Dynamic Binding: A Basis for the Representation of Shape by Neural Networks[1]

John E. Hummel and Irving Biederman

Department of Psychology, University of Minnesota, Minneapolis, MN 55455

Abstract

A neural network model for object recognition based on Biederman's (1987) theory of Recognition by Components (RBC) is described. RBC assumes that objects are recognized as configurations of simple volumetric primitives called *geons*. The model takes a representation of the edges in an object as input and, as output, activates an invariant, entry-level representation of the object that specifies the object's component geons and their interrelations. Local configurations of image edges first activate cells representing local viewpoint-invariant properties (*VIPs*), such as vertices and 2-D axes of parallelism and symmetry. Once activated, VIPs are bound into sets through temporal synchrony in the firing patterns of cells representing the VIPs and image edges belonging to a common geon. The synchrony is established by a mechanism which operates only between pairs of a) collinear, b) parallel, and c) coterminating edge and VIP cells. This design for perceptual organization through temporal synchrony is a major contribution of the model. A geon's bound VIPs activate independent representations of the geon's major axis and cross section, location in the visual field, aspect ratio, size, and orientation in 3-space. The relations among the geons in an image are then computed from the representations of the geons' locations, scales and orientations. The independent specification of geon properties and interrelations uses representational resources efficiently and yields a representation that is completely invariant with translation and scale and largely invariant with viewpoint. In the final layers of the model, this representation is used to activate cells that, through self-organization, learn to respond to individual objects

Introduction

Within the limits of visual resolution, and excluding so-called "accidental" viewpoints (i.e., singularities of viewing angle that project misleading images on the retina, such as viewing a cylinder from an angle that makes it appear to be a rectangle), an object's image may be projected on the retina in any location, in any size, and from any viewing angle, and the object will still be readily recognized. Biederman's (1987) theory of Recognition by Components (RBC) explains these fundamental phenomena of object recognition by positing that objects are represented as structured configurations of viewpoint-invariant volumetric primitives called *geons*. This paper introduces a neurally plausible model of object recognition based upon RBC. Although the model described herein is a complete model of recognition, this paper describes only those parts explicitly concerned with the deriviation of an object's structural description in terms of geons and their relations.

To derive a viewpoint-invariant representation of the geons and relations in an image of an object, a neural network (NN) must solve three related problems: (1) For any image

---

[1]This research was supported by AFOSR Research Grant 88-0231 to I.B. and an NSF Graduate Fellowship to J.E.H. Correspondences should be addressed to J.E.H. (Bitnet: EQZ6628@UMNACVX) or I.B. (Bitnet: PSYIRV@UMNACVX)

containing more than one geon, it must determine which image features belong with which geons; (2) it must recognize the geons and represent them in a manner that, while invariant with location and viewpoint, expresses the location and orientation of each geon; and (3), it must derive the relations among the different geons in the image and bind those relations to the geons to which they apply. These tasks are all manifestations of the *Binding Problem*, a problem that has not been adequately handled by artificial neural networks. We describe a solution to binding which allows the present model to solve each of these problems.

The Binding Problem The term *binding* refers to the represention of feature conjunctions. For example, how can a NN represent an image edge that is at a particular location in the image *and* at a particular orientation *and* with a particular curvature, etc.? The predominant approach is to allocate a cell (or specific pattern of activity over a set of cells) to respond to edges with the desired combination of properties. Likewise, other cells or patterns would be allocated to respond to all other combinations. We will use the term *enumerated* to refer to representations of this type because feature conjunctions are represented by enumerating all possible combinations and allocating separate cells for each.

Despite its popularity in NNs, enumeration suffers critical shortcomings as a general solution to binding. Its most serious difficulty is that the cells representing a given feature conjunction must be dedicated *prior* to the occurrence of that conjunction in the system's input. In addition to inefficiency of representation (most cells are unused most of the time), this requirement precludes *dynamic* binding. *Dynamic binding* refers to conjoining stimulus properties that are represented in *different* cells or even different parts of the brain. The problem of dynamic binding is typically illustrated in the context of conjoining an object's color and shape, and its solution is usually described in terms of an attentional mechanism that operates by somehow "gluing" together different properties that are linked to a common point in some sort of "location map" (e.g., Kahneman & Treisman, 1984; Treisman & Gelade, 1980).

But the problem of dynamic binding has implications far beyond conjoining color and shape by reference to common location. First, any image projected on the retina will exist over a range of locations, so even assembling the various features defining a shape (Problem 1 above) entails binding features occurring at *different* locations. In this context, binding is referred to as *image parsing* or *perceptual grouping* (although whether the binding is performed dynamically or by pre-dedicated cells is rarely addressed explicitly). Representing geons in a manner invariant with location and viewpoint while still expressing these properties (Problem 2) also requires a mechanism for dynamic binding since, to be invariant with location, the representation of a geon must be independent of the represention of its location. Therefore, conjoining these separate representations to express the location of a particular geon requires dynamic binding. The same logic applies to binding geons and relations (Problem 3). Unless a different cell is to be posited for each possible geon in each possible relation with every other geon, binding geons and relations entails dynamically binding features represented in separate cells. Thus, the dynamic binding problem underlies each of the above difficulties posed by a NN approach to viewpoint-invariant recognition.

Binding Through Synchrony In the present model, independent features are dynamically bound by establishing synchronous firing in the cells representing those features. Although synchrony as a basis for binding was first described by von der Malsburg (1981,1987) and later by others (e.g., Crick, 1984), this article presents an original proposal for establishing synchrony among the basic features of complex shapes. Specifically, we posit the existence of *Fast Enabling Links* (*FELs*) that induce

synchronous firing in active units sharing them. In the model's first two layers, visual features represented by cells sharing FELs are grouped into coherent shapes. The form of the grouping, fundamental to the model's capacity for representing shape, is determined by the specific set of FELs. The resulting synchrony is then used in higher layers both to bind the independent properties of geons, and to bind relations to the geons they describe.

The Model

Overview The complete model is a 7 layer connectionist network that takes as input a representation of a line drawing of an object and, as output, activates a cell representing the entry-level category of the object. An overview of its architecture is shown in Figure 1. The model's first layer (L1) is composed of a mosaic of cell clusters distributed over the model's visual field. The cells within an L1 cluster respond to image edges in terms of their orientation, curvature, and whether they terminate within the cluster's receptive field. The model's second layer (L2) is also composed of a mosiac of cell clusters. These cells respond to configurations of edges that define vertices, 2-D axes of parallelism and symmetry, and oriented, elongated blobs at particular locations in the visual field. Cells in L1 and L2 group themselves into sets (or assemblies) describing geons by establishing temporal synchrony among their spikes of output. Cells tend to fire in synchrony if they represent features of the same geon and out of synchrony if they represent features of different geons.

Cells in L3 respond to properties of complete geons. These cells take their input directly from L2, but because of the binding achieved in L1 and L2, each of the geon properties is represented independently of every other property. For example, the shape of a geon's major axis (straight or curved) is represented in one vector of cells and the geon's location is represented in another vector. Consequently, the representation of the geon's axis does not change when the geon is moved in the visual field. The fourth and fifth layers compute the relations among the geons represented in L3. These computations are performed on the basis of the geons' coarsely coded metric properties (i.e., location in the visual field, scale and orientation in 3-space). The relations among the geons are bound to the geons they describe by the same synchrony of firing that binds image features together for geon recognition. The output of layers 3 and 5 together describe an object in terms of its geons and their relations. This representation is invariant with scale and translation, and largely invariant with viewpoint (orientation in the visual plane and in depth).

The representations and processes employed in layers 1 to 5 constitute the major contribution of this effort, but the model has two additional layers that use the output of layers 3 and 5 as a basis for invariant recognition. These layers integrate the outputs of L3 and L5 over time and self-organize to recognize complete objects. This paper will emphasize the processes employed in L1 and L2 for image parsing and the processes employed in L4 for computing relations.

Image Parsing

Image parsing is among the first problems to confront a geon-based model of visual recognition because the VIPs in an image must be grouped before the geons they define can be identified. For example, correctly parsing the image in Figure 2a entails grouping vertices V1,V2 with segments S1,S2 as features of one geon, and V3 with S3 as features of the other. This model performs parsing by establishing temporal synchrony among the cells in L1 and L2 representing features of a common geon. The synchrony is established through local interactions among edge- and VIP-sensitive cells.

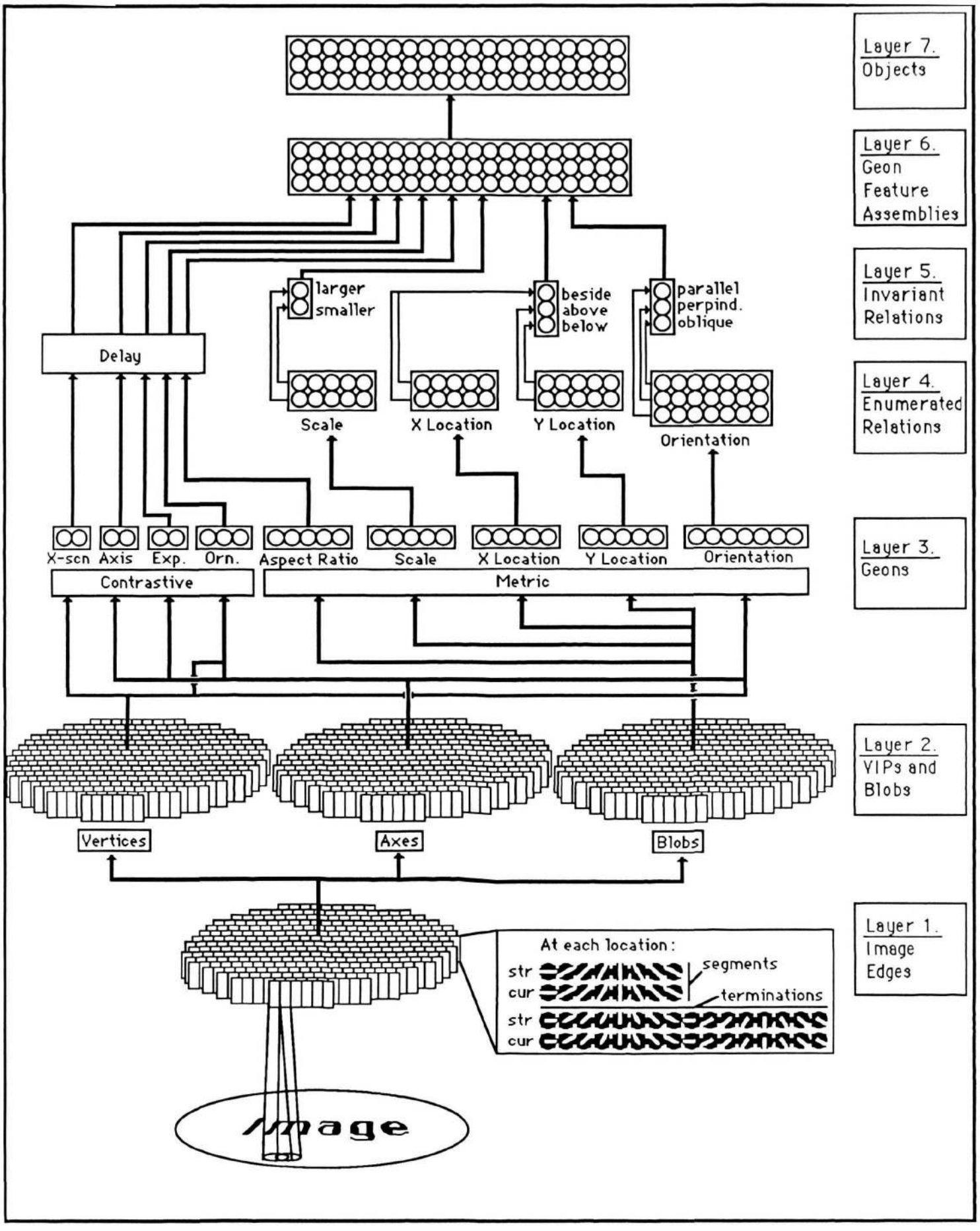

Figure 1. Overall Architecture

Cells in L1 share two types of connections with other cells in the network: typical *activation* connections (or simply *connections*), which spread excitation and inhibition from one cell to another, and *fast enabling links* (*FELs*), which are assumed to operate approximately an order of magnitude faster than the duration of a "time slice" (the temporal period within which cells sum their inputs), and which propagate only binary *enabling* signals between cells. FELs induce synchronous firing in pairs of active cells as follows: Each $cell_i$ in L1 has an output refractory R_i which prevents it from generating a continual train of output spikes. If $cell_i$ is active, it will generate a spike of output only when R_i goes below the refractory threshold (0). R_i is assumed to decay linearly. When $R_i \leq 0$, it is reset to its maximum, and $cell_i$ fires (it generates a spike of output and sends a signal out along all its FELs). Because of the speed with which FELs propagate, an enabling signal will tend to arrive at its destination within the same time slice it was generated. When it arrives at an active $cell_j$, its effect is to set R_j immediately to zero, causing $cell_j$ to go through the same sequence of events as $cell_i$ (i.e., reset its refractory, and generate an output and an enabling signal). If the enabling signal arrives at an inactive cell, nothing happens.

Because active cells sharing FELs tend to fire in synchrony, they organize themselves into groups defined by their temporal firing patterns. As shown in figure 2b, FELs are posited between five types of cell pairs in L1 and L2, reflecting four general constraints on the formation of edge-based images:

1) Image edges usually extend beyond the receptive field of a single edge-sensitive neuron, so a given image edge will tend to excite collinear, adjacent edge-sensitive cells. Such cells are synchronized by FELs *a* and *b* in Figure 2b.
2) The receptive fields of adjacent edge-sensitive cells overlap, so a given image edge will tend to excite parallel, adjacent edge-sensitive cells. These cells are synchronized by FEL *c*.
3) Edges that coterminate in an intra-geon vertex (a fork, arrow, L or tangent Y) tend to correspond to edges of a common geon. These features are grouped by FELs between termination and vertex cells in corresponding locations of L1 and L2 (FEL *d*).
4) An edge occluded by a surface will appear as collinear segments that, excluding accidental alignments, do *not* coterminate with other edges at the points of occlusion. Complimentary, collinear termination cells in distant locations are synchronized by FEL *e* in Figure 2b.

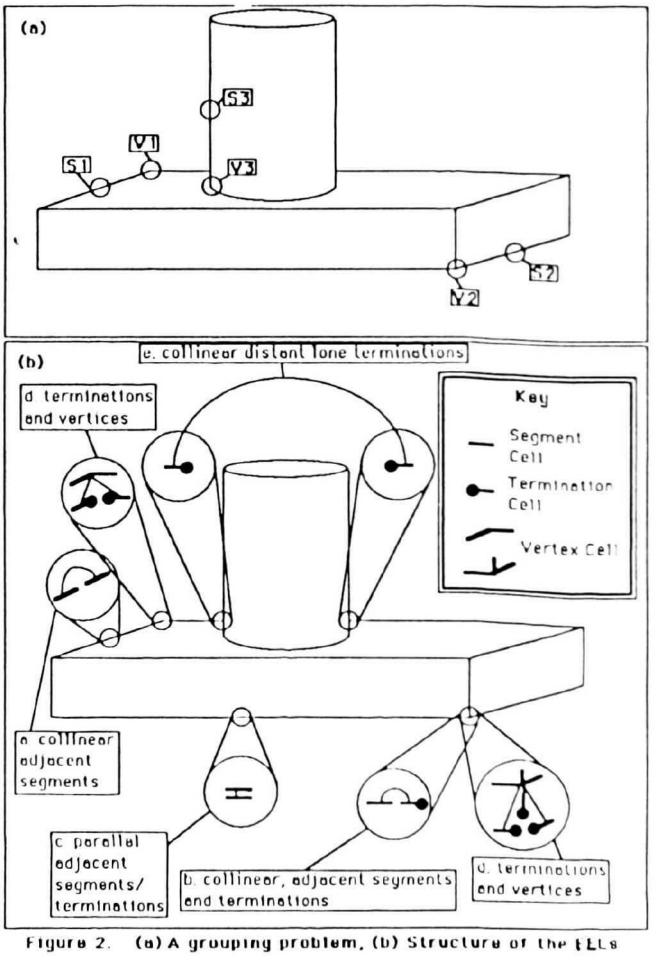

Figure 2. (a) A grouping problem. (b) Structure of the FELs

FELs *a*, *b* and *c* cause all the cells representing a continuous edge to fire synchronously. Termination-to-vertex FELs (d) group vertices with the edges to which they are attached. Since FELs are bidirectional, all the edges coterminating at a given vertex will also be grouped together (an enabling signal will enter a vertex from one termination and be passed, via that vertex, to the other terminations at that location). The distant termination-to-termination FELs (e) operate through vertex cells in L2 which respond specifically to *lone terminations* (edges that do not terminate with other edges, such as the stem of a T vertex). FELs between collinear lone terminations allow the visible pieces of occluded edges to group with one another. Because the L2 lone termination cells are inhibited by inconsistent terminations in L1, this type of "gap jumping" cannot occur when more than one edge terminates at a point. This restriction prevents edges belonging to different geons from being grouped just because they happen to be collinear. Note also that this set of FELs implicitly excludes grouping edges that form T vertices. T vertices are formed at the junction of separate geons, the "top" belonging to one geon, and the "stem" to another, and therefore constitute image features whose constituent parts should *not* be grouped.

Applied iteratively, locally grouping edges and vertices causes all features belonging to a common geon to be grouped, and since blobs and axes receive their inputs from edge cells, the blobs and axes belonging to a given geon will fire on the same time slices as the geon's edges and vertices. In this manner, the local computations performed by the FELs parse an image into its constituent geons. Unlike a top-down or knowledge-driven mechanism, these computations require no information about what volumes are in the image and where they are located. This is an important advantage, because a parsing mechanism that required such information would effectively require that the image already be parsed.

Representing Geons The cells of the model's third layer represent the properties of geons (shown in Figure 1) which have the following characteristics: (1) Geon properties are divided into two classes: *contrastive* properties (such as straight axis vs. curved axis) and *metric* properties (such as location in the image). The former are used directly for recognition while the latter are used to compute the relations among the various geons in an image. (2) Geon properties are activated by the VIPs activated in layer 2. For example, all L2 cells representing Tangent Y vertices activate the cell which responds to curved cross section geons. (3) Each L3 cell responds independently to a particular value on a particular dimension over which geons can vary. For example, the L3 cell that responds to the value *curved* on the dimension *shape-of-major axis* will fire in response to any geon with a curved axis, such as a large curved brick in the upper left of the visual field or a small curved cone in the lower right. Thus, *each geon property is represented in a manner that is invariant with every other geon property.* This invariance, made possible by the binding achieved in L1 and L2, is a crucial aspect of the model's design.

Deriving Relations Among Geons Of the properties derived in L3, only the contrastive properties are used directly for geon classification. The metric properties (size, location, and orientation) are used to determine the relations among the geons in the image, such as relative size, relative location, and relative orientation (Figure 1).

Determination of inter-geon relations is performed in two steps. In L4, relations are computed separately for each value of each dimension. The relation *below* will be used to illustrate how the L4 cells operate, but the logic generalizes to all relations. Consider Figure 3. Associated with every position in Y (Y_p), there is an L3 cell which becomes active when that position is occupied by a geon ($L3_{y=p}$), and there are two L4 cells: one that becomes active when Y_p is below another occupied position ($L4_{\text{below at } y=p}$), and

one that becomes active when Y_p is above another occupied position ($L4_{above\ at\ y=p}$). In L5, there is only one cell for each relation, each of which receives excitation from every corresponding L4 cell. For example, $L5_{below}$ receives excitation from $L4_{below}$ at $y=1$, $L4_{below}$ at $y=2$, etc. Each L4 cell receives two types of input: an excitatory input and an enabling signal (through an FEL). $L4_{below}$ at $y=1$ receives an enabling singal from $L3_{y=1}$ and excitatory input from $L3_{y=2} \ldots L3_{y=5}$. To determine its activation, each L4 cell sums its excitatory inputs over time. Suppose there is a geon at Y_1 ($geon_1$) and another at Y_3 ($geon_3$). $L3_{y=1}$ will fire on the same time slices as the other properties of $geon_1$, and $L3_{y=3}$ will fire in synchrony with the other properties of $geon_3$. $L4_{below}$ at $y=1$ receives an excitatory input from $L3_{y=3}$ because Y_1 is below Y_3. Therefore, when $L3_{y=3}$ fires, $L4_{below}$ at $y=1$ will receive an excitatory input, and its activation will go above zero. When $L3_{y=1}$ fires, it sends an enabling signal to $L4_{below}$ at $y=1$, causing it to fire. Note that $geon_1$ is positioned below $geon_3$. *Below* is therefore a property that describes $geon_1$, and $L4_{below}$ at $y=1$ is now firing in synchrony with $geon_1$'s other properties. Since $L4_{below}$ at $y=1$ sends an excitatory signal to $L5_{below}$, $L5_{below}$ will also fire in synchrony with $geon_1$'s other properties.

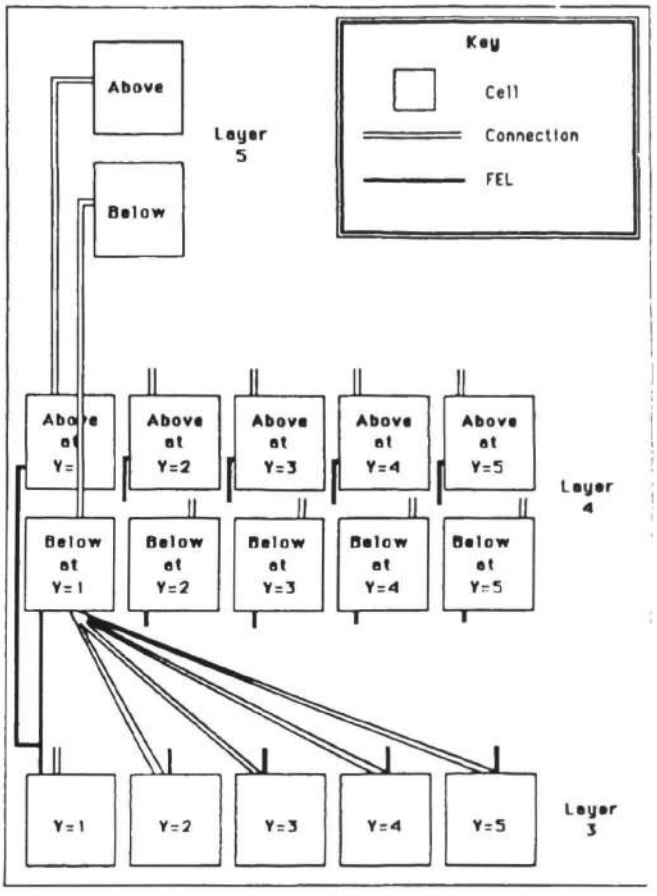

Figure 3. Computing Relations

Recognizing Objects The model's sixth layer receives inputs from L3 cells describing geons, and from L5 cells describing inter-geon relations. L6 and L7 perform two functions: they use the temporal synchrony of inputs within an assembly to ensure that the geons in a object are in the appropriate configuration to define that object, and they combine information from different time slices into an interpretation of a single object.

Preliminary Results and Discussion

Simulations with the model described here have shown that the model is capable of parsing line drawings of simple (even unfamiliar) objects and deriving descriptions of their geons and relations that are completely invariant with scale and translation, and largely invariant with viewpoint. As a consequence, it demonstrates complete translation and scale invariance in recognizing each of the objects with which it is familiar (the objects on which it was allowed to self-organize), and demonstrates rotation invariance resembling that of the human in experimental situations with nonsense objects. That is, it tolerates rotations in depth better than rotations in the visual plane, and its performance on rotations in the plane degrades as a function of the degree of rotation.

References

Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, 94, 2, 115-147.

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

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). New York: Academic Press.

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

von der Malsburg, C. (1981). The correlation theory of brain function. Internal Report 81-2. Department of Neurobiology, Max-Plank-Institute for Biophysical Chemistry

von der Malsburg, C. (1987). Synaptic plasticity as a basis of brain organization. In J. P. Chaneaux & M. Konishi (Eds.), *The Neural and Molecular Bases of Learning* (pp. 4111-432). New York: Wiley