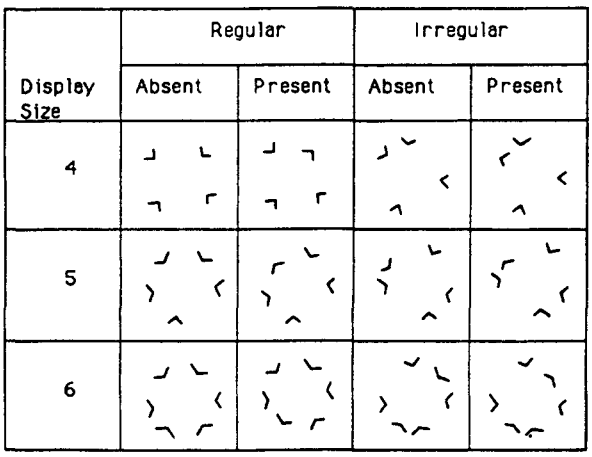


Figure 41. Displays for a search task in which the target was an outward-pointing vertex and the terminators of the distractors were collinear only through a vertex. Search times increased with the number of vertices. (From "Parallel Computation of Primitive Shape Descriptions" by N. Donnelly, G. W. Humphreys, and M. J. Riddoch, 1991, *Journal of Experimental Psychology: Human Perception and Performance, 17,* Figure 3, p. 565. Copyright 1991 by The American Psychological Association. Reprinted by permission.)

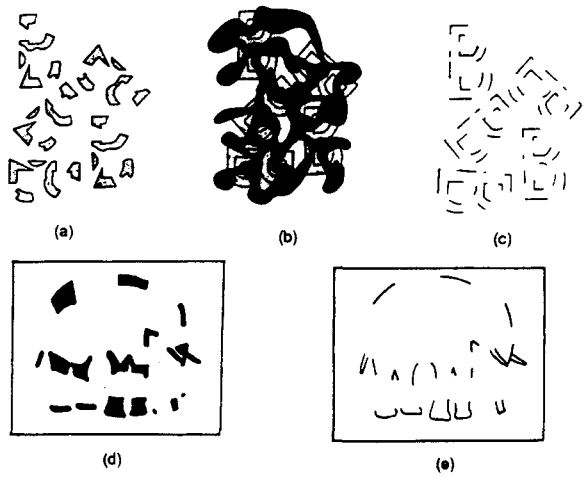


Figure 42. Two demonstrations of the inhibition of collinear grouping of end-stopped segments by L vertices. (a: These image fragments [Bregman, 1982] are not identifiable as familiar forms. b: Bregman showed that with the addition of an occluder, the *B*s become apparent. Blickle (1989) noted that the image fragments in Bregman's non-occluded image (Panel a) contained accidental L vertices, formed where contour from the occluder was left in the figure, that impede the fragments' grouping into objects. When the L vertices are removed, as in Panel c (Blickle, 1989) the *B*s readily appear, even in the absence of the occluder. Blickle performed a similar analysis of the difficulty in recognizing the Leeper (1935) figures. d: Leeper's elephant. e: Like Bregman's *B*s, Blickle's removal of the accidental L vertices made the objects' shape more apparent. (Panels a and b are from "Asking the 'What for' Question in Auditory Perception," pp. 106–107, by A. S. Bregman, 1982, in M. Kubovy and J. R. Pomerantz, *Perceptual Organization,* Hillsdale, NJ: Erlbaum. Copyright 1982 by Erlbaum. Adapted by permission. Panel d is from "A Study of a Neglected Portion of the Field of Learning: The Development of Sensory Organization" by R. Leeper, 1935, *Journal of Genetic Psychology, 46,* Figure 2, p. 50. Reprinted with permission of the Helen Dwight Reid Educational Foundation. Published by Heldref Publications, 1319 18th Street, NW, Washington, DC 20036-1802. Copyright 1935. Panels c and e are from "Recognition of Contour Deleted Images" by T. W. Blickle, 1989, p. 20, Unpublished doctoral dissertation, State University of New York at Buffalo. Adapted by permission.)

the effects of amount and locus of contour deletion on recognition performance. Error rates and RTs increased with increases in the amount of contour deleted from an image, but for a constant amount of contour deletion, performance suffered more when the contour was removed from the vertices than when it was removed at midsegment. Recall that JIM groups the contours comprising a geon using the vertices produced where those contours coterminate. As such, it is unable to group contours into geons when all the vertices have been removed and predicts that recognition should fail. By contrast, when contour is removed at midsegment, the vertices can still be grouped by means of the FELs between collinear distant lone terminations. Although midsegment deletion partially removes the axes of symmetry in a geon's image, the remaining features can still be grouped, and recognition should be possible, though slowed. Thus, JIM predicts the difference between performance with vertex-deleted and midsegment-deleted stimuli. JIM cannot currently predict that recognition is possible at all in the vertex-deleted condition. This failure likely reflects JIM's current inability to exploit information at different spatial scales. It is possible that other sources of information, such as an object's axis structure, can be used for classification even when the edge contour has been removed near the points where the edges coterminate.

The final effect we shall note here was observed by Malcus (1982). Malcus found that extraneous contour introduced into an image impeded recognition performance more when it passed through the vertices in the image than when it passed through edge midsegments. If an extraneous contour crosses a contour of the original image at midsegment, an X vertex will be produced. Recall that JIM does not group contours that meet at vertices with more than three prongs. Contours forming X vertices (containing four prongs) will therefore remain independent; that is, the geon to which the original contour belongs will not fire in synchrony with the extraneous contour and will therefore be processed normally in JIM's third layer. By contrast, if an extraneous contour is passed through a vertex in the image of a geon, that vertex will no longer be detected in the model's second layer (e.g., a three-pronged vertex with an extra contour through it becomes a five-pronged vertex) and will not be available for grouping the contours that actually belong to the geon. If enough vertices are disrupted in this manner, the geon will be unrecoverable.

Implications of Single Place-Predicate Relations of JIM

In this section, we describe two effects in human visual recognition predicted by JIM's treatment of relations. JIM represents the relations among objects as single-place predicates, that is, predicates that take only one argument. For example, the representation of above specifies its subject (i.e., of which geon "aboveness" is true) by firing in synchrony with the other attributes of that geon, but it does not specify its object (which geon the subject is above). Thus, the representation of Geon A in Figure 43 would specify that it is above another geon, but not whether it is above B or above C. The object of the relation *above* is specified only implicitly as the subject of *below*: That A is above B can be determined only because *above* fires in synchrony with A and *below* with B. As we have argued earlier, implicit specification of a property (a relation, a binding, or, in this case, a case role assignment) is prone to serious weaknesses. Here, the weakness is a propensity for confusions when multiple geons in an object share the same relations.

Consider, for example, a totem pole-like object in which multiple geons are stacked vertically, as shown in Figure 44. For totem poles of two or three geons, each geon will be described by a unique combination of above and below (Figure 44a), so there should be little possibility for confusion between one totem pole and the same geons stacked in a different order. However, when a fourth geon is added, the two central geons will share the same above–below description; that is, both geons will be bound to both *above* and *below*, as shown in Figure 44b. As such, they could switch places and the representation of the totem pole would not change (Figure 44c). JIM thus predicts greater confusability for totem poles with four or more geons than for totem poles with two or three geons. Confusability will likely increase with the number of geons, but it should jump markedly between three and four. Also, confusability should be greater when geons in the center of the stack change positions than for changes involving geons on either end.

JIM makes similar predictions for horizontal arrays of geons (Figure 45) except that, for an equivalent number of geons, confusability should be higher in horizontal arrays than for vertical arrays. This prediction derives from JIM's not discriminating left–right relations, rather both are coded simply as *beside*. As such, all the geons in a horizontal array will have the beside relation, with their respective serial positions in the array unspecified.

Limitations of the Current Implementation

The current implementation of JIM is limited in a number of respects. Some of these limitations derive from simplifying assumptions we have made to keep the implementation manageable, and others reflect fundamental questions that remain to be addressed. We shall focus the discussion primarily on the latter.

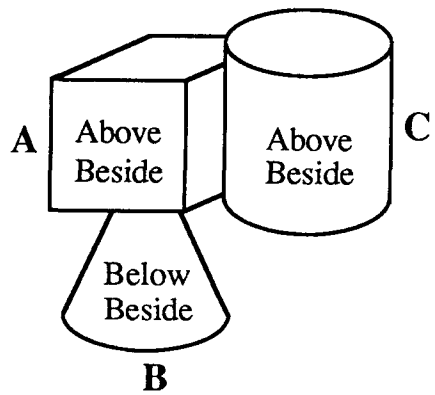


Figure 43. JIM would represent Geon A as above something, but it would not specify that it is above Geon B rather than Geon C.

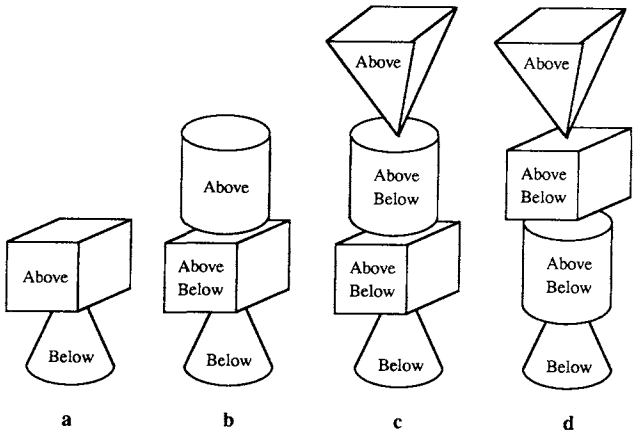


Figure 44. Example stimuli in the proposed "totem pole" experiments. (a: For totem poles with two or three geons, each geon is described by a unique combination of *above* and *below.* The geons in these totem poles could not change places without changing their representation in JIM's third and fifth layers. b: The middle geons in a four-geon totem pole have the same above–below description. c: The middle geons in a four-geon totem pole can change places without changing their representation in Layers 3 and 5.)

Limitations of the Binding Mechanism

For many images, the set of FELs posited is sufficient to group local features into geons while allowing the features of separate geons to remain independent, as discussed earlier. Some images JIM cannot parse were presented earlier (Figures 18 and 19). These shortcomings may constitute empirical predictions in that human subjects may evidence greater difficulty recognizing images that JIM cannot segment than those that it can. However, it is clear that the degree of difficulty JIM would demonstrate with such images is unrealistically great.

To remedy these difficulties, at least two major extensions to the current grouping mechanism will likely be required. First, the FELs need to be generalized to allow cells to actively resist firing in phase. Such a mechanism could help the model deal with accidental alignments and parts that meet at L vertices by allowing many desynchronizing effects to overcome the synchronizing effects that occur at the points where separate geons meet. Naturally, the conditions under which image features should resist synchrony would have to be specified. The second extension that would improve the model's capacity for feature grouping is to provide contingencies for global constraints for grouping. For example, axes of symmetry could be used for grouping, as described in Mohan (1989), or convex image patches, as in Vaina and Zlateva (1990). However, how such grouping constraints could be plausibly implemented in a neural network architecture remains an open question.

Other Limitations of the Current Implementation

Several other extensions suggest themselves for exploration as well. Among the first is the compatibility of the proposed constraints for parsing and recognition with natural images. In this implementation, we have assumed that recognition starts with a clean representation of the surface and depth discontinuities describing a line drawing of an object. Implicitly, we have also assumed that this representation could be achieved without top-down influences from geons or objects. However, deriving a representation of orientation and depth discontinuities from natural images is clearly a nontrivial problem, and it may be the case that it cannot be solved without top-down mediation. These considerations motivate two lines of exploration. The first is the compatibility of our FELs with output from available edge-detection algorithms, and the second is the use of FELs and geons to interactively derive depth discontinuities from natural images. In particular, edge definition may be facilitated by constraints from geons (Biederman, 1987b).

JIM's capacity for representing shape is most limited by its impoverished vocabulary of geons and relations. JIM is capable of discriminating eight geon types, whereas RBC posits 24. The most important relations omitted from JIM's current vocabulary are the connectedness relations, which describe how the various geons in an image are connected to one another. The simplest of these is the binary relation *connected* versus *not connected.* JIM does not discriminate two geons that are physically joined from two that are not. For example, its representation of a table would express only the relative angles, locations, and sizes of the various parts (such as the top and the legs), neglecting that the top is connected to each of the legs but that the legs do not touch one another. Other connectedness relations include whether two geons are joined end to end (like the cylinders of a flashlight) or end to side (like the join between a camera body and lens) and centeredness (whether one geon connects to another near the center of the latter's side or near an edge).

Expanding the current architecture to capture more relational and shape attributes will require additional structures. In particular, it is not clear that the current architecture in L4 and L5 could be applied directly to the problem of deriving connectedness relations. However, given that architectures capable of

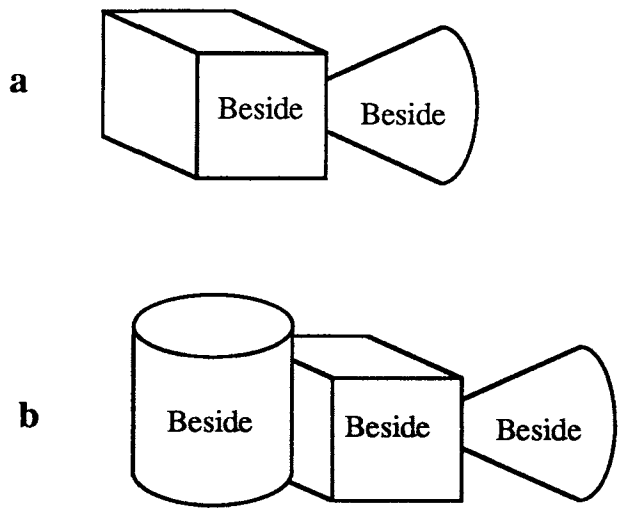


Figure 45. All geons in a horizontal array have the same relative position description: They are all described as beside something.

deriving them can be described, new attributes can be added to JIM's representation at a cost of one cell per attribute. None of the additional properties seem incompatible with the existing structure. Therefore, there is no reason to think that expanding the model to capture them will require violating any important assumptions underlying its current design.

Among JIM's most serious weaknesses is an inability to deal with multiple objects in an image. Expanding it to deal with multiple objects will almost certainly entail addressing questions of scale, visual attention, and additional problems in grouping such as figure–ground segmentation. Although we do not expect these extensions to be straightforward, we also do not expect that they will require abandoning the basic tenets underlying JIM's current design. If we regard JIM's current domain as a subset of the processing that occurs within the focus of attention, then its failure with multiple objects is not psychologically unrealistic. Biederman, Blickle, Teitelbaum, and Klatsky (1988) found that search time for a target object in a nonscene display (objects were arrayed in a circular arrangement, as the numbers on the face of a clock) was a linear function of the number of distractor objects in the display, suggesting that subjects were attending to one object at a time. This finding seems anomalous in the context of the finding that complete scenes can be recognized in the same time it takes to recognize a single object (Biederman, 1981). However, Mezzanotte (1981) demonstrated that scene recognition does not require that the individual objects in the scene be identifiable in isolation. Rather, as argued by Biederman (1988), familiar groups of interacting objects may be treated by the visual system as a single object, that is, as configurations of geons in particular relations.

Summary and Conclusions

We have described a neural net architecture that takes as input a retinotopically mapped representation of the surface and depth discontinuities in an object's image and activates a representation of the object that is invariant with translation and scale and largely invariant with viewpoint. JIM's behavior conforms well to empirical data on human object recognition performance. The fundamental principle underlying its design is that an object is represented as a structural description specifying its parts and the relations among them. This design principle frees JIM from trading off attribute structures for an implicit representation of the relations among those attributes. Also, it permits shape representation to be achieved with a remarkably small number of units. JIM's capacity for structural description derives from its solution to the dynamic binding problem. Dynamic binding is thus critical for shape representation, but it is subject to intrinsic capacity limitations. In the case of binding through synchrony, the limits derive from the temporal parameters of cells and the links among them. We speculate that observed limitations on visual attention in human subjects may reflect the limits of a natural dynamic binding mechanism.

References

Abeles, M. (1982). *Local cortical circuits: Studies of brain function* (Vol. 6), New York: Springer.

Atiya, A., & Baldi, P. (1989). Oscillations and synchronizations in neural networks: An exploration of the labeling hypothesis. *International Journal of Neural Systems, 1,* 103–124.

Baldi, P., & Meir, R. (1990). Computing with arrays of coupled oscillators: An application to pre-attentive texture discrimination. *Neural Computation, 2,* 459–471.

Biederman, I. (1981). On the semantics of a glance at a scene. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 213–263). Hillsdale, NJ: Erlbaum.

Biederman, I. (1987a). Matching image edges to object memory. *Proceedings of the First International Conference on Computer Vision* (pp. 384–392). Washington, DC: IEEE.

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

Biederman, I. (1988). Aspects and extensions of a theory of human image understanding. In Z. Pylyshyn (Ed.), *Computational processes in human vision* (pp. 370–428). Norwood, NJ: Ablex.

Biederman, I., Blickle, T., Teitelbaum, R., & Klatsky, G. (1988). Object search in nonscene displays. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 456–467.

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 human shape recognition. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 121–133.

Biederman, I., & Hilton, H. J. (1991). Metric versus nonaccidental determinants of object priming, Manuscript in preparation.

Blickle, T. W. (1989). *Recognition of contour deleted images.* Unpublished doctoral dissertation, State University of New York at Buffalo.

Bower, J. (1991, April). Oscillations in cerebral cortex: The mechanisms underlying an epiphenomenon. Paper presented at the Workshop on Neural Oscillations. Tucson, AZ.

Brady, M. (1983). Criteria for representations of shape. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and machine vision.* (pp. 39–84). San Diego, CA: Academic Press.

Brady, M., & Asada, H. (1984). Smoothed local symmetries and their implementation. *International Journal of Robotics Research, 3,* 36–61.

Bregman, A. S. (1981). Asking the "what for" question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 99–118). Hillsdale, NJ: Erlbaum.

Crick, F. H. C. (1984). The function of the thalamic reticular spotlight: The searchlight hypothesis. *Proceedings of the National Academy of Sciences, 81,* 4586–4590.

Crick, F. H. C., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in Neuroscience, 2,* 263–275.

Donnelly, N., Humphreys, G. W., & Riddoch, M. J. (1991). Parallel computation of primitive shape descriptions. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 561–570.

Eckhorn, R., Bauer, R., Jordan, W., Brish, M., Kruse, W., Munk, M., & Reitboeck, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? Multiple electrode and correlation analysis in the cat. *Biological Cybernetics, 60,* 121–130.

Eckhorn, R., Reitboeck, H., Arndt, M., & Dicke, P. (1990). Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex. *Neural Computation, 2,* 293–307.

Edelman, S., & Poggio, T. (1990, April). Bringing the grandmother back into the picture: A memory-based view of object recognition. MIT Artificial Intelligence Memo No. 1181.

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

Engel, A. K., Konig, P., Kreiter, A. K., & Singer, W. (1991). Interhemispheric synchronization of oscillatory neuronal responses in cat visual cortex. *Science, 252,* 1177–1179.

Feldman, J. A. (1982). Dynamic connections in neural networks. *Biological Cybernetics, 46,* 27–39.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6,* 205–254.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 1–71.

Gerhardstein, P. C., & Biederman, I. (1991, May). *Priming depth-rotated object images: Evidence for 3D invariance.* Paper presented at the Meeting of the Association for Research in Vision and Ophthalmology, Sarasota, FL.

Gray, C. M., Konig, P., Engel, A. E., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-column synchronization which reflects global stimulus properties. *Nature, 338,* 334–337.

Gray, C. M., & Singer, W. (1989). Stimulus specific neuronal oscillations in orientation columns of cat visual cortex. *Proceedings of the National Academy of Sciences, 86,* 1698–1702.

Gross, C. G. (1973). Visual functions of inferotemporal cortex. In R. Jung (Ed.), *Handbook of sensory physiology: Vol. VII/3. Central processing of information. Part B. Visual centers in the brain.* (pp. 451–482). Berlin: Springer-Verlag.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, 52,* 217–257.

Grossberg, S., & Somers, D. (1991). Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks, 4,* 453–466.

Guzman, A. (1971). Analysis of curved line drawings using context and global information. *Machine Intelligence, 6,* 325–375. Edinburgh, Scotland: Edinburgh Press.

Hinton, G. E. (1981). A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the 7th International Joint Conference on Artificial Intelligence.* Symposium conducted at the University of British Columbia, Vancouver, British Columbia, Canada.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: Foundations* (pp. 77–109). Cambridge, MA: MIT Press/Bradford Books.

Hoffman, D. D., & Richards, W. A. (1985). Parts of recognition. *Cognition, 18,* 65–96.

Hummel, J. E., & Biederman, I. (1990a). Dynamic binding: A basis for the representation of shape by neural networks. In M. P. Palmarini, *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 614–621). Hillsdale, NJ: Erlbaum.

Hummel, J. E., & Biederman, I. (1990b, November). *Binding invariant shape descriptors for object recognition: A neural net implementation.* Paper presented at the 21st Annual Meeting of the Psychonomics Society, New Orleans, LA.

Hummel, J. E., & Biederman, I. (1991, May). *Binding by phase locked neural activity: Implications for a theory of visual attention.* Paper presented at the Annual Meeting of the Association for Research in Vision and Ophthalmology, Sarasota, FL.

Hummel, J. E., Biederman, I., Gerhardstein, P. C., & Hilton, H. J. (1988). From image edges to geons: A connectionist approach. In D. Touretsky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 462–471), San Mateo, CA: Morgan Kaufman.

Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition, 13,* 289–303.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman, R. Davies, & J. Beatty (Eds.), *Varieties of attention* (pp. 29–62). San Diego, CA: Academic Press.

Keele, S. W., Cohen, A., Ivry, R., Liotti, M., & Yee, P. (1988). Test of a temporal theory of attentional binding. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 444–452.

Konig, P., & Schillen, T. B. (1991). Stimulus-dependent assembly formation of oscillatory responses: I. Synchronization. *Neural Computation, 3,* 167–178.

Lange, T., & Dyer, M. (1989). *High-level inferencing in a connectionist neural network.* (Tech. Rep. No. UCLA-AI-89-12), Los Angeles: University of California, Computer Science Department.

Leeper, R. (1935). A study of a neglected portion of the field of learning: The development of sensory organization. *Journal of Genetic Psychology, 46,* 41–75.

Lindsay, P. H., & Norman, D. A. (1977). *Human information processing: An introduction to psychology* (2nd ed.). San Diego, CA: Academic Press.

Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision, 1,* 57–72.

Malcus, L. (1982). *Contour formation, segmentation, and semantic access in scene and object perception.* Unpublished doctoral dissertation, State University of New York at Buffalo.

Malik, J. (1987). Interpreting line drawings of curved objects. *International Journal of Computer Vision, 1,* 73–103.

Marshall, J. A. (1990). A self-organizing scale-sensitive neural network. *Proceedings of the International Joint Conference on Neural Networks, 3,* 649–654.

Mezzanotte, R. J. (1981). *Accessing visual schemata: Mechanisms invoking world knowledge in the identification of objects in scenes,* Unpublished doctoral dissertation, State University of New York at Buffalo, Department of Psychology.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review, 81,* 521–535.

Mohan, R. (1989). *Perceptual organization for computer vision.* Unpublished doctoral dissertation, University of Southern California, Department of Computer Science.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229,* 782–784.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Pinker, S. (1986). Visual cognition: An introduction. In S. Pinker (Ed.), *Visual cognition* (pp. 1–63). Cambridge, MA: MIT Press.

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

Sejnowski, T. J. (1986). Open questions about computation in cerebral cortex. In J. L. McClelland, & D. E. Rumelhart. (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 372–389). Cambridge, MA: MIT Press.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. *Symposium on the Mechanism of Thought Processes.* London: Her Majesty's Stationery Office.

Selfridge, O. G., & Neisser, U. (1960). Pattern recognition by machine. *Scientific American, 203,* 60–68.

Shastri, L., & Ajjanagadde, V. (1990). *From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings* (Tech. Rep. No. MS-CIS-90-05). Philadelphia: University of Pennsylvania, Department of Computer and Information Sciences.

Smolensky, P. (1987). *On variable binding and the representation of symbolic structures in connectionist systems* (Internal Rep. No. CU-CS-355-87). Boulder, CO: University of Colorado, Department of Computer Science & Institute of Cognitive Science.

Strong, G. W., & Whitehead, B. A. (1989). A solution to the tag-assign-

ment problem for neural networks. *Behavioral and Brain Sciences, 12,* 381–433.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation dependence in shape recognition. *Cognitive Psychology, 21,* 233–283.

Thorpe, S. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. In R. Eckmiller, G. Hartmann, & G. Hauske (Eds.), *Parallel processing in neural systems and computers* (pp. 91–94). Amsterdam: North-Holland.

Treisman, A. (1982). Perceptual grouping and attention in visual search for objects. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 194–214.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* 97–136.

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

Vaina, L. M., & Zlateva, S. D. (1990). The largest convex patches: A boundary-based method for obtaining object parts. *Biological Cybernetics, 62,* 225–236.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal Rep. No. 81-2). Gottingen, Germany: Max-Plank-Institute for Biophysical Chemistry, Department of Neurobiology.

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. 411–432). New York: Wiley.

Waltz, D. (1975). Generating semantic descriptions from drawings of scenes with shadows. In P. Winston (Ed.), *The psychology of computer vision* (pp. 19–91). New York: McGraw-Hill.

Wang, D., Buhmann, J., & von der Malsburg, C. (1991). Pattern segmentation in associative memory. *Neural Computation, 2,* 94–106.

Winston, P. (1975). Learning structural descriptions from examples. In P. Winston (Ed.), *The psychology of computer vision* (pp. 157–209). New York: McGraw-Hill.

Appendix

Response Metrics

Max_i represents the highest activation value (A_i) achieved by a given target cell i during a run of N time slices. $Mean_i$ is calculated as the arithmetic mean of the target cell's activation (A_i) over n time slices. Max and mean provide raw estimates of the recognizability of an object in a particular condition (i.e., the match between the structural description activated in response to the image of an object in a given condition and the structural description used for familiarization with that object). Because they consider the activation of the target cell only, max and mean will tend to be misleading if all object cells achieve high activations in response to every image.

P is a response metric designed to overcome this difficulty. P_i for a given target cell i on a given run is calculated as the target cell $Mean_i$ divided by the sum of all above-zero object cell mean activations:

$$P_i = [Mean_i/(Mean_i + \Sigma_j Mean_j)]^+, \quad Mean_j > 0,$$

where j corresponds to a nontarget object cell. This metric provides a measure of the discriminability of the target object from the population of nontargets given a particular image. Although it is not subject to the same criticism as max and mean, P is a less sensitive metric when the mean activation of nontarget object cells is low. For example, assume that the baseline view of a given object produces a mean target cell activation of 0.50 (with all nontarget object cell means below zero), and another view produces a mean of 0.01 (nontarget means below zero). With these values, P would provide the misleading impression that the model had performed identically with the two views: 0.50/(0.50 + 0 + . . . 0) = 0.01/(0.01 + 0 + . . . 0). An additional difficulty with P as a metric is that it is sensitive only to differences among object cells with mean activations above zero.

The final response metric, MP, is designed to reflect both the raw recognizability of a view of an object and its discriminability from the other objects in the set. MP_i is calculated as the product of $Mean_i$ and P_i. This metric suffers the same insensitivity to negative numbers as does P, but is prone neither to P's tendency to mask large differences between conditions in the face of inactive nontargets, nor to the tendency of max and mean to mask indiscriminate responding.

Received July 30, 1990
Revision received July 30, 1991
Accepted August 12, 1991 ■