Visual Segmentation and Feature Binding without Synchronization

Robert Thau (rst@ai.mit.edu)*
MIT Dept. of Brain and Cognitive Sciences
MIT, room NE43-711
Cambridge, Ma., 02139

INTRODUCTION

There has recently been a sizable amount of work on visual segmentation in neural networks. Many of these networks have functioned by "tagging" the outputs of feature detectors as being features of one object or another. In particular, several models have appeared in which visual segmentation is performed by a network of oscillators; different objects are discriminated by setting the oscillators representing the features of a given object in phase with each other, but out of phase with the features of other objects in the scene [4, 2, 7]. However, it is still not entirely clear whether synchronization among units, be they phase-locked or otherwise, actually plays a role in the functioning of primate visual cortex, nor what role that might be.

The network described here exploits behavioral constraints on the segmentation task to solve it with units whose outputs have no special features for binding or tagging at all. The network uses a process analogous to constraint propagation among Waltz labels for the scene — in particular, the propagation of a "focus" label for the object being sent to higher centers, and the placement of other labels on features of surroundings which describe spatial relations between them and the object in focus.

The particular constraint we exploit most fully is the limit on the capacity of the primate visual system to process multiple objects in an image simultaneously. For instance, when searching for complicated targets amidst distractors, human subjects are forced to do a linear search. (The exceptions are that such primitive features as "tilted left" or "red", which can apparently be detected without any segmentation whatever). So, the segmentation task, in primates, appears to be constrained to serve

*This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences, and at the Artificial Intelligence Laboratory. This research is sponsored by grants from the Office of Naval Research under contracts N00014-91-J-1270 and N00014-92-J-1879; by a grant from the National Science Foundation under contract ASC-9217041 (funds provided by this award include funds from DARPA provided under the HPCC program); and by a grant from the National Institutes of Health under contract NIH 2-S07-RR07047. Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Sumitomo Metal Industries, and Siemens AG. Support for the A.I. Laboratory's artificial intelligence research is provided by ONR contract N00014-91-J-4038.

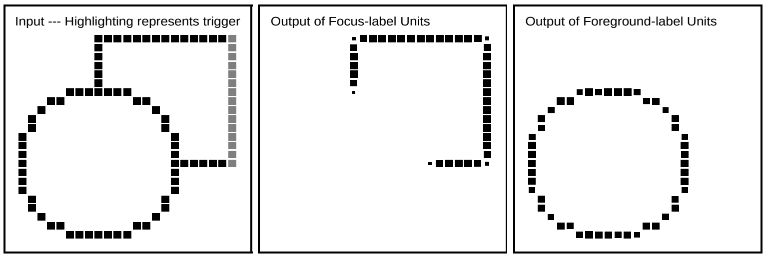

Figure 1: Activity in the network once it has finished segmentation. Activity of input edge detectors is indicated on the left. Highlighted items indicate loci of trigger activity — that is, the signal to the network that it should focus on the object containing the features at those locations. Output of "focus" labeled features after the network has settled is in the middle; output of "foreground" labeled features is on the right.

up one object at a time to higher centers — in other words, to cause them to attend to one object in the scene and ignore the others.

THE VISUAL SEGMENTATION TASK RECONSIDERED

Reducing the segmentation task to attentional subselection does not render it trivial. The structure of images has an important role in human visual perception — good parses of this structure can aid perception; misparses can block it completely. Kanizsa [5] gives numerous examples, and there is also direct psychophysical evidence of the importance of a grouping-related process in recognition and related activities [8, 1].

In particular, a simple spotlight can capture inappropriate features. For example, consider the object in the background of the sample input in figure 1. While we cannot know the exact shape of that object (part of it being obscured), it is unlikely, on the general-view assumption, that it really is a square with a quarter-circle wedge cut out of it. Therefore, we should exclude the features of the circle in the foreground when figuring out what this object actually is. A fixed-shape, convex spotlight of attention cannot do this — there is no place to put it which captures all the features of the square without grabbing a piece of the circle as well. More sophisticated machinery is needed to do the job properly.

On the other hand, if one ties the attentional mechanism very closely to object recognition per se (as in Fukushima and Imagawa [3]), one has trouble segmenting scenes containing unfamiliar objects.

NETWORK STRUCTURE

Putting all of the above criteria together, the task of the network is to segregate out one object from the surroundings, and to characterize its relationship with those surroundings.

We now wish to create a network to solve this problem. In overview, we do this as follows: We create feature maps representing features such as "there is a horizontal

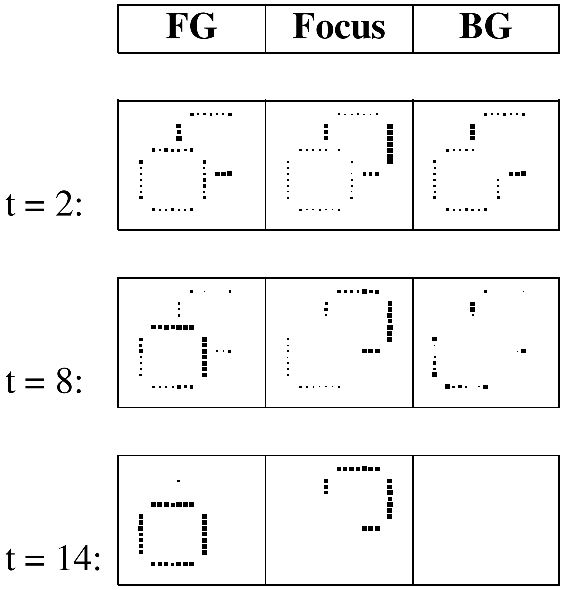

Figure 2: Activity within the network's three sets of feature maps over several timesteps. Initially, activity in all feature maps is identical except for the small region (one line at upper right) covered by the trigger input. Over a small number of time steps, the mutually excitatory and inhibitory connections cause the network to settle into a correct segmentation.

edge segment here which is part of the attended object" (called here the *focus label*), or "... part of an object in the foreground". The network includes excitatory connections between units representing compatible hypotheses, and mutually inhibitory connections between units representing incompatible hypotheses. There are also detectors for figural primitives, such as corner detectors, which play a role in arbitrating between competing hypotheses.

Most of the inputs to this network are simply oriented edge detectors. There is also an input (the "trigger" input) from the putative higher centers, pointing out features which probably belong to an object of interest, whose exact boundaries are not even roughly known to the higher centers. Thus, the network in figure 1 is triggered with one line, at the edge of the image — it is the task of the network to figure out what is connected to that line (via corners) and what is not (even though it may be physically adjacent — e.g. via tee junctions).

These inputs feed into the mutually reinforcing and inhibiting feature detectors described above. After some amount of (simulated) time, the only units left active are those representing a single, consistent parse of the scene, with the entire object of interest bearing the "focus" label. Higher centers that wish to further process the features of that object (e.g., to classify it) may simply read the outputs of those units. A history of the network processing a simple stimulus is shown in figure 2.

Patterns of Connections

The bulk of the network is organized into the three sets of feature maps visible in figure 2. Once the network has reached statsis, these represent the features of the object in focus (in the "focus" maps), and the features of adjacent objects in front of it and behind it (in the "foreground" and "background" maps, respectively). These are, in turn, divided up into oriented submaps. Therefore, most units of the network represent oriented borders of the object in focus, or some releated object.

When the network begins operating, of course, almost all of the features on the input have *not* been classified as to their proper placement. Doing that classification is, after all, the network's job. (The exceptions are the specific features which have been singled out by the trigger input as forming the "nucleus" of the object in focus). And indeed, looking at the history of network activity in figure 2 shows that, initially, the activity in the three sets of maps is more or less the same. The network must therefore move from a state in which only the features covered by the trigger are classified, to a state in which everything is classified. The connections within and between the feature maps are set up to accomplish this.

For instance, consider the case of colinear features. If part of a line is known to be part of the object focus, a colinear feature farther along the line should also be in focus. We therefore create mutually excitatory connections between the units representing colinear features within the "focus" map. Likewise, since colinear features should probaly not be assigned to the foreground and background maps, we create mutually inhibitory connections between colinear features in these maps.

Network Micro-Architecture

Most connections in the network consist of this straightforward mutual excitation or inhibition, as in the connections between units representing compatible or competing explanations of adjacent, collinear edge segments (which are mutually excitatory or inhibitory, respectively). However, in some cases this just isn't adequate.

For instance, if there is a corner at a particular spot, then the line segments forming the sides of the corner are probably part of the same object, and we would want the focus-map units on one edge of the corner to excite the (differently-oriented) focus-map units on the other. However, extending one of the lines forming the corner yields a T-junction. In this case, the same primitive edge segments are all present (as well as a few new ones), but they are now very probably parts of different objects. So, with the T, we would *not* want the mutual reinforcement to occur.

We deal with this by creating the excitatory connection, but gating it with the output of a corner detector, using a multiplicative mechanism. The inputs from feature units "around the corner" are multiplied by the output of an appropriately oriented corner detector; if the corner detector is off (if it is being suppressed by the presence of a T-junction), the mutual reinforcement is ineffective. The use of multiplication, and the mechanism we use, is biologically plausible [9]. In any case, it is possible to achieve the same effects without invoking multiplication at all by creating interneurons with gerrymandered thresholds (as we did in an earlier version of the network, at the price of a substantial number of basically superfluous units).

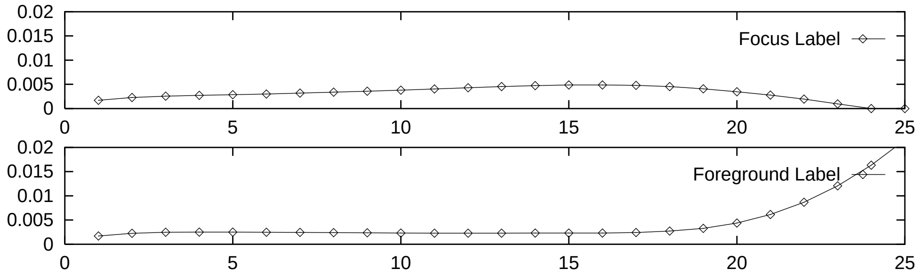

Figure 3: Activity traces from units representing alternative labelings of the same feature; units along the X axis are simulated timesteps. Labels traced are for the lower left of the circle in figure 1, the farthest point in the image from the trigger. Note that the focus-label unit dies off, and the foreground-label unit (representing the correct classification) is reinforced, once constraint propagation has reached this point in the image.

Details of Individual Units

The individual units in the network (aside from inputs whose values are fixed externally) are simple integrators, whose outputs follow the differential equation

$$\frac{dx_i}{dt} = \sigma_\lambda(\mathrm{I}_i) - \epsilon x_i$$

where x_i is the activation of the i'th unit, I_i is its aggregate input, σ_λ is the usual sigmoid squashing function, and ϵx_i is a damping term to keep the activity of a unit which is being continually excited from growing to infinity.

The only unusual thing about these units is the form of the input I_i. In addition to a weighted sum of the activity of other units,

$$\mathrm{I}_i = \sum w_{ij} x_j + \cdots$$

where w_{ij} is the matrix of connection weights among the units, we allow products of such sums:

$$\mathrm{I}_i = \sum w_{ij} x_j + (\sum u^1_{ij} x_j)(\sum v^1_{ij} x_j) + (\sum u^2_{ij} x_j)(\sum v^2_{ij} x_j) + \cdots$$

where $u^1_{ij}, v^1_{ij}, u^2_{ij}, v^2_{ij}$ are again matrices of connection weights. These are the products of sums of inputs discussed above. Figure 3 shows activity of two units from the lower left of the circle in figure 1; note that there is no oscillatory activity necessary for them to achieve their correct final states.

POSSIBLE PHYSIOLOGICAL CORRELATES

Figure 3 displays outputs of two units representing alternative labelings of the same edge feature. In this case, the feature in question is at the lower left of the circle, i.e., in the foreground; note that the inappropriate focus label for this feature dies off once constraint propagation has made its way from the triggered features through the tee junctions and around the circle.

In an actual brain, most of the units of the form pathway would presumably be devoted to the analysis of the attended object. If a mechanism such as this were operating in the brain, we would therefore expect most units in the temporal (i.e., the form) pathway to exhibit the same characteristics as the focus-labeled units here, that is, their responses to non-attended objects would be reduced or suppressed when the animal was looking elsewhere.

In fact, there have been reports of exactly this sort of thing in the literature on primate visual cortex for some time. For instance, Moran and Desimone [6] studied responses of intermediate-level feature detectors in the primate form pathway (in particular, V4 and IT). They find responses of cells responding to the features of one object are indeed selectively suppressed when the monkey is required to attend to a different object. No such suppression was found in primary visual cortex (nor is it required of the primitive edge detectors in our model).

References

[1] Mercedes Bachilon Ben-av, Dov Sagi, and Jochen Braun. Visual attention and perceptual grouping. *Perception and Psychophysics*, 52(3):277–294, 1992.

[2] Leif Finkel and Paul Sajda. Object discrimination based on depth from occlusion. *Neural Computation*, 4(6):901–921, 1992.

[3] Kunihiko Fukushima and Taro Imagawa. Recogntion and segmentation of connected characters with selective attention. *Neural Networks*, 6(1):33–42, 1993.

[4] John E. Hummel and Irving Biederman. Dynamic binding in a neural network for shape recognition. Unpublished ms., July 1991.

[5] Gaetano Kanizsa. *Organization in Vision: Essays on Gestalt Perception*. Praeger, New York, 1981.

[6] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[7] Michael Mozer, Richard Zemel, Marlene Behrmann, and Christopher Williams. Learning to segment images using dynamic feature binding. *Neural Computation*, 4(5):650–65, 1992.

[8] Ken Nakayama, Shinsuke Shimojo, and Gerald H. Silverman. Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception*, 18(1):55–68, 1989.

[9] Tomaso Poggio and Christof Koch. Multiplying with synapses and neurons. In T. McKenna, J. Davis, and S.F. Zorenter, editors, *Single Neuron Computation*, pages 315–345. Academic Press, Boston, 1992.