1994-10

Neural Network for Dynamic Binding with Graph Representation: Form, Linking, and Depth-From-Occlusion

NEURAL NETWORK FOR DYNAMIC BINDING WITH GRAPH REPRESENTATION: FORM, LINKING, AND DEPTH-FROM-OCCLUSION

James R. Williamson

October 1994

Technical Report CAS/CNS-94-032

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
111 Cummington Street
Boston, MA 02215

Neural Network for Dynamic Binding with Graph Representation: Form, Linking, and Depth-from-Occlusion

James R. Williamson[1]

Department of Cognitive and Neural Systems
Boston University, 111 Cummington Street, Boston, MA 02215

Abstract

A neural network is presented which explicitly represents form attributes and relations between them, thus solving the binding problem without temporal coding. Rather, the network creates a graph representation by dynamically allocating *nodes* to code local form attributes and establishing *arcs* to link them. With this representation, the network selectively groups and segments in depth objects based on line junction information, producing results consistent with those of several recent visual search experiments. In addition to depth-from-occlusion, the network provides a sufficient framework for local line-labelling processes to recover other 3-D variables, such as edge/surface contiguity, edge slant, and edge convexity.

1 Introduction

Visual object recognition in humans is largely invariant to viewpoint. The 2-D projection of local object features and their relations changes dramatically with change in viewpoint. However, the 3-D structural relationships which compose the object do not. Therefore, viewpoint invariant form representations should explicitly encode the local form attributes and relations between them, so that the stable 3-D *structural description* can be recovered from the unstable 2-D information.

These computational considerations, backed up by many psychophysical results, support structural description models of human visual shape classification, which have independent, explicit representations of form attributes, and relations between these attributes. Alternative approaches, such as template matching and feature list matching, instead "trade off the capacity to represent attribute structures with the capacity to represent relations" (Hummel & Biederman, 1992).

Models of early vision typically use a topographic representation, in which several features that compose a local form attribute are coded at each position in a 2-D lattice. These features are explicitly *bound* by the architecture to their spatial position, but feature conjunctions at the same position, as well as structural relations between positions, such as "next to", "same edge", or "belongs to", are left implicit. To explicitly represent these relations requires *binding*, which is problematic in a neural network architecture due to cross-talk between the highly interconnected representational units (Hummel & Biederman, 1992; Finkel & Sajda, 1992; Barlow, 1981; Feldman & Ballard, 1982). As an example, suppose a red circle is at location x1, a green square is at location x2, and color and shape are represented independently. Then, higher order units that

[1]Supported in part by AFOSR (F49620-92-J-0225), ARPA (N00014-92-J-4015), NSF (IRI-90-24877 and IRI-90-00530), and ONR (N00014-91-J-4100).

detect shape/color conjunctions, invariant to position, are unable to determine which color belongs to which shape (Feldman & Ballard, 1982).

One general binding approach is *spatial coding*, in which each attribute conjunction is explicitly coded by a separate unit (Feldman & Ballard, 1982; Hinton, McClelland, & Rumelhart, 1986). However, using conjunctive codes at each lattice position results in an unacceptable combinatorial explosion of units. The combinatorial explosion can be alleviated with spatial coarse-coding, although this is only effective given a spatially sparse distribution of attributes, since the degree of confusion between different attributes varies with the coarseness of the code, or inversely with the spacing between attributes (Hinton et al., 1986).

Another binding approach is *temporal coding*, in which attribute conjunctions are represented by temporal correlations in the outputs of different neurons (Hummel & Biederman, 1992; Engel, König, Kreiter, Schillen, & Singer, 1992). Support for this idea stems from roughly 50 Hz oscillations found in cortex, with high temporal correlations between neurons responding to spatially distant parts of the same edge of a moving bar, but low correlations between neurons responding to the edges of different bars moving in opposite directions (Engel et al., 1992). This approach suffers from intrinsic capacity limitations, however, due to the limited temporal resolution of neurons. Only a small number of different bindings can be simultaneously encoded, and the length of time required to measure neural temporal correlations may be too great to subserve real-time visual binding (Hummel & Biederman, 1992).

A general consideration in evaluating neural binding mechanisms must be the available neural resources. Neurons have poor temporal resolution, but have large dendritic fields and large fan in/fan out of connections. Therefore, a spatial coding scheme which takes advantage of these neural qualities might provide a more powerful binding mechanism than a temporal coding scheme because it better utilizes the available architectural strengths.

A simple spatial coding scheme (Feldman & Ballard, 1982) uses units with large dendritic fields and simple dendritic processing that make them receptive to local feature conjunctions, invariant to position in the sampling lattice. Here, a combinatorial explosion of conjunctive units is avoided, but at the cost of spatial uncertainty and inability to distinguish the number of copies of the same feature conjunction at different spatial positions. One solution to this problem would be to dynamically *assign* each conjunctive unit to a different, but useful, lattice position.

A neural network model, the GRAF (Graph of Relations And Form) model, takes this approach by dynamically binding nodes coding local form attributes to critical image locations, and then binding the nodes into links with each other, thus producing a graph representation capable of coding 3-D structure. In contrast to the many recent models of dynamic binding that use temporal codes, the GRAF model creates bindings with a combination of dynamic gating of signals to large dendritic fields and competitive interactions.

The GRAF model uses local form attribute information to guide linking across gaps, linking behind objects (amodal linking), and depth segmentation based on occlusion, thereby producing results consistent with those of many recent visual search experiments (Donnelly, Humphreys, & Riddoch, 1991; Enns & Rensink, 1994; Rensink & Enns, 1994). In addition, the GRAF model's representation is sufficient to support local line-labelling processes for recovering 3-D variables

such as surface contiguity, edge slant, and edge convexity (Rensink, 1992; Enns & Rensink, 1991).

1.1 Graph Representations

In coding a visual scene, graph nodes can represent local scene attributes, while graph arcs can represent binding relationships between the local attributes, such as "connected to", "belongs to", "part of", etc. As well as explicitly specifying relations between local attributes, arcs can control communication between the nodes, necessary for relaxation labelling (Hummel & Zucker, 1983).

Similarly, the GRAF model approaches the binding problem by creating a graph representation in which nodes correspond to local form attributes and connecting arcs are explicitly coded, so that relations between nodes are explicit, and communication between a pair of nodes occurs only if their arc is active. Note that I use the term *node* to refer to graph nodes, and the term *neuron* to refer to the basic processing units of the neural network. The graph representation is coded by two sets of neurons corresponding to nodes and arcs. The first set consists of many neural groups, each group making up a node. A node codes local form attributes and 3-D structural information where it is dynamically bound. The second set also consists of neural groups, each group making up an arc, which codes links between nodes. Each arc joins two nodes representing different spatial locations, and controls all communication between them. Unlike the graph representations typically used in machine vision systems, the GRAF model establishes its representation via a parallel processing neural network.

1.2 Overview of GRAF Model

In order for a neural network to create such a graph representation, certain implementational constraints need to be realized. **First**, neurons coding local form attributes, which compose the nodes, should be spatially flexible. Static allocation of a node at each lattice position results in a combinatorial explosion of neurons, since each node requires many neurons to code a full conjunctive set of form features. Rather, nodes are dynamically allocated to spatial positions as a function of featural salience. Neurons composing a node are thus capable of coding local information from many possible spatial positions.

Second, each node represents a unique spatial position, so that the mapping from coded spatial positions to nodes is one-to-one. On the other hand, a many-to-one mapping entails losing spatial and featural identity, while a one-to-many mapping results in an inefficient allocation of nodes.

Third, links between pairs of nodes are explicitly coded. The strength of a link is based on the spatial positions and attributes coded by its two nodes. Due to the spatial flexibility of nodes, relative spatial position is recovered only after the nodes are dynamically spatially allocated. If the spatial flexibility of nodes is unlimited, then linkage must be possible between arbitrary pairs of nodes. Presuming that links have spatial limits, then, as the spatial flexibility of nodes is restricted, certain links can be ruled out because some pairs of nodes can never code sufficiently nearby spatial positions to be linked.

Limiting the number of allowed links by restricting the spatial flexibility of nodes helps to avoid a combinatorial explosion, since the GRAF model statically allocates an arc for each possible pairing of nodes, which in the fully connected case is $\frac{N(N-1)}{2}$ arcs given N nodes. After nodes are dynamically allocated to different spatial positions, and code the local feature conjunctions, a competitive selection process establishes links (active arcs), thus specifying binding relationships between pairs of nodes.

The GRAF model is broadly illustrated in Figure 1. In Figure 1a, four "potential" nodes (circles), and six arcs (dotted lines), wait to code an input. Given a visual input of the triangle in Figure 1b, a saliency map of the important lattice positions (Figure 1c) is activated, and three of the nodes are dynamically allocated to the positions of local maxima in the saliency map, coding the local form attributes (Figure 1d). Based on spatial relations and form attributes of the nodes, the appropriate arcs become activated, as shown by bold dotted lines in Figure 1e. The information coded by the nodes and arcs is schematically shown in Figure 1f, demonstrating a compressed representation of the object with explicit bindings between line junctions.

1.3 Psychophysical Data

What is the psychophysical evidence that sophisticated form representations are obtained in a purely bottom-up "preattentive" manner, as suggested above? Until recently, the prevailing view was that binding of local features between dimensions and across space, necessary to determine 3-D structure, requires attention (Treisman, 1985).

Recent experiments using the visual search paradigm have shown, however, that preattentive representations of 3-D structure are obtained which require integrating complex information from localized regions, such as line junctions, across objects (Enns & Rensink, 1991). In addition, preattentive structures are obtained by grouping disconnected figures, where the grouping is again dependent on complex information at localized regions (Donnelly et al., 1991; Enns & Rensink, 1994; Rensink & Enns, 1994). Therefore, rapid, preattentive vision is capable of tasks once thought to require attentive processes. To explain the experimental results, these rapid visual processes must be, for the most part, parallel and automatic.

1.4 Physiological Data

What is the physiological support for the GRAF model? The model predicts that extrastriate cells should be found with classical receptive fields much larger than the size of their optimal stimulus. These cells should, at any one time, respond to only a portion of their receptive field, ignoring the rest.

Many extrastriate cells, in areas V2 and V4, have been found with classical receptive fields much larger than the size of their optimal stimuli (Desimone & Schein, 1987; Hubel & Livingstone, 1985). Many V4 receptive fields can apparently be restricted to a subregion that corresponds to an attended location (Moran & Desimone, 1985). Models to explain these data posit attentional gating of receptive fields (Van Essen & Anderson, 1990; Desimone, 1992), or a feature-based sup-

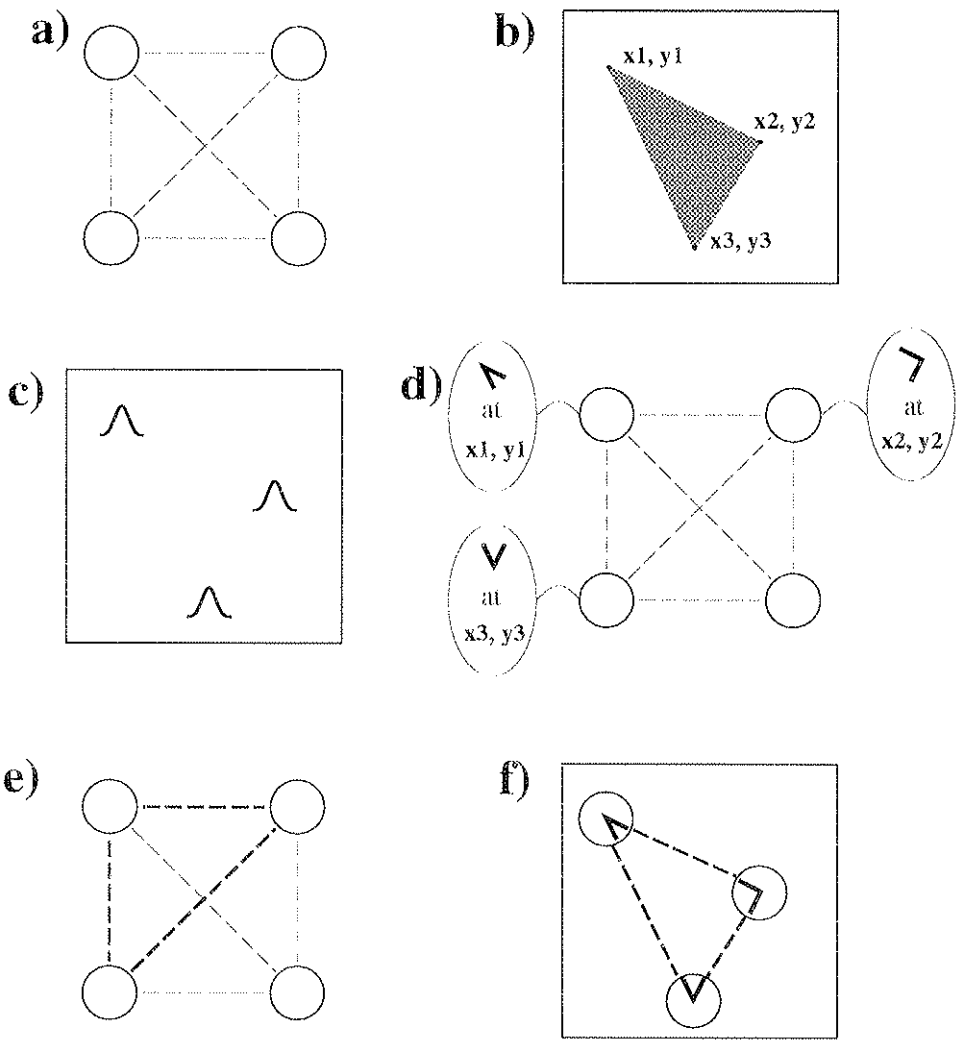


Figure 1: Illustration of GRAF model: **a)** four unallocated nodes (circles), and six arcs (dotted lines); **b)** input image of triangle; **c)** resulting activity of saliency map, with local maxima at line junctions; **d)** three nodes are allocated to local maxima of saliency map, coding position and local form attributes; **e)** based on codes of graph nodes, arcs between appropriate pairs are activated; **f)** information coded by network is depicted by placing nodes at their coded positions, iconically representing form attributes coded by each node, and showing active arcs between nodes.

pression mechanism (Desimone, 1992). The GRAF model predicts that stimulus-driven receptive field restriction also occurs independent of focused attention.

In the remainder of the paper, the GRAF model is described in detail and simulations of the model are shown which account for many of the psychophysical results discussed above. the GRAF model consists of four stages: 1) *Feature Extraction*, in which simple features are topographically represented; 2) *Feature Abstraction*, in which the topographic feature representation is spatially abstracted into nodes, which code form attributes and their spatial positions; 3) *Attribute Linking*, in which nodes establish pairwise links (activated arcs) based on their form attributes and spatial relations; 4) *Depth Segmentation*, in which relative depth is initially estimated at each node, based on local form attributes, and estimates are subsequently refined by relaxation between linked nodes.

2 Feature Extraction

The first stage of the GRAF model is feature extraction within retinotopic coordinates. The output of this stage consists of local form measurements from oriented complex and end-stopped cells at each 2-D lattice position. Oriented complex cells represent smooth object boundaries, while end-stopped cells represent boundary discontinuities or segments of high curvature. Many types of boundaries, such as texture boundaries and illusory contours, are currently ignored for the sake of simplicity.

The complex and end-stopped cell responses are obtained using oriented filters and subsequent nonlinearities in a process adapted from Heitger, Rosenthaler, von der Heydt, Peterhans, & Kubler (1992). The responses are combined across orientation, separately for complex and end-stopped cells, to produce two saliency maps. The first is a *Continuation Saliency Map* (from complex cells), and the second a *Junction Saliency Map* (from end-stopped cells). The respective saliency maps provide the bases for allocation of *Continuation* and *Junction* nodes to appropriate spatial positions in the subsequent Feature Abstraction stage. The Continuation and Junction nodes respectively represent smooth boundaries and boundary discontinuities.

The Feature Extraction stage is schematically illustrated in Figure 2, which shows the Continuation and Junction Saliency Maps resulting from an example input image. The details of the Feature Extraction stage are described in Appendix A.

3 Feature Abstraction

In the Feature Abstraction stage, nodes are allocated to lattice positions, where they code local form attributes, as illustrated in Figure 1d. This process occurs in parallel for two sets of nodes, Continuation and Junction nodes, which code smooth boundaries and boundary discontinuities, respectively. Continuation nodes are allocated to activated locations of the Continuation Saliency Map, while being prevented from coding the same locations as Junction nodes, which are allocated to activated locations of the Junction Saliency Map.

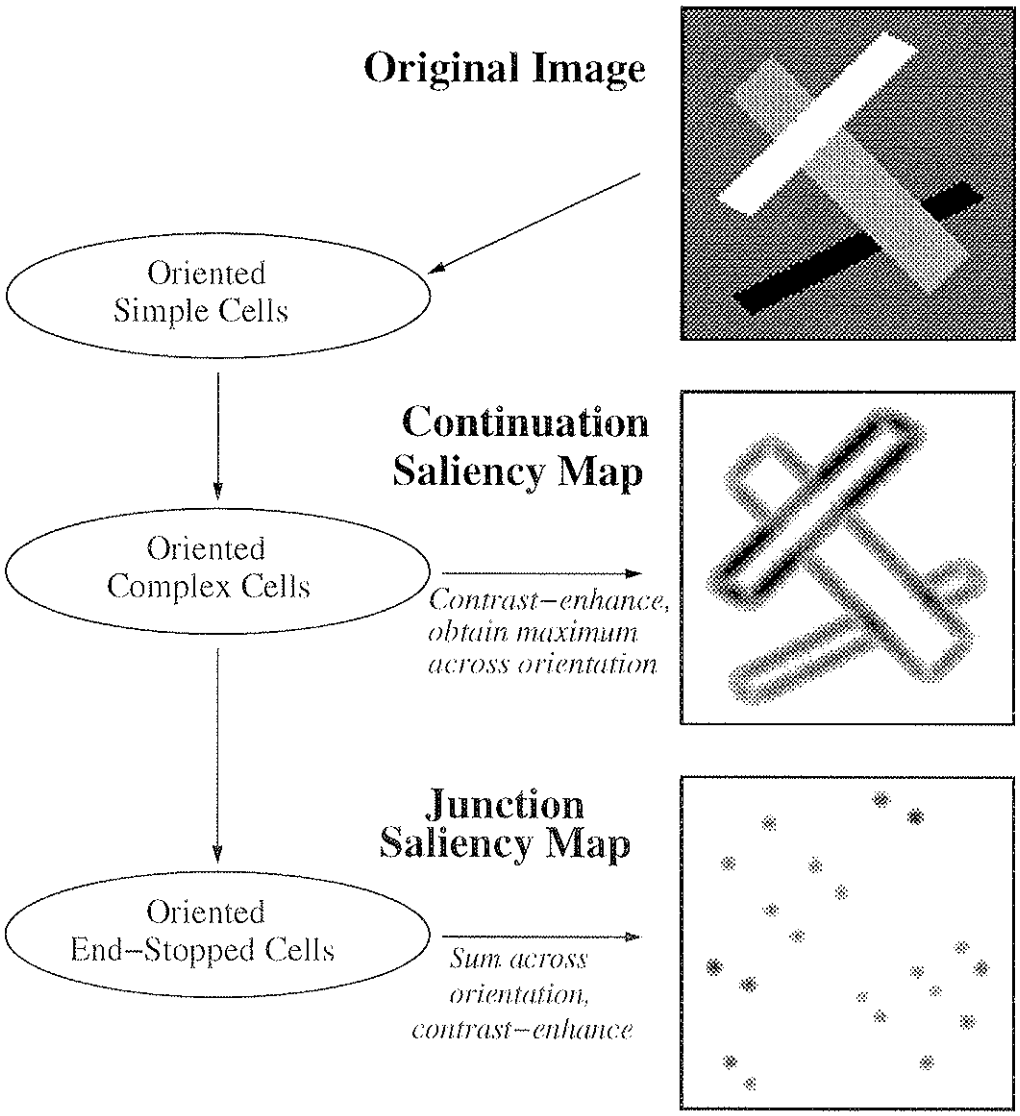


Figure 2: Feature Extraction example, given a 100x100 pixel input image of three overlapping bars. Activity in Continuation and Junction Saliency Maps is shown.

3.1 Composition of Nodes

To understand the Feature Abstraction process, it is necessary to understand the neural composition of nodes. Each node consists of three classes of neurons, *Position*, *Form Attribute*, and *3-D Structure* neurons.

3.1.1 *Position* Neurons

Each node has its own Position Map, made up of topographically organized Position neurons in one-to-one correspondence with a spatially offset chunk of a Saliency Map. With the activation of a single Position neuron, a Position Map codes the location where its node is dynamically allocated. Position Maps of neighboring nodes overlap, thereby providing flexibility in how the nodes are spatially dynamically allocated. Figure 3a shows how three neighboring Position Maps overlap with respect to a Saliency Map.

3.1.2 *Form Attribute* Neurons

Form Attribute neurons represent conjunctions of complex and end-stopped cells from the Feature Extraction stage, at the spatial location specified by the Postion Map. By coding conjunctions of low-level retinotopic features, the precise form attribute is explicitly represented by the best-matching Form Attribute neuron, which inhibits other Form Attribute neurons of the same node.

3.1.3 *3-D Structure* Neurons

3-D Structure neurons can potentially code many variables such as relative depth, convexity, slant, surface contiguity, etc, although the current GRAF model only codes relative depth. The other variables can be effectively estimated using local operations, as shown in (Rensink, 1992), and so could be added to the model. Initial local estimates of these variables can be made based on each node's form attribute. This estimate can then be updated, based on information passed across the arcs between different nodes, through relaxation processes. The GRAF model currently demonstrates this process with the relative depth variable.

3.2 Allocation of Nodes

Each neuron in a (Continuation or Junction) Saliency Map, which is in absolute, retinotopic coordinates, sends output to a single Position neuron in a node's Position Map. The Position Maps of nearby nodes spatially overlap, as shown in Figure 3a. A single node is allocated to an active Saliency Map location via two simultaneous competitive processes. The *first* is winner-take-all competition between different nodes for the same general location of the Saliency Map. This is accomplished by feedback suppression from Position neurons to the nearby output signals from the Saliency Map that feed to the other nodes, as well as to nearby Position neurons of the same node. The *second* is winner-take-all competition for different spatial locations between neurons of

each Position Map. These two inhibitory processes are illustrated in Figure 3a. The allocation process establishes one-to-one mappings between salient locations and nodes. The active Position neuron determines the image location where the "templates" of the Form Attribute neurons of the same node are centered, and thus the location coded by the node. This process is described in greater detail in Appendix B.1.

3.3 Coding of Local Form Attributes

Each node has many Form Attribute neurons, each of which is receptive to different conjunctions of complex and end-stopped cells, and so represents different precise form attributes. Each Form Attribute neuron applies a template to the input pattern made up of signals from complex and end-stopped cells. The Form Attribute neuron with best matching template to the input pattern inhibits all other Form Attribute neurons.

Each Form Attribute neuron can "apply" its template anywhere within a large spatial extent, because it has a large dendritic field which receives several spatially separated input copies. A Form Attribute neuron can only respond where the node is dynamically allocated, however, because its dendritic field is gated by the active Position neuron, as illustrated in Figure 3b. Activation of Form Attribute neurons is described in more detail in Appendix B.2.

An example of the Feature Abstraction result, given the input image and Feature Extraction shown in Figure 2, is shown in Figure 4. Here, nodes are illustrated in their spatially allocated positions. Junction nodes are depicted with circles, with icons showing their coded form attributes. Continuation nodes only code boundary orientation, so they are depicted with oriented bars.

4 Linking Form Attributes

Following the dynamic allocation of nodes and subsequent coding of form attributes, pairs of nodes are linked together, based on their spatial relations and form attribute codes. The linking of nodes is coded by arcs, which are composed of neurons coding the distance and angle between nodes, as well as neurons coding the strength and direction of inter-node linking.

4.1 Initial Estimate of Linking Strength

Each pair of nodes has a statically allocated arc. As soon as the two nodes are spatially allocated, the spatial relation between the nodes is recovered. Together, the nodes' active Position neurons activate, via 2nd order connections, Distance and Angle neurons of the inter-node arc, so that the angle θ and distance d between the nodes' loci is represented, as shown in Figure 5a.

Once the arc codes the spatial relation between the two nodes, its strength is initially estimated based on the degree of colinearity and cocircularity between the outwardly continuing boundaries of the form attributes coded at the nodes. This is determined by excitatory input to Linking neurons, which code links in all possible directions between nodes. First, the Distance and Angle

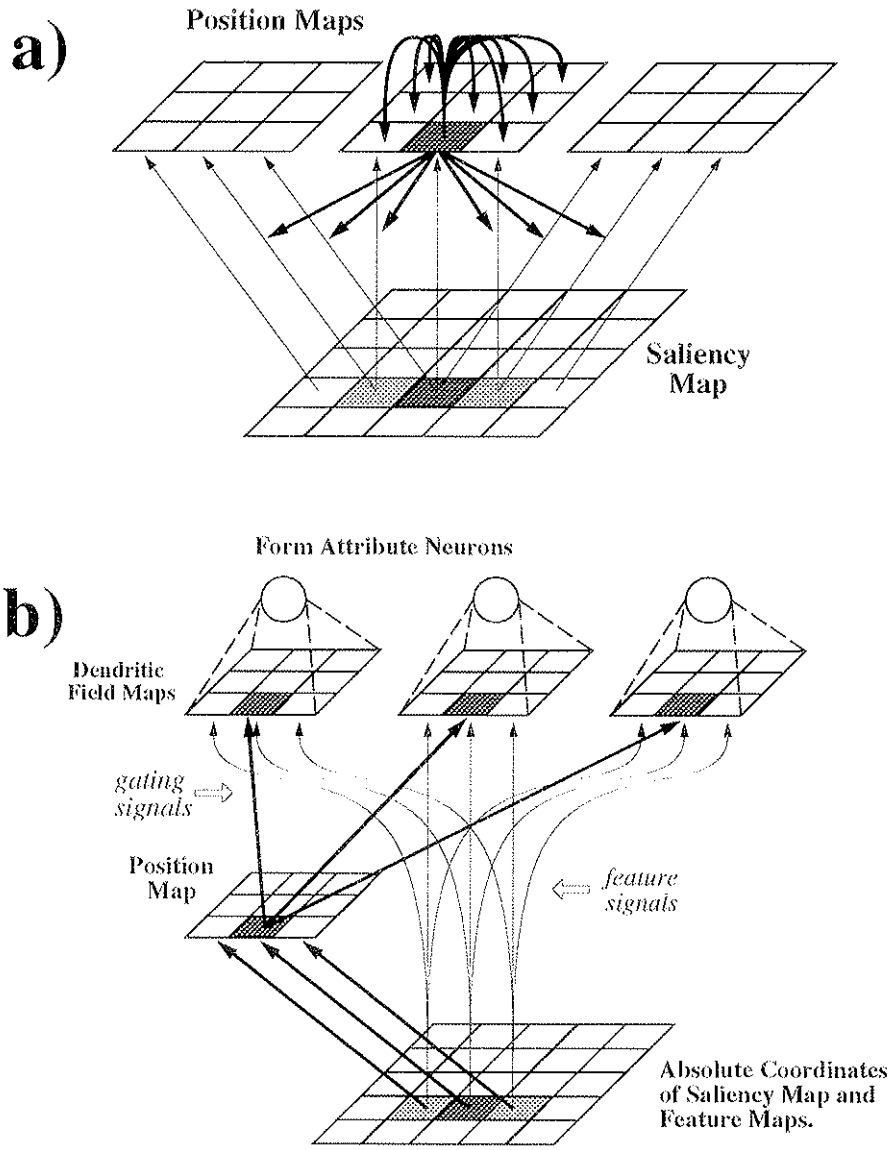


Figure 3: a) Competitive processes that insure one-to-one mappings from local maxima in a Saliency Map to a single neuron in a Position Map are shown. Inhibitory signals are depicted with bold lines. The active Position Map neuron inhibits all other neurons in its map, and also suppresses output signals from nearby positions in the Saliency Map to its map and to nearby, overlapping, Position Maps. b) Positional gating of Form Attribute neurons by an active Position Map neuron is shown. Bold lines indicate positional "where" signals, while thin lines indicate featural "what" signals.

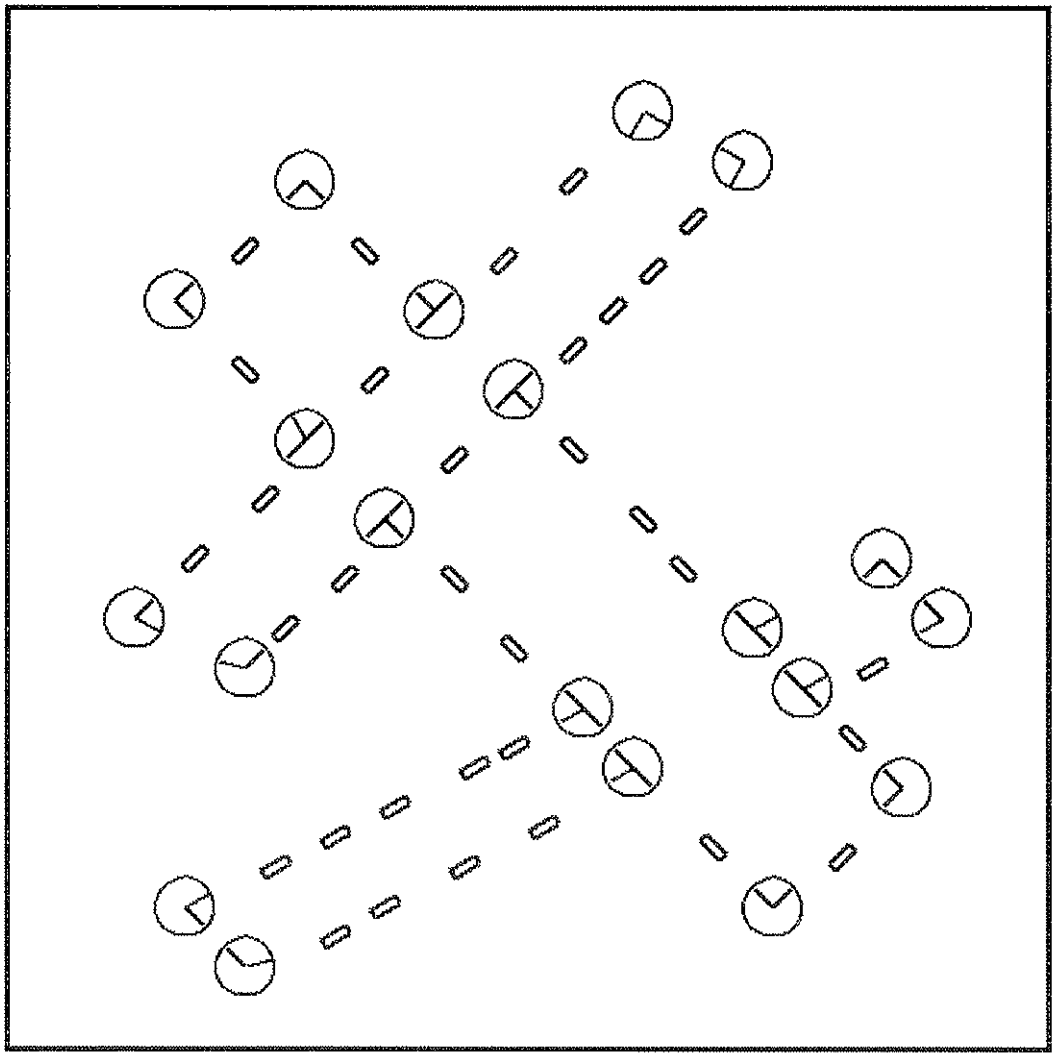


Figure 4: A Feature Abstraction example, given the Feature Extraction result shown in Figure 4 as input, is shown. Junction nodes are depicted by circles with icons showing the coded Form Attribute. Continuation nodes are depicted by oriented bars. Nodes are shown in their bound positions.

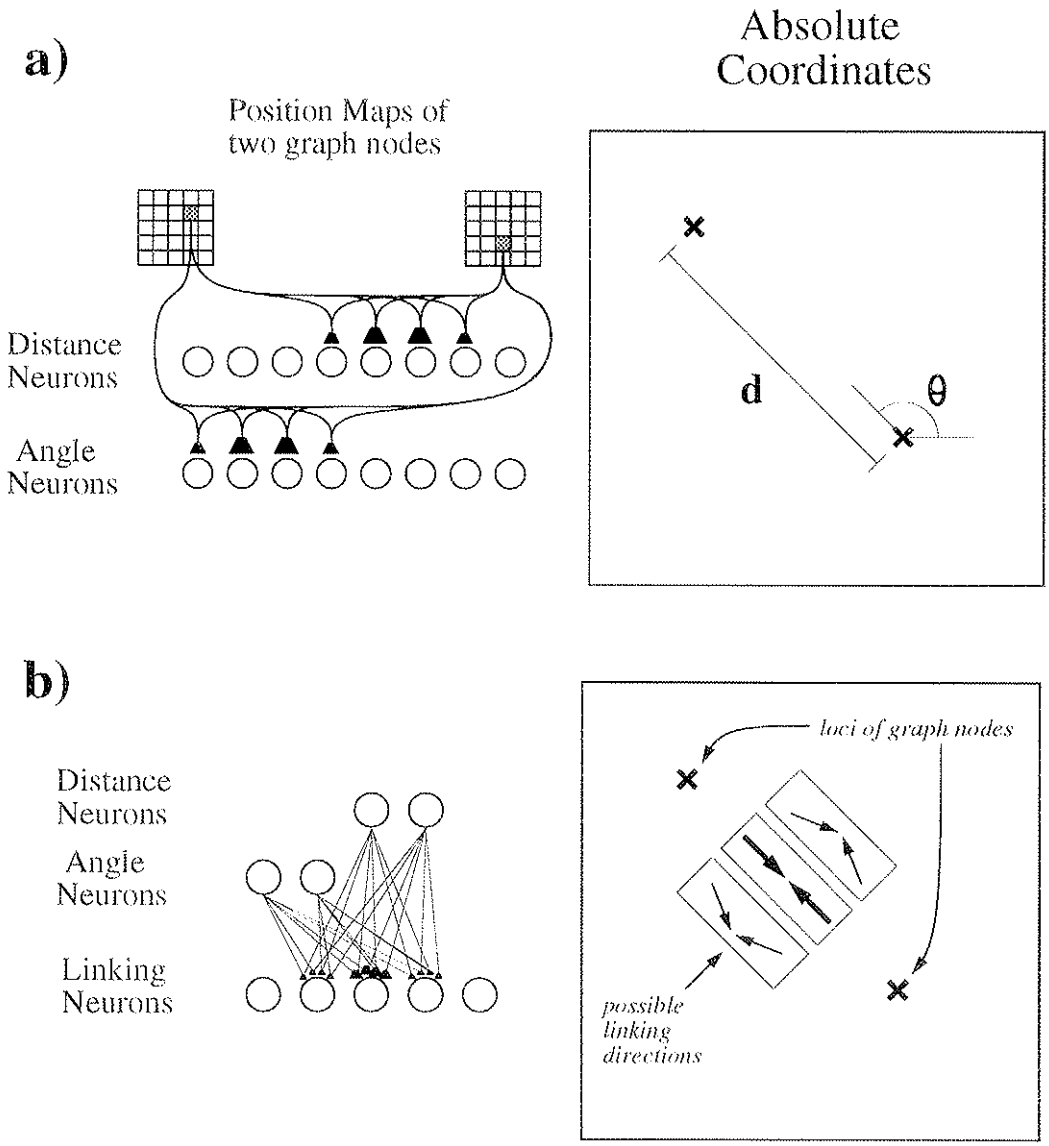


Figure 5: a) The spatial location coded by each node's active Position neuron is combined, via 2nd order connections, with that of the other node to activate Distance and Angle neurons of the connecting arc which specify the *current* spatial relation between the nodes. The spatial relation consists of the distance d and angle θ between the nodes, as shown to the right. b) Linking neurons are gated by active Distance and Angle neurons, forming a dynamic template of possible linking directions, given the spatial relation of the nodes' loci, as shown to the right.

neurons excitatorily gate Linking nodes for appropriate possible linking directions, forming a dynamic template, as shown in Figure 5b. Next, each node sends excitatory grouping signals, based on its coded form attributes, to the Linking neurons. The degree to which these signals match with the excitatory gating of the Linking neurons determines the strength of the net input signal, as illustrated in Figure 6a. Note that the nodes in Figure 6a can potentially link vertically and horizontally based on their form attributes, yet the dynamic template in the Linking neurons (indicated by shading) allows only horizontal linking.

Linking is based on colinearity and cocircularity of the boundary inducers. The linking constraints obey spatial grouping rules similar to those of other models, following the principle of *good continuation* (Grossberg & Mingolla, 1985; Parent & Zucker, 1989; Kellman & Shipley, 1991). Two types of linking exist, corresponding to 1) modal grouping and 2) amodal grouping. Modal grouping is illustrated in Figure 6b, which shows spatially "fuzzy" colinear grouping signals of Continuation and Junction nodes. Amodal grouping only takes place between Junction nodes which code T-junctions and Terminations, form attributes that indicate occlusion. Amodal grouping must often be spatially long-range, so an additional long-range cocircular component is used, as shown in Figure 6b. The process outlined above of initially estimating the linking strength is described in detail in Appendix C.1.

4.2 Competition Between Links

The initial link estimates of different arcs must compete with each other so that, locally, the most likely links are selected. To appreciate the importance of this, consider two nearby parallel boundaries. Links should only exist along each of these boundaries, yet some initial activation of links *between* the different boundaries is inevitable. The existence of strong links along each boundary should thus suppress all links between the boundaries. In general, links between different boundaries should only code properties such as parallelism or common surface ownership, yet these types of links are beyond the scope of the current GRAF model. Currently, links only represent the continuation of a single boundary. How do two different arcs "know" if they should compete with each other? First, competition is only possible if the two arcs share a common node. Second, arcs compete only if they code links in similar directions. This constraint is enforced by direct competition between the Linking neurons of different arcs. This competition is illustrated in Figure 7a, where suppression of "cross-boundary" links is shown. A simulated example of linking, based on the Feature Abstraction example shown in Figure 4, is shown in Figure 7b, before and after inter-arc competition. The process of competition between links is described in detail in Appendix C.2.

5 Depth Segmentation

Now that the nodes are linked together, a framework exists for information propagation to refine local 3-D estimates, based on more global information. Much work has been done on the use of local operations for line-labelling and occlusion-based depth segmentation (Rensink, 1992; Finkel & Sajda, 1992; Grossberg, 1994). A key to this process is proper control of communication between

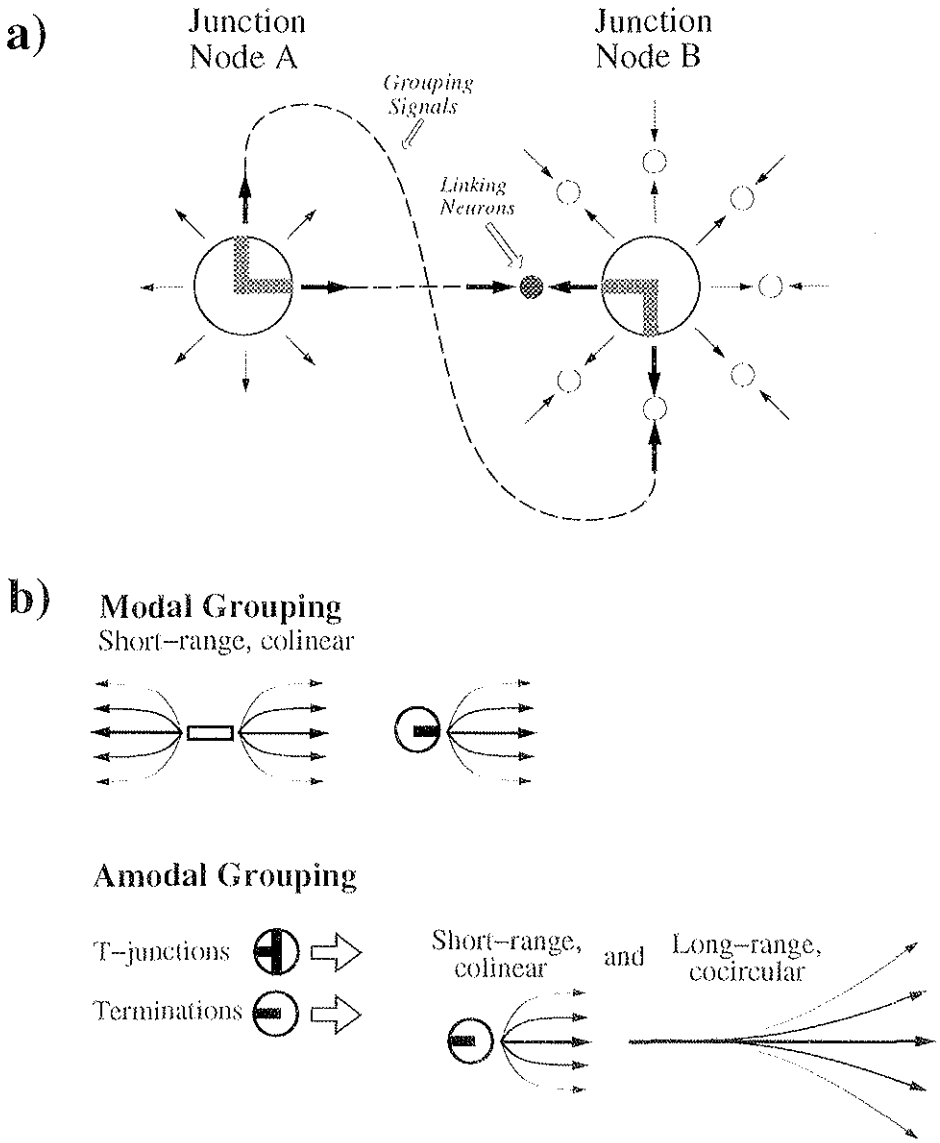


Figure 6: **a)** Nodes send grouping signals based on their coded form attributes. If these signals match with the dynamic template established at the Linking neuron(s) (indicated by shading), then the Linking neurons are activated. Thus, the nodes depicted above link horizontally but not vertically. **b)** Modal grouping signals of Continuation and Junction nodes are colinear with spatial fuzziness. Amodal grouping signals only come from Junction nodes coding T-junctions or Terminations. These signals have an added long-range cocircular component.

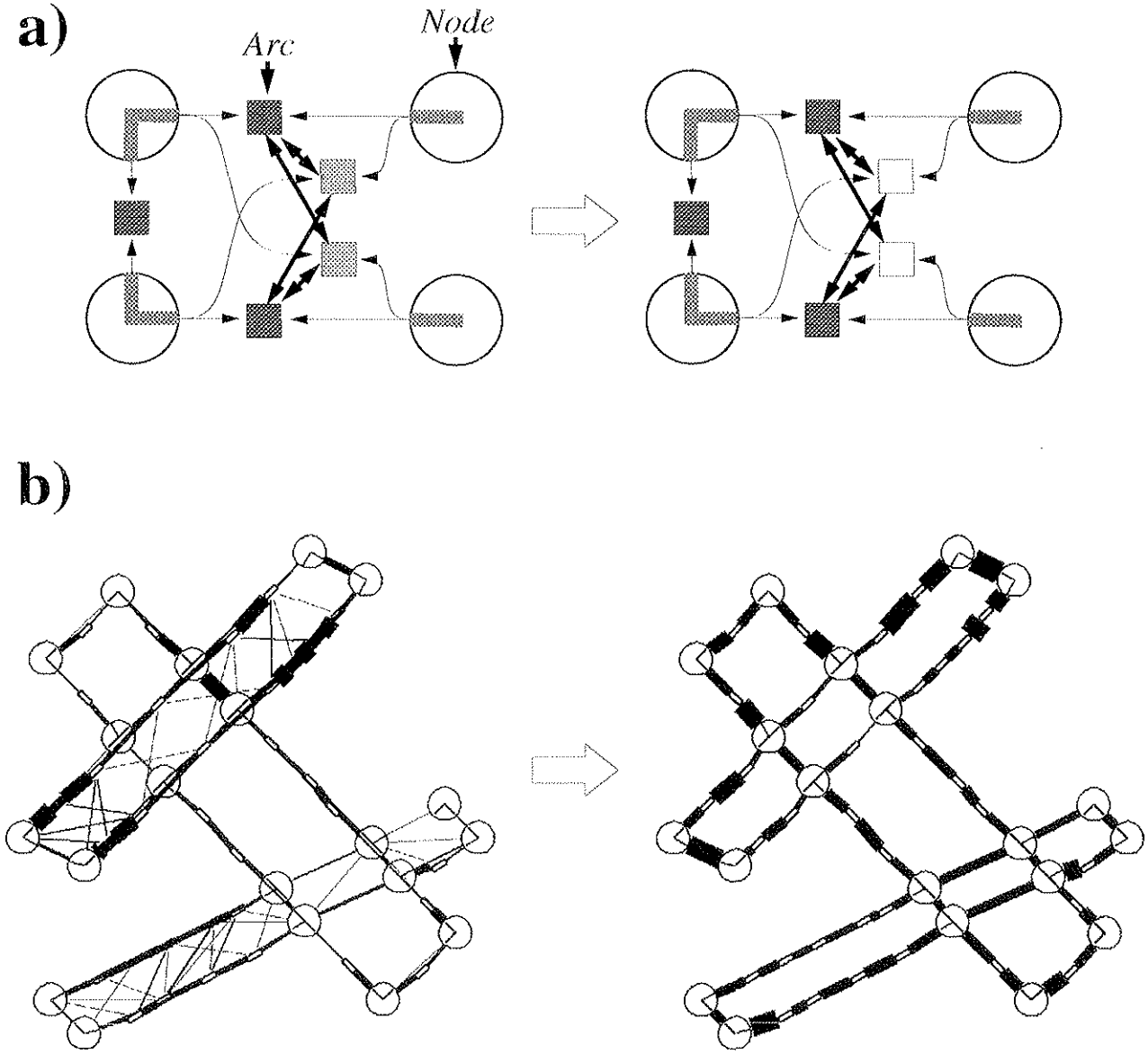


Figure 7: **a)** Demonstration of inter-arc competition is shown. Arcs are depicted with squares, with activity level depicted by shading. Initial activation of arcs is based on local grouping information (left). The most likely links are selected by competition between arcs sharing a node *and* coding similar direction (right). **b)** Example of linking of nodes, in which links are depicted with lines, and line width corresponds to the strength of the link: (left) before competition between arcs; (right) after competition between arcs.

local representations. The GRAF model achieves this control with its explicitly coded arcs, which bind the nodes in appropriate relations with each other.

Currently, the only 3-D structure processing performed by the GRAF model is relative depth segmentation based on occlusion relationships. This processing is illustrative, however, of how other 3-D variables could be recovered, such as surface contiguity, edge slant, and edge convexity (Rensink, 1992). The GRAF model's depth segmentation mechanism is similar to that of Finkel & Sajda (1992), who achieve binding with tags. Their tags do not have a neural implementation, although the authors suggest a temporal code implementation. Depth segmentation in their model is triggered at "tag junctions", which occur at T-junctions. Similarly, the GRAF model also segments boundaries in depth at T-junctions, which are explicitly coded. In addition, the GRAF model can be easily expanded to include line-labelling processes that require explicit coding of L-, Y-, and Arrow-junctions.

Any additional 3-D structure processes in the GRAF model would be controlled by the same gating mechanism that controls depth segmentation. Communication between a pair of nodes is gated by the activation of an arc, as depicted in Figure 8a. For depth segmentation, communication between nodes enforces the same depth, while at a Junction node coding a T-junction, communication between two depth representations enforces higher depth for the top bar, and lower depth for the bottom stem.

The details of these depth signals are illustrated in Figure 8b. At each node, several *3-D Structure* Depth neurons coarse code depth. The depth signals between linked nodes are on-center/off-surround, thus encouraging linked nodes to code the same depth. In a Junction node coding a T-junction, two sets of Depth neurons represent the stem and top bar of the T, respectively. The competitive interactions between these neurons "push" the top bar to a higher depth, and "push" the stem to a lower depth.

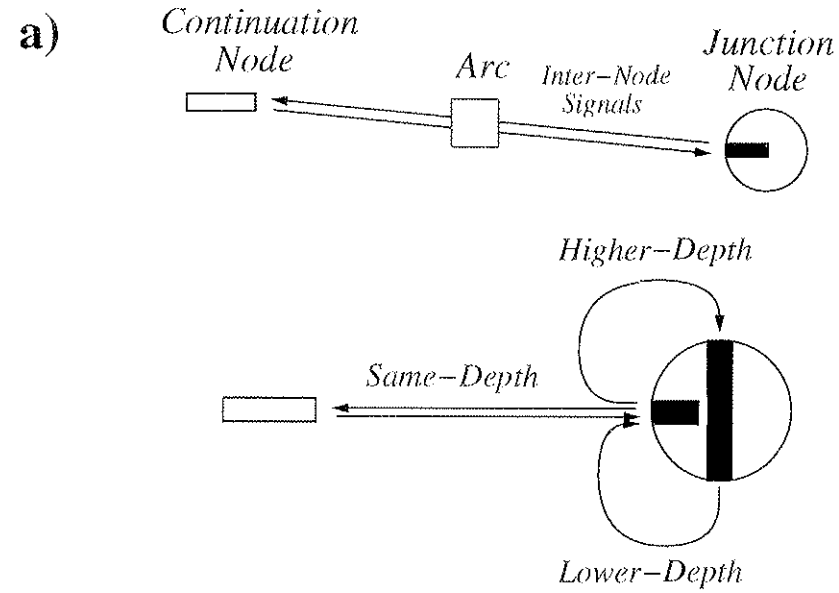
A simulated example of depth segmentation, given the final linked representation depicted in Figure 7b, is shown in Figure 9, which shows Junction nodes (points) and locally maximum arcs (lines) in a 3-D plot, in which the ordinate represents depth as coarse coded by the depth nodes. Over time, the occluding and occluded bars are pushed away from each other, due to the cooperative/competitive relaxation between Depth neurons of linked nodes. The process of depth segmentation is described in detail in Appendix D.

6 Simulations

Simulations use equations and parameters specified in the Appendices. The same set of parameters are used in all the simulations, as well as in the examples shown earlier.

6.1 Grouping across gaps

Rensink and Enns (1994) used visual search tasks to show that an apparent length illusion induced by Muller-Lyer stimuli is obtained preattentively. Thus, if a target Muller-Lyer figure is "wings-



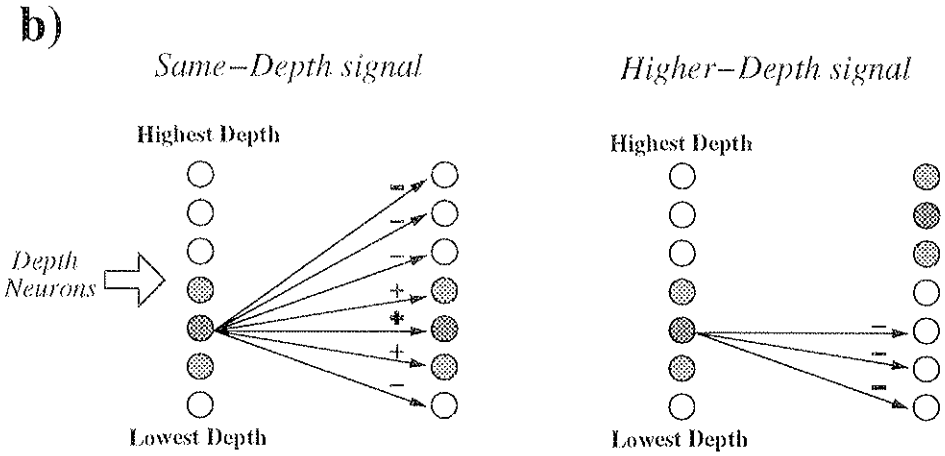


Figure 8: **a)** Communication between nodes is gated by arc activation. Communication for relative depth segmentation consists of "same-depth" between linked nodes, and "higher-" or "lower-depth" between the stem and top bar of a Junction node coding a T-junction form attribute. **b)** The *same-depth* signal between linked nodes, of coarse coding Depth neurons, is shown (left), and the *higher-depth* signal, within a Junction node, from Depth neurons representing the T-junction stem, to Depth neurons representing the T-junction top-bar, is shown (right).

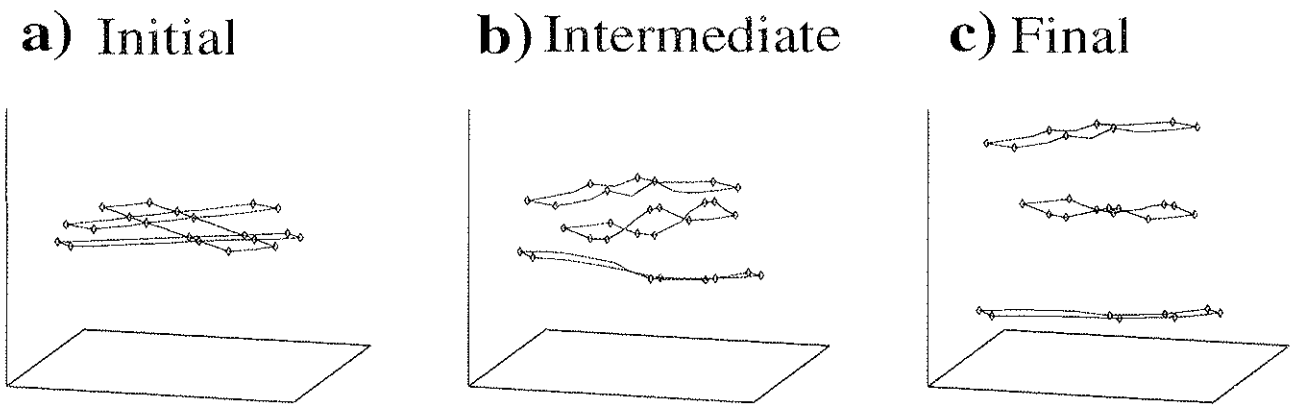


Figure 9: Depth segmentation of Junction nodes (points) and connecting arcs (lines), given the linked representation shown in Figure 7b, is shown. The ordinate represents depth, as coarse coded by Depth neurons.

out" and distractor figure is "wings-in", then search is slow only if the lengths of the entire figures, including the wings, are the same. They used this basic result to explore conditions under which preattentive grouping across gaps binds contour fragments together.

Figure 10a shows example target/distractor pairs, adapted from Rensink & Enns (1994), in which the entire distractor wings-in figure is shorter than the target wings-out figure, but in which each figure contains gap(s). If the gap is in the middle of the connecting bar, as shown in the top two cases, then search is equally fast, regardless of whether the gap is small or large, indicating that the pieces are bound together across the gaps. The GRAF model produces consistent results, shown to the right, in which the figures are linked across the gaps. On the other hand, if two gaps are placed so that the Y-junctions become L-junctions, as in the bottom two cases, then search is slow regardless of whether the gaps are large or small. This indicates that the pieces are not bound together. The GRAF model produces consistent results, in which the pieces are not linked across the gaps.

Figure 10b shows variations involving a center gap. Here, the overall target and distractor figures are of the same length, so fast search indicates a lack of binding of the pieces, and slow search indicates binding of the pieces; see Rensink and Enns (1994) for details. If the right-hand pieces are shifted vertically, so that cocircular interpolation of the segments across the gap is removed, as shown for the top two cases, then binding only takes place if the gap is small (as indicated by the search results). If both segments are bent so that cocircular interpolation still joins them, as shown in the bottom two cases, the pieces are bound whether the gap is small or large. The GRAF linking results are consistent with these experimental results, as shown in Figure 10b. The ability of the GRAF model to produce these results derives from the use of two separate constraints for amodal grouping: 1) short-range, fuzzy colinear grouping; and 2) long-range, cocircular grouping.

Donnelly et al. (1991) showed that a target is much easier to find if the distractors group easily. Figure 11a shows target absent (top) and target present (bottom) cases, in which the target is a "flipped" chevron, and the distractor chevrons line up with each other. Search for target absent and target present are both very fast, indicating that the chevrons are preattentively grouped, or bound, together. The GRAF model produces consistent results, shown alongside the stimuli.

Figure 11b shows configurations in which all the chevrons are flipped with respect to those of Figure 11a. Here, search for target absent and target present are both very slow, indicating that the distractors are not grouped together. Again, the GRAF model produces consistent results, shown alongside.

6.2 Amodal Completion and Depth Segmentation

The example in Figures 7b and 9 shows the GRAF model's "recovery" from occlusion, which consists of amodal completion and depth segmentation. Enns and Rensink (1994) showed that introducing small gaps between the occluding and occluded objects causes dramatic changes in visual search, presumably because the two segments of the occluded object are no longer bound together. Figure 12a and 12b illustrates this finding, where occlusion in Figure 12(a) results in amodal linking of the occluded object and subsequent depth segmentation of the objects. In Figure 12(b), on the other hand, small gaps separating the objects result in no amodal linking and no depth segmentation.

Although the GRAF model uses no surface representation, it still demonstrates some rudimentary intelligence in its amodal linking, working entirely at the level of boundaries. Figure 12c shows an input image similar to that of Figure 12a, except that the right-hand segment is shifted vertically by half its height. The result is that before competition in the Linking stage, three different amodal links are equally strong (middle). Due to inter-arc competition, however, the two correct links are chosen (right).

The depth segmentation examples so far have been globally consistent. Given a scene which is globally inconsistent assuming planar objects with no 3-D slant, the GRAF model finds the best compromise between local cues for different-depth at T-junctions, and same-depth along boundaries, producing figures bent in depth in a spline-like way (Figure 12d).

7 Discussion

Hummel & Biederman's (1992) dynamic binding model parses images into geons in explicit relations with each other, using low-level features including line junctions. Like the GRAF model, their model represents each possible one-, two-, and three-pronged junction. Unlike the GRAF model, however, their model has a full set of these representations at each 2-D lattice position, resulting in a combinatorial explosion of units. By dynamically binding nodes to positions, on the other hand, the GRAF model reduces the combinatorial requirement by about two orders of magnitude, with the actual reduction depending on parameters determining the number of Junction nodes relative to the size of the 2-D lattice. As well as avoiding a combinatorial explosion of units,

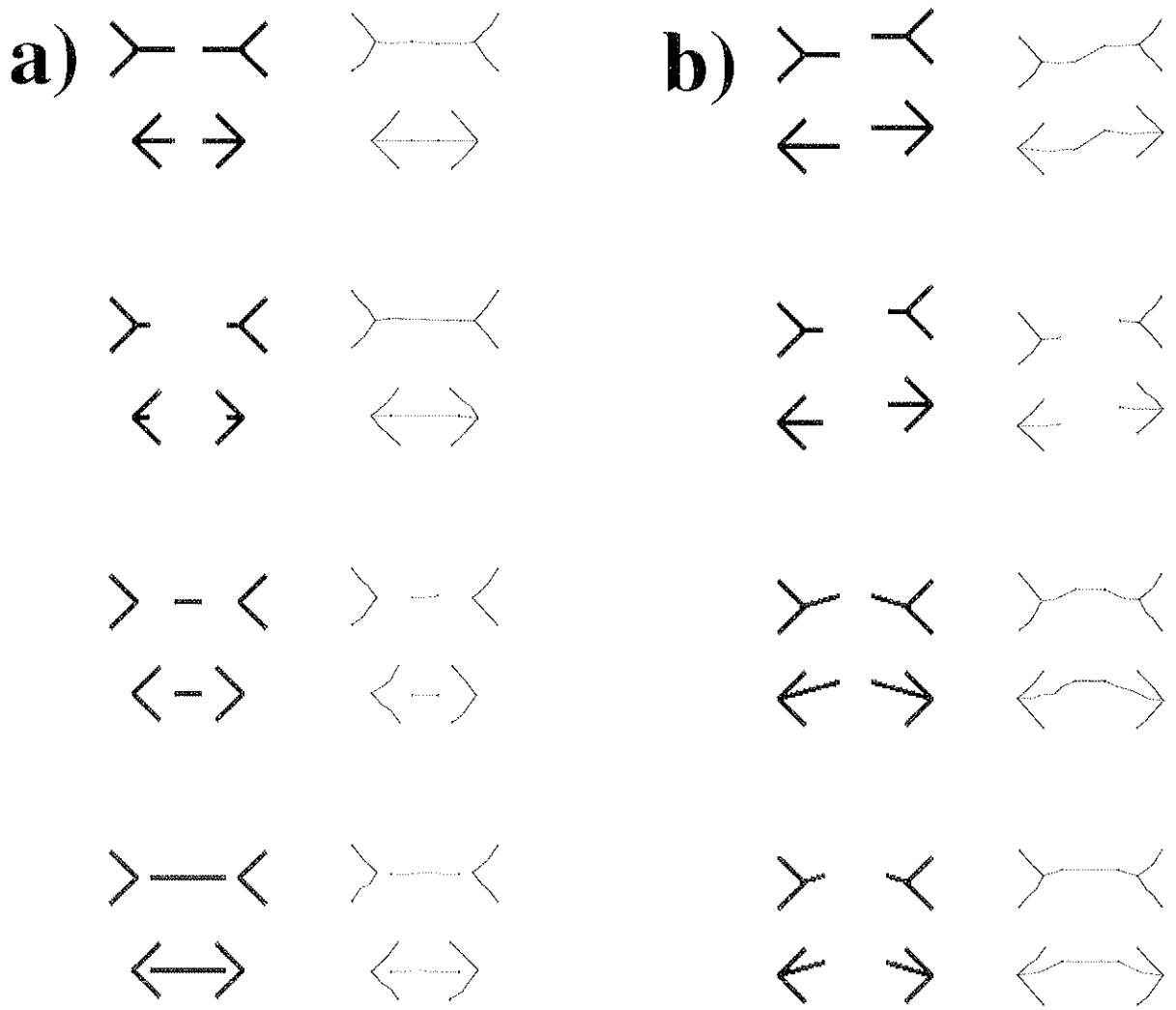


Figure 10: Simulation results for Muller-Lyer stimuli adapted from (Rensink & Enns, 1994), is shown. Input stimuli are shown on the left of each column, with linked GRAF representation shown on the right, with Junction nodes (points), and locally maximum arcs (lines) plotted. Linking across gaps obtained by the GRAF model is consistent in all of these cases with the experimental results of (Rensink & Enns, 1994).

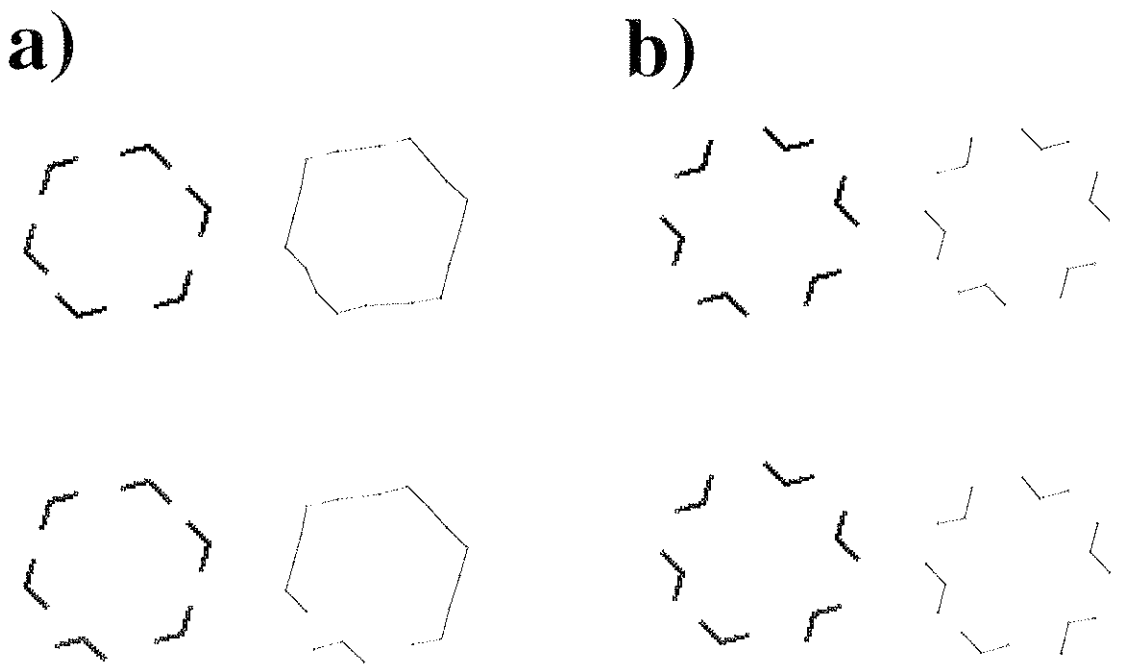


Figure 11: Simulation results for stimuli adapted from (Donnelly et al., 1991) are shown. Input stimuli are shown on the left of each column, with linked GRAF representation shown on the right. Linking across gaps obtained by the GRAF model is consistent in all of these cases with the experimental results of (Donnelly et al., 1991).

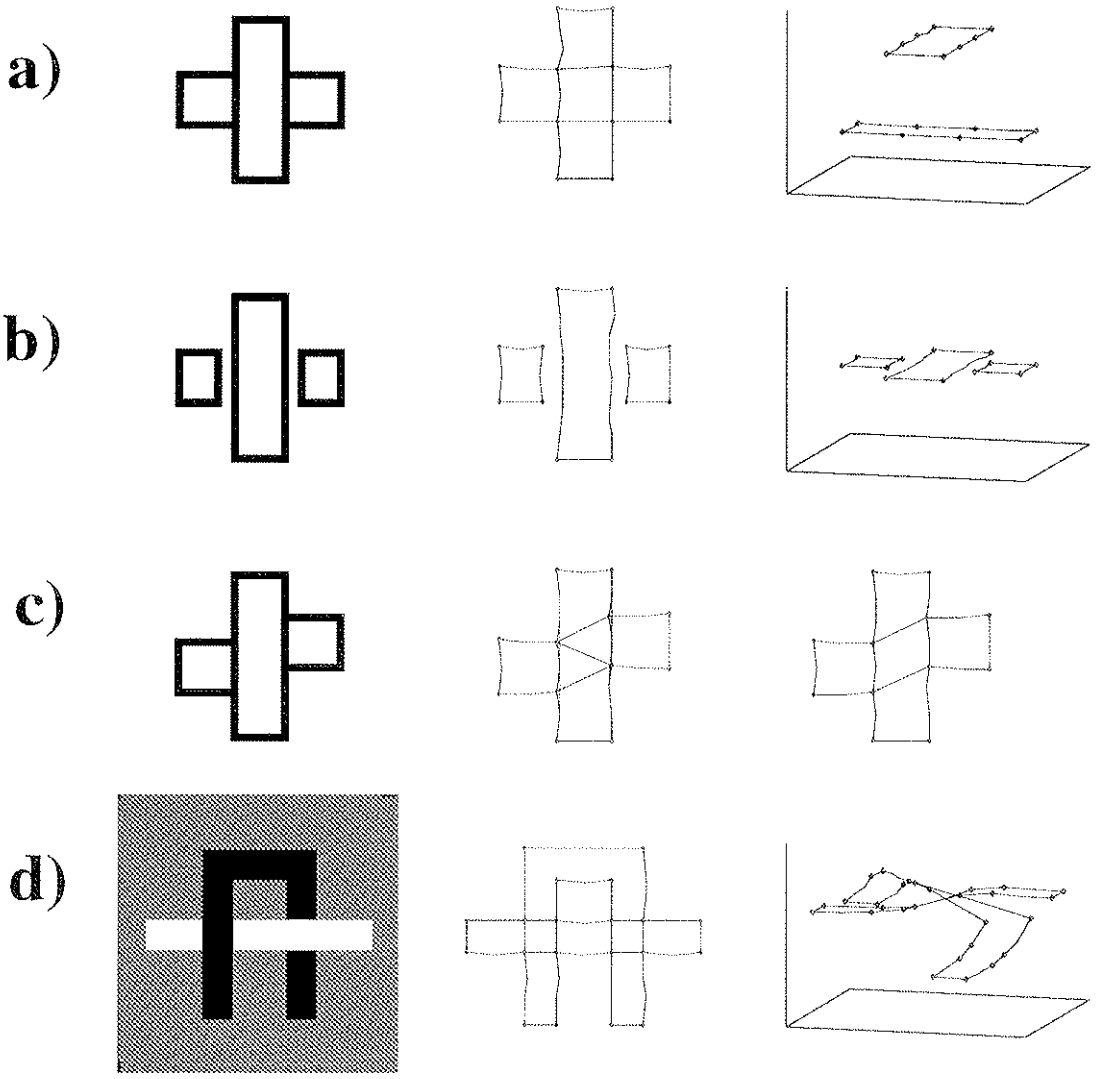


Figure 12: Simulation results demonstrating amodal completion and depth segmentation are shown. The left column shows input images. **a)** Amodal linking of occluded bar (middle), and depth segmentation (right) is shown. **b)** No amodal linking (middle), nor depth segmentation (right), occurs if there are gaps between the segments. **c)** If the right-hand segment is shifted vertically by half its height, then initial amodal links are inconsistent (middle), but inter-arc competition produces correct amodal linking (right). **d)** Depth segmentation bends in a spline-like way to satisfy local depth contraints for a globally inconsistent figure.

the GRAF model can recover 3-D structural information using depth segmentation and (potentially) line labelling processes, which would greatly aid the formation of higher-level compressed representations such as geons.

Finkel & Sajda's model (1992) computes direction of figure, generates illusory contours, and produces occlusion-based depth segmentation. All computations in their model take place in a 2-D lattice, and line junctions are not explicitly represented. A key to their model is the use of tags to "identify elements as belonging to the same object. Tags linking units responding to the same contour are used to determine the direction of figure and to change the perceived depth of the entire contour based on occlusion relationships detected at isolated points (the tag junctions)". While not committing to a mechanism for realizing tags, the authors suggest various temporal coding alternatives. Thus, their model is subject to the intrinsic capacity limitations of temporal coding. Whie avoiding these capacity limitations, the GRAF model in one respect produces a more powerful representation. By explicitly representing junctions, as well as links between them, the GRAF model provides a framework for formation of higher-level representations such as geons, as well as for local line-labelling. Rensink (1992) showed that using purely local processes, constraints on contiguity, convexity, and slant sign, based on junction type, combined with information propagated between line junctions, can reliably produce good (but not perfect) line labelling, resulting in recovery of 3-D orientation compatible with recent psychophysical results. The GRAF model provides a sufficient framework for these processes, which Finkel & Sajda's model does not.

Unlike the above models, the GRAF model achieves bindings without temporal coding, by dynamically binding nodes to locations, then linking nodes together based on their context. With its novel binding architecture, the GRAF model provides a new interpretation for the role of large dendritic fields and fan-in/fan-out of connections besides mere filtering or feature detection. The main innovations of the GRAF model are thus 1) dynamically binding nodes to locations to explicitly code form attributes without a combinatorial explosion, and 2) competing between links which bind form attributes into explicit relationships to provide a powerful framework for depth segmentation and line-labelling processes.

Appendix

All Appendix equations use the indices (x, y) to denote 2-D retinotopic coordinates, (g, h) to denote the 2-D coordinates of nodes, (i, j) to denote the 2-D coordinates of node Position neurons, where are in one-to-one correspondence with spatially offset sections of retinotopic coordinates, and k to denote orientation.

A Feature Extraction

A.1 Complex and End-stopped Neurons

Retinotopic feature extraction produces complex and end-stopped responses (see Heitger et al., 1992), by first applying oriented even- and odd-symmetric Gabor-like filters, at K orientations, to a 100x100 pixel image, I,

$$S^{\lambda}_{x,y,k} = I_{x,y} * D^{\lambda}_{k}, \quad \lambda = (\text{odd}, \text{even}), \tag{1}$$

the outputs of which are combined to produce an initial complex cell response,

$$C_{x,y,k} = \sqrt{(S^{\text{odd}}_{x,y,k})^2 + (S^{\text{even}}_{x,y,k})^2}, \tag{2}$$

which is partially normalized across orientation to produce the final complex cell response,

$$C'_{x,y,k} = \frac{C_{x,y,k}}{\alpha + \sum_{k'=1}^{K} C_{x,y,k'}}. \tag{3}$$

The initial end-stopped response is computed by a differencing operation in each direction along the long axis of complex cells followed by half-wave rectification,

$$E_{x,y,k} = [C'_{x-\Delta_x(k,\delta),y-\Delta_y(k,\delta),k} - C'_{x+\Delta_x(k,\delta),y+\Delta_y(k,\delta),k}]^{+}, \tag{4}$$

$$E_{x,y,k+K} = [C'_{x+\Delta_x(k,\delta),y+\Delta_y(k,\delta),k} - C'_{x-\Delta_x(k,\delta),y-\Delta_y(k,\delta),k}]^{+}, \tag{5}$$

and the final end-stopped response produced by applying side-inhibition to attenuate end-stopped responses next to smooth boundaries,

$$E'_{x,y,k} = [E_{x,y,k} - \beta S_{x,y}]^{+}, \tag{6}$$

where the side inhibition term is defined as

$$S_{x,y} = \frac{1}{2K} \sum_{k=1}^{2K} [\gamma C'_{x+\Delta_y(k,\delta'),y+\Delta_x(k,\delta'),k} - C'_{x,y,k}]^{+}, \tag{7}$$

and spatial offsets are defined as

$$\Delta_x(k, \tau) = \tau \cos((k\pi)/K), \quad \Delta_y(k, \tau) = \tau \sin((k\pi)/K). \tag{8}$$

Filters used in (1) are

$$D_k^{\text{odd}}(p,q) = (2\pi\sigma_x\sigma_y)^{-1}\sin\Big((k\pi)/K\,\upsilon r(\exp(-\lambda r^2/\sigma_y^2)+1)\Big), \tag{9}$$

$$D_k^{\text{even}}(p,q) = (2\pi\sigma_x\sigma_y)^{-1}\cos\Big((k\pi)/K\,\upsilon r(\exp(-\lambda r^2/\sigma_y^2)+1)\Big), \tag{10}$$

where

$$r = \sqrt{(p\cos((\pi k)/K) - q\sin((\pi k)/K))^2 + (p\sin((\pi k)/K) + q\cos((\pi k)/K))^2}. \tag{11}$$

Parameters are $K=12$, $\alpha=4.0$, $\beta=6.0$, $\delta=3.5$, $\delta'=3.0$ $\gamma=1.2$, $\lambda=0.0775$, $\upsilon=0.2$, $\sigma_x=3.0$, $\sigma_y=2.0$.

A.2 Junction and Continuation Saliency Maps

The Junction Saliency Map, SM^J, is computed with partially normalized center/surround contrast enhancement,

$$SM^J = \Gamma^J \frac{[E'_{x,y} * A_1 \;-\; \zeta E'_{x,y} * A_2]^+}{\alpha \;+\; E'_{x,y} * A_1 \;+\; \zeta E'_{x,y} * A_2}, \tag{12}$$

of end-stopped responses averaged across orientation,

$$E'_{x,y} = \frac{1}{2K}\sum_{k=1}^{2K} E'_{x,y,k}. \tag{13}$$

The Connection Saliency Map, SM^C, is computed from the maximum response of complex cells, after contrast enhancement in the direction of their long axis,

$$SM^C = \Gamma^C \max_k([C'_{x,y,k} * B_k]^+). \tag{14}$$

Filters in (12) and (14) are

$$A_\lambda(p,q) = G(p,q,\sigma_\lambda),\ \lambda=(1,2), \tag{15}$$

$$\begin{aligned} B_k(p,q) = {} & 2G(p,q,\sigma_1) - G(p+\Delta_y(k,\sigma_1),\ q+\Delta_y(k,\sigma_1),\ \sigma_1) \\ & - G(p-\Delta_y(k,\sigma_1),\ q-\Delta_y(k,\sigma_1),\ \sigma_1), \end{aligned} \tag{16}$$

where $G(p,q,\sigma) = (2\pi\sigma^2)^{-1}\exp(-(p^2+q^2)/(2\sigma^2))$, and parameters are $\alpha=0.01$, $\zeta=1.25$, $\Gamma^J=10.0$, $\Gamma^C=35.0$, $\sigma_1=2.0$, $\sigma_2=3.0$.

B Feature Abstraction

B.1 Binding Nodes to Locations

A node is bound to a location by activation of a single Position neuron in its Position Map. The term ω indicates whether a variable applies to a Junction node ($\omega=J$) or a Continuation node

($\omega = C$). The coordinates (g, h) of nodes, (i, j) of Position neurons, and (x, y) of Saliency and feature maps are related by

$$(x,\ y) = (g\Lambda^\omega + i,\ h\Lambda^\omega + j), \tag{17}$$

where Λ^ω is the distance between centers of neighboring Position Maps. The GRAF model is applied with a geometry of 9x9 Continuation nodes, each with a 20x20 Position Map, and 7x7 Junction nodes, each with a 40x40 Position Map. Thus, fewer Junction nodes than Continuation nodes are used, but the former are more spatially flexible. Position neurons obey a competitive shunting differential equation (Grossberg, 1973),

$$\frac{d}{dt}P^\omega_{g,h,i,j} = -\alpha P^\omega_{g,h,i,j} + (1 - P^\omega_{g,h,i,j})S^\omega_{g,h,i,j} - P^\omega_{g,h,i,j}\sum_{(i',j')\neq(i,j)} f_P(P^\omega_{g,h,i',j'}), \tag{18}$$

in which the feedback function is threshold linear,

$$f_P(\tau) = \beta[\tau - \gamma]^+, \tag{19}$$

the input signal strength is determined by input from the corresponding saliency map location, modulated by the pathway strength minus feedback suppression,

$$S^\omega_{g,h,i,j} = [SM^\omega_{x,y}(E^{\omega,*}_{i,j} - \zeta F^\omega_{g,h,i,j})]^+, \tag{20}$$

where the input pathway strength,

$$E^{\omega,*}_{i,j} = G^\omega(i - \Upsilon^\omega/2,\ j - \Upsilon^\omega/2) + r^*, \tag{21}$$

has a symmetry breaking stochastic component, $r^* = \mathrm{U}[-0.01 : 0.01]$, and decreases gradually from the center of the Position Map, with

$$G^\omega(i, j) = [1 - \sqrt{i^2 + j^2}/\upsilon]^+, \tag{22}$$

and the feedback suppression of nearby input pathways,

$$F^\omega_{g,h,i,j} = \sum_{\omega'\in\Omega(\omega)}\sum_{g',h',i',j'} f^\omega_P(P^\omega_{g',h',i',j'})H((g - g')\Lambda^\omega + i - i',\ (h - h')\Lambda^\omega + j - j'), \tag{23}$$

in which $(\omega', g', h', i', j') \neq (\omega, g, h, i, j)$, decreases with distance,

$$H(i, j) = \exp(-(i^2 + j^2)/(2\sigma_h^2))\ \text{ if } |i|, |j| \leq \Gamma,\ \ 0 \text{ otherwise.} \tag{24}$$

Feedback suppression is asymmetrical between Junction and Continuation nodes, Junction nodes enjoying a competitive advantage with $\Omega(J) = (J)$, and $\Omega(C) = (C, J)$. Parameters are $\alpha = 0.2$, $\beta = 1.0$, $\gamma = 0.1$, $\upsilon = 100.0$, $\sigma_h^2 = 6.0$, $\Lambda^J = \Lambda^C = 10$, $\Upsilon^C = 20$, $\Upsilon^J = 40$, $\Gamma = 6$.

B.2 Activation of Form Attribute Neurons

Each node contains a set of M^ω Form Attribute (FA) neurons. The FA neuron with best matching template for complex and end-stopped cells, applied where the node is bound, is chosen.

$$FA^\omega_{g,h,m} = 1 \text{ if } \text{Match}^\omega(g,h,m) > \text{Match}^\omega(g,h,m') \; \forall \; m' \neq m, \;\; 0 \text{ otherwise,} \tag{25}$$

where

$$\begin{aligned}\text{Match}^\omega(g,h,m) = \sum_{i,j} f^\omega_P(P^\omega_{g,h,i,j}) \Bigg(& \sum_{k=1}^{K} W^\omega_{g,h,i,j,m,k} C'_{x,y,k} + \\ & \sum_{k=1}^{2K} W'^\omega_{g,h,i,j,m,k+K} C'_{x+\Delta_x(x,\delta),y+\Delta_y(y,\delta),k} + \\ & \sum_{k=1}^{2K} W^\omega_{g,h,i,j,m,k+3K} E'_{x,y,k} \Bigg),\end{aligned} \tag{26}$$

and a node is bound only if it has a Position neuron above threshold,

$$f^\omega_P(\tau) = f_P(\tau) \;\text{ if } f_P(\tau) > \Psi^\omega, \;\; 0 \text{ otherwise.} \tag{27}$$

Parameters are $\delta = 7.0$, $\Psi^J = 0.2$, $\Psi^C = 0.16$

The details of the weight templates W are not specified. Note that only $M^J + M^C$ different templates exist, however, each consisting of $5K$ components. The indices (g,h) indicate which node the template belongs to, and (i,j) the template's spatial position within a dendritic field. All spatial positions within a dendritic field contain a copy of the same template.

C Linking Form Attributes

C.1 Initial Estimate of Linking Strength

C.1.1 Grouping neurons at each node

Each node has $2K$ grouping neurons which code the grouping strength in each of the $2K$ directions. Grouping strength is a function of a combination of the winning FA neuron, indexed by M, and the feature signals from complex (C') and end-stopped (E') neurons.

$$\begin{aligned}G'^\omega_{g,h,k} = \; & \Phi^\omega_1 \sum_{k' \in \Theta(M,k)} W^\omega_{g,h,i,j,m,k'} f_G(C'_{x,y,k'}) + \\ & \Phi^\omega_2 \sum_{k' \in \Theta(M,k)} W^\omega_{g,h,i,j,m,k'+K} f_G(C'_{x+\Delta(x,\delta),y+\Delta(y,\delta),k'}) + \\ & \Phi^\omega_3 \sum_{k' \in \Theta(M,k)} W^\omega_{g,h,i,j,m,k'+3K} f_G(E'_{x,y,k'}),\end{aligned} \tag{28}$$

where featural signals are compressed, $f_G(\tau) = \tau/(\alpha + \tau)$, amodal grouping takes place in the direction opposite to a termination,

$$\Theta(M,k) = \begin{cases} (k,\ (k+K) \bmod 2K) \text{ if } M \text{ indexes a Termination or T-junction,} \\ (k) \text{ otherwise,} \end{cases} \tag{29}$$

and parameters are $\alpha = 0.1$, $\delta = 7.0$, $\Phi_1^J = 0.0$, $\Phi_1^G = 1.0$, $\Phi_2^J = \Phi_2^G = 0.5$, $\Phi_3^J = \Phi_3^G = 1.0$.

C.1.2 Recovery of spatial relation between nodes

Activation of Distance and Angle neurons is approximated by directly calculating the distance and angle between nodes based on their maximally active Position neurons, $P^{\omega}_{g,h,i,j}$ and $P^{\omega'}_{g',h',i',j'}$,

$$\text{Dist}(g,h,g',h') = \sqrt{(x-x')^2 + (y-y')^2}, \tag{30}$$

$$\text{Angle}(g,h,g',h') = K/\pi \arctan(y-y', x-x'), \tag{31}$$

where

$$(x,y),\ (x',y') = (g\Lambda^{\omega} + i, h\Lambda^{\omega} + j),\ (g'\Lambda^{\omega'} + i', h'\Lambda^{\omega'} + j'). \tag{32}$$

An arc only exists if the maximum possible distance between its pair of nodes falls below a limit, which is implemented as 30.0, 100.0, and 200.0 for arcs joining two Continuation nodes, a Continuation node with a Junction node, and two Junction nodes, respectively, resulting in a total of about 9,000 arcs.

C.1.3 Input to arc's Linking neurons

The initial estimate of linking strength is the input to a Linking neuron, $I^{\omega,\omega'}_{g,h,g',h',k}$, which is determined by the angle and distance between the nodes, combined with the nodes' grouping signals. The identity of the two nodes are given by (g,h,ω) and (g',h',ω').

$$\begin{aligned} I^{\omega,\omega'}_{g,h,g',h',k} \ = \ & G^{i\omega}_{g,h,k} G^{i\omega'}_{g',h',(k+K)\bmod 2K} \\ & \exp\left(-(\Theta_{\text{diff}}(k, \text{Angle}(g,h,g',h')))^2/(2\sigma_\theta^2) - (\text{Dist}(g,h,g',h'))^2/(2\sigma_d^2)\right) \\ & + \text{Amodal}(\omega,g,h,\omega',g',h',k), \end{aligned} \tag{33}$$

where the angle difference function is

$$\Theta_{\text{diff}}(k,k') = \min\left((k-k'+2K) \bmod 2K,\ |k-k'-2K| \bmod 2K\right), \tag{34}$$

and the amodal component is

$$\begin{aligned} \text{Amodal}(\omega,g,h,\omega',g',h',k) = & \sum_{k'} G^{i\omega}_{g,h,k'} G^{i\omega'}_{g',h',k''} \\ & \exp\left(-(\Theta_{\text{diff}}(k, \text{Angle}(g,h,g',h')))^2/(2\sigma_{\theta'}^2) \ - \ (\text{Dist}(g,h,g',h'))^2/(2\sigma_{d'}^2)\right) \end{aligned} \tag{35}$$

where k', k'' correspond to orientations of amodal grouping signals, and are cocircularly related,

$$k'' = (k' + K + 2\Theta_{\text{diff}}(k, \text{Angle}(g,h,g',h'))) \bmod 2K. \tag{36}$$

Parameters are $\sigma_\theta = 6.0$, $\sigma_d = 12.5$, $\sigma_{\theta'} = 3.8$, $\sigma_{d'} = 25.0$.

C.2 Competition Between Links

Activation of Linking neurons L obey an on-center off-surround competitive shunting differential equation,

$$\frac{d}{dt}L^{\omega,\omega'}_{g,h,g',h',k} = -L^{\omega,\omega'}_{g,h,g',h',k} + (1 - L^{\omega,\omega'}_{g,h,g',h',k})\text{Excite} - L^{\omega,\omega'}_{g,h,g',h',k}\text{Inhib}, \tag{37}$$

where

$$\text{Excite} = I^{\omega,\omega'}_{g,h,g',h',k} + (L^{\omega,\omega'}_{g,h,g',h',k})^2, \tag{38}$$

$$\text{Inhib} = \sum_{\omega''\in(J,C)}\sum_{g''h''}\sum_{k'}(L^{\omega,\omega''}_{g,h,g'',h'',k'} + L^{\omega'',\omega'}_{g'',h'',g',h',k'})\Theta_{\text{falloff}}(k,k') - 2L^{\omega,\omega'}_{g,h,g',h',k}\Theta_{\text{falloff}}(k,k), \tag{39}$$

in which inhibition decreases with orientational difference,

$$\Theta_{\text{falloff}}(k,k') = 2(\sqrt{2\pi}\sigma_\theta)^{-1}\exp(-(\Theta_{\text{diff}}(k,k'))^2/(2\sigma_\theta^2)), \tag{40}$$

and $\sigma_\theta = 1.5$.

D Depth Segmentation

Each node contains a set of N depth neurons which coarse code depth. A node's Depth neurons are excited and inhibited by Depth neurons of other nodes it is linked to. In addition, if a node codes a T-junction, then it has two sets of N Depth neurons, one set representing the depth at the top bar of the T, and the other set representing the depth at the terminating stem of the T.

Depth signals between nodes are gated by the summed activity of Linking neurons at their connecting arc,

$$\text{Link}^{\omega,\omega'}_{g,h,g',h'} = \sum_{k=1}^{2K} L^{\omega,\omega'}_{g,h,g',h',k}. \tag{41}$$

Input to a node is distinguished by the Form Attribute stem it belongs to (an L junction has 2 stems, an Arrow junction has 3 stems). Different inputs to the same stem s are summed, and the summed inputs to different stems are multiplied. Finally, excitatory and inhibitory input to the nth Depth neuron is determined by convolution of other Depth neurons, centered at n, with the excitatory and inhibitory interaction kernels.

$$\frac{d}{dt}D^{\omega}_{g,h,n} = -\alpha D^{\omega}_{g,h,n} + \beta + (1 - D^{\omega}_{g,h,n})\text{Excite} - D^{\omega}_{g,h,n}\text{Inhib}, \tag{42}$$

where

$$\text{Excite} = \prod_s \sum_{g',h',\omega'} \text{Link}^{\omega,\omega'}_{g,h,g',h'} \sum_{n'=1}^{N} \text{Cen}(n-n')D^{\omega'}_{g',h',n'}, \tag{43}$$

$$\text{Inhib} = \prod_s \sum_{g',h',\omega'} \text{Link}^{\omega,\omega'}_{g,h,g',h'} \sum_{n'=1}^{N} \text{Sur}(n-n')D^{\omega'}_{g',h',n'} \tag{44}$$

in which stem membership functions (not specified here) are $s \in \text{Stems}(g,h,\omega)$ and $g',h',\omega' \in \text{Stem}(\omega,g,h,s)$. The excitatory (center) and inhibitory (surround) kernels are

$$\text{Cen}(n) = \gamma_c \exp(-n^2/(2N^2\sigma_c^2)), \tag{45}$$

$$\text{Sur}(n) = \gamma_s(1 - \exp(-n^2/(2N^2\sigma_s^2))). \tag{46}$$

Finally, cross inhibition is added to Inhib in (44) if the form attribute is a T-junction, in order to push the top bar of the T-junction to a higher depth, and the terminating stem of the T-junction to a lower depth,

$$\text{Cross}(g,h,\omega) = \begin{cases} \sum_{n'=1}^{N} \text{Up}(n-n') D'^{,\omega}_{g,h,n'} \text{ if top bar,} \\ \sum_{n'=1}^{N} \text{Down}(n-n') D'^{,\omega}_{g,h,n'} \text{ if bottom stem,} \end{cases} \tag{47}$$

where

$$\text{Up}(n) = 1 + n/(\varepsilon + |n|), \quad \text{Down}(n) = 1 - n/(\varepsilon + |n|), \tag{48}$$

and D' denotes the other set of Depth neurons at the same node. Parameters are $N = 10$, $\alpha = 0.1$, $\beta = 0.1$, $\gamma_c = 3.0$, $\gamma_s = 1.0$, $\varepsilon = 0.5$, $\sigma_c = \sigma_s = 0.075$.

Reference

Barlow, H. B. (1981). Critical limiting factors in the design of the eye and visual cortex. *Proc. R. Soc. (London), B212*, 1–34.

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

Desimone, R. (1992). Neural circuits for visual attention in the primate brain. In G. A. Carpenter, S. G. (Ed.), *Neural networks for vision and image processing*, pp. 343–364. Cambridge, MA: MIT Press/Bradford Books.

Desimone, R., & Schein, S. J. (1987). Visual properties of neurons in area v4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology, 57*, 835–868.

Donnelly, N., Humphreys, G., & Riddoch, M. (1991). Parallel computation of primitive shape descriptions. *Journal of Experimental Psychology, 17*, 561–570.

Engel, A., König, P., Kreiter, A., Schillen, T., & Singer, W. (1992). Temporal coding in the visual cortex: new vistas on integration in the nervous system. *Trends in Neurosciences, 15*, 218–226.

Enns, J. T., & Rensink, R. A. (1991). Preattentive recovery of three-dimensional orientation form line drawings. *Psychological Review, 98*, 335–351.

Enns, J. T., & Rensink, R. A. (1994). An object completion process in early vision.. In A. Gale (Ed.), *Visual search III*. London: Taylor & Francis.

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

Finkel, L., & Sajda, P. (1992). Object discrimination based on depth-from-occlusion. *Neural Computation, 4*, 901–921.

Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics, LII*, 213–257.

Grossberg, S. (1994). 3-d vision and figure-ground separation by visual cortex. *Perception & Psychophysics, 55(1)*, 48–120.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of perceptual grouping: Textures, boundaries, and emergent segmentations. *Perception and Psychophysics, 38*, 141–171.

Heitger, F., Rosenthaler, L., von der Heydt, R., Peterhans, E., & Kubler, O. (1992). Simulation of neural contour mechanisms: from simple to end-stopped cells. *Vision Research, 32(5)*, 963–981.

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

Hubel, D. H., & Livingstone, M. S. (1985). Complex-unoriented cells in a subregion of primate area 18. *Nature Lond.*, *315*, 325–327.

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

Hummel, R. A., & Zucker, S. W. (1983). On the foundation of relaxation labeling processes. *IEEE PAMI*, *5*, 267–287.

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, *23*, 141–221.

Lowe, D. (1987a). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, *31*, 355–395.

Lowe, D. (1987b). The viewpoint consistency constraint. *Int. J. Computer Vision*, *1*, 57–72.

Mackworth, A. (1976). Model-driven interpretation in intelligent vision systems. *Perception*, *5*, 349–370.

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

Mohan, R., & Nevatia, R. (1992). Perceptual organization for scene segmentation and description. *IEEE PAMI*, *14*, 616–635.

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

Morrone, M., & Burr, D. (1988). Feature detection in human vision: A phase-dependent energy model. *Proc. R. Soc. Lond. B*, *235*, 221–245.

Nakayama, K., & Shimojo, S. (1990). Toward a neural understanding of visual surface representation. *Cold Spring Harbor Symp. Quant. Biol.*, *40*, 911–924.

Parent, P., & Zucker, S. W. (1989). Trace inference, curvature consistency, and curve detection. *IEEE PAMI*, *11*, 823–839.

Rensink, R. A. (1992). The rapid recovery of three-dimensional orientation from line drawings.. Ph.D. Thesis (also Technical Report 92-25), Department of Computer Science, University of British Columbia, Vancouver, BC, Canada.

Rensink, R. A., & Enns, J. T. (1994). Pre-emption effects in visual search: Evidence for low-level grouping.. *Psychological Review*.

Treisman, A. (1985). Preattentive processing in vision.. *Computer Vision, Graphics, and Image Processing*, *31*, 156–177.

Van Essen, D. C., & Anderson, C. H. (1990). Information processing strategies and pathways in the primate retina and visual cortex. In Zornetzer, Davis, L. (Ed.), *An introduction to neural and electronic networks*, pp. 43–72. Academic Press.

Wang, D., Buhmann, J., & von der Malsburg, C. (1990). Pattern segmentation in associative memory. *Neural Computation*, *2*, 94–106.

Williamson, J. (1993). Dynamic binding of visual contours without temporal coding. *WCNN93*, *1*, 97–100.