Discovering and using perceptual grouping principles in visual information processing

Michael C. Mozer
Department of Computer Science &
Institute of Cognitive Science
University of Colorado
Boulder, CO 80309-0430

Richard S. Zemel
Department of Computer Science
University of Toronto
Toronto, Ontario M5S 1A4

Marlene Behrmann
Department of Psychology & Rotman
Research Institute of Baycrest Centre
University of Toronto
Toronto, Ontario M5S 1A1

Abstract

Despite the fact that complex visual scenes contain multiple, overlapping objects, people perform object recognition with ease and accuracy. Psychological and neuropsychological data argue for a *segmentation* process that assists in object recognition by grouping low-level visual features based on which object they belong to. We review several approaches to segmentation/recognition and argue for a bottom-up segmentation process that is based on feature grouping heuristics. The challenge of this approach is to determine appropriate grouping heuristics. Previously, researchers have hypothesized grouping heuristics and then tested their psychological validity or computational utility. We suggest a basic principle underlying these heuristics: they are a reflection of the structure of the environment. We have therefore taken an adaptive approach to the problem of segmentation in which a system, called MAGIC, *learns* how to group features based on a set of presegmented examples. Whereas traditional grouping principles indicate the conditions under which features should be bound together as part of the same object, the grouping principles learned by MAGIC also indicate when features should be segregated into different objects. We describe psychological studies aimed at determining whether limitations of MAGIC correspond to limitations of human visual information processing.

Recognizing an object in a visual scene involves matching a collection of visual features in the scene that correspond to the object against stored object models. In scenes that contain multiple objects, the matching process alone is insufficient for recognition because it presumes that the features are partitioned by object. Consequently, a complete model of scene recognition requires the ability to *group* features of an object together, or equivalently, to *segment* the scene into regions corresponding to different objects. Psychophysical and neuropsychological evidence suggests that the human visual system possesses such an ability (Duncan, 1984; Farah, 1990; Kahneman & Henik, 1981; Treisman, 1982).

Models of scene recognition can be divided roughly into three classes (Figure 1). *Interactive* models are based on the observation that the scene cannot be properly segmented until object identities are known, yet objects cannot be properly identified until they are segmented. Consequently, segmentation and matching form an iterative cycle in which the matching system can propose refinements of the initial segmentation, which in turn refines the output of the matching system, and so forth (Hinton, 1981; Hinton, Williams, & Revow, 1992; Hanson & Riseman, 1978; Waltz, 1975). The problem with this approach is that it involves a simultaneous search for a good segmentation and a good interpretation of the data. We are skeptical about the computational feasibility of such massive combinatorial searches; they are slow and often lead to local optima in the search space (e.g., Hinton & Lang, 1985).

Bottom-up models are based on the premise that matching processes can be devised that do not require a precise segmentation (e.g., Mozer, 1992). Consequently, segmentation can be viewed as an early heuristic process that depends solely on low-level features. The results of segmentation are fed to the matching system, but the matching system does not directly influence segmentation. Although the heuristics used to group features will not be infallible, the hope is that they will suffice for most recognition tasks (Enns & Rensink, 1992). In cases where recognition fails the first time around, the segmentation can be adjusted and the process restarted. Although this restarting procedure is iterative, iteration is the exception, in contrast to the interactive model which intrinsically relies on an iterative constraint-satisfaction process to perform segmentation and matching jointly. The difficulty with the bottom-up approach is that an adequate set of grouping heuristics is required. We return to this issue later.

*This research was supported by NSF PYI award IRI-9058450, grant 90-21 from the James S. McDonnell Foundation, and DEC external research grant 1250 to MM, and by a National Sciences and Engineering Research Council Postgraduate Scholarship to RZ. Our thanks to Chris Williams, Paul Smolensky, Radford Neal, Geoffrey Hinton, and Jürgen Schmidhuber for helpful comments regarding this work.

Interactive and bottom-up models attempt to achieve object-based segmentation of the scene. That is, features of an object are collected together even if the features are noncontiguous in space and overlap with features of other objects. An alternative approach, *location-based* segmentation, simply determines a coherent region of space that is sufficiently large to be assured of containing all features of a single object, even if the features of other objects are present in that region. The hope is then to devise a matching process that can ignore irrelevant context surrounding the object of interest. It would seem quite difficult to achieve this robust a matching process. Recently, however, Rumelhart (1992; Keeler & Rumelhart, 1992) have proposed such a system using neural net learning techniques. The claim is that learning will find cues reliably indicating the presence of an object regardless of the context in which it is embedded. Even if such cues exist for real-world scenes—and of this we are skeptical—this class of model is inconsistent with the previously mentioned data indicating that people perform object-based grouping of featural information.

Note that the interactive and bottom-up models do not deny the possibility of location-based selection. Indeed, prior to the operation of these models, a spatial focus of attention may well be applied to select the general region of interest. Such a preselection stage would simplify the object-based segmentation task.

Our conviction is that the interactive and location-based models have serious complications both in terms of computational feasibility and psychological validity. We have thus turned to the bottom-up model and attempted to overcome its limitations. The primary concern is whether, based on information from the scene alone, a set of grouping heuristics exist that can determine which features belong together.

Gestalt psychologists have suggested a variety of grouping principles that govern human perception. In exploring how people group elements of a display, evidence has been found for grouping of elements that are close together in space or time, that appear similar, that move together, or that form a closed figure (Rock & Palmer, 1990). There is a long history of attempts by the computer vision community to turn these principles into grouping heuristics, with a fair degree of success (e.g., Lowe & Binford, 1982). The degree of success depends on the ingenuity of the researchers in proposing an adequate set of heuristics.

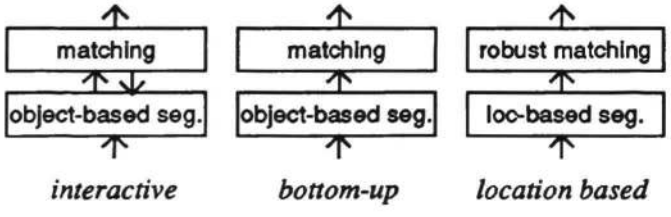

Figure 1: Three classes of object recognition models

We believe there is a more basic principle that underlies these heuristics and that can be used to suggest better heuristics. Namely, these grouping heuristics are a reflection of the structure of the environment. Preliminary evidence from neurobiology suggests that experience can indeed affect the strength of synaptic connections that may play a role in perceptual grouping (Löwel & Singer, 1992).

We have therefore taken an *adaptive* approach to the problem of segmentation in which a system learns how to group features based on a set of examples. We call our system MAGIC, an acronym for multiple-object adaptive grouping of image components. In many cases MAGIC discovers grouping heuristics similar to those proposed in earlier work, but it also has the capability of finding nonintuitive structural regularities in scenes. Hummel and Biederman (1992) have also discussed the possibility of discovering grouping heuristics based on environmental regularities.

MAGIC is trained on a set of presegmented scenes containing multiple objects. By "presegmented", we mean that each feature is labeled as to which object it belongs. MAGIC learns to detect configurations of the scene features that have a consistent labeling in relation to one another across the training examples. Identifying these configurations then allows MAGIC to label features in novel, unsegmented scenes in a manner consistent with the training examples.

This training procedure is a form of supervised learning. Of course, the real world does not directly provides such examples to a learner. However, there is a wealth of information in the environment that can supply the supervision. Perhaps the most important piece of information is motion. A rigid object moving in the plane perpendicular to the line of sight will have the property that all of its features travel across the visual field with the same velocity vector. Thus, by designing the learning system to treat velocity information as a training signal, the system can discover grouping principles that will also apply to stationary objects. Evidence from developmental psychology indeed suggests that the representation of object unity is initially derived from motion (Spelke, 1990).

The Domain

Our initial work has been conducted in the domain of two-dimensional geometric contours. The contours are constructed from four primitive feature types—oriented line segments at 0°, 45°, 90°, and 135°—and are laid out on a 25 × 25 grid. At each location on the grid are units, called *feature units*, that represent each of the four primitive feature types. In our present experiments, scenes contain two contours. We exclude scenes in which the two contours share a common edge. This permits a unique labeling of each feature. Examples of several randomly generated scenes containing rectangles and diamonds are shown in Figure 2. Although the scenes we have tested are composed only of

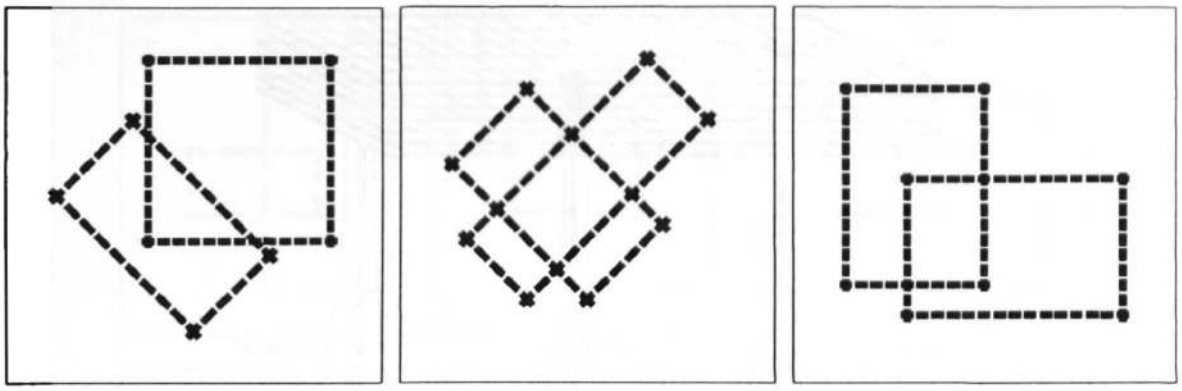

Figure 2: Examples of randomly generated two-dimensional geometric contours

shape features, MAGIC could be directly used to learn grouping principles based on color, texture, etc.

Representing Feature Labelings

Before describing MAGIC, we must first discuss a representation that allows for the labeling of features. Von der Malsburg (1981; von der Malsburg & Schneider, 1986), Gray et al. (1989), Eckhorn et al. (1988), and Strong and Whitehead (1989), among others, have suggested a biologically plausible mechanism of labeling through temporal correlations among neural signals, either the relative timing of neuronal spikes or the synchronization of oscillatory activities in the nervous system. The key idea here is that each processing unit conveys not just an activation value—average firing frequency in neural terms—but also a second, independent value which represents the relative phase of firing. The dynamic grouping or *binding* of a set of features is accomplished by aligning the phases of the features.

In MAGIC, the activity of a feature unit is a complex value with *amplitude* and *phase* components. The phase represents a labeling of the feature, and the amplitude represents the confidence in that labeling. The amplitude ranges from 0 to 1, with 0 indicating a complete lack of confidence and 1 indicating absolute certainty. There is no explicit representation of whether a feature is present or absent in a scene. Rather, absent features are clamped off—their amplitudes are forced to remain at 0—which eliminates their ability to influence other units, as will become clear when the activation dynamics are presented later.

The Architecture

When a scene is presented to MAGIC, units representing features absent in the scene are clamped off and units representing present features are set to a small amplitude and random initial phases. MAGIC's task is to assign appropriate phase values to the units. Thus, the network performs a type of pattern completion.

The network architecture consists of two layers of units, as shown in Figure 3. The lower (input) layer contains the feature units, arranged in spatiotopic arrays with one array per feature type. The upper layer contains hidden units that help to align the phases of the feature units; their response properties are determined by training. There are interlayer connections, but no intralayer connections. Each hidden unit is reciprocally connected to the units in a local spatial region of all feature arrays. We refer to this region as a *patch*; in our current simulations, the patch has dimensions 4×4. For each patch there is a corresponding fixed-size *pool* of hidden units. To achieve uniformity of response across the scene, the pools are arranged in a spatiotopic array in which neighboring pools respond to neighboring patches and the patch-to-pool weights are constrained to be the same at all locations in the array.

The feature units activate the hidden units, which in turn feed back to the feature units. Through a relaxation process, the system settles on an assignment of phases to the features. One might consider an alternative architecture in which feature units were directly connected to one another (Hummel & Biederman, 1992). However, this architecture is in principle not as powerful as the one we propose because it does not allow for higher-order contingencies among features.

Once MAGIC reaches equilibrium, grouped features can be passed on to an object matching system (Figure 1, middle panel). Essentially, this involves considering all phases in a particular range as belonging to a single object. A filter situated between the segmentation system and the matching system permits only features having phases in this range to pass through. The determination of how many objects are present and their range of phases can easily be made using Hough transforms (Ballard, 1981).

Network Dynamics

We summarize here the activation dynamics and learning algorithm. Further justification and intuitions underlying each are presented in Mozer, Zemel, Behrmann, & Williams (1992).

The response of each feature unit i, x_i, is a complex

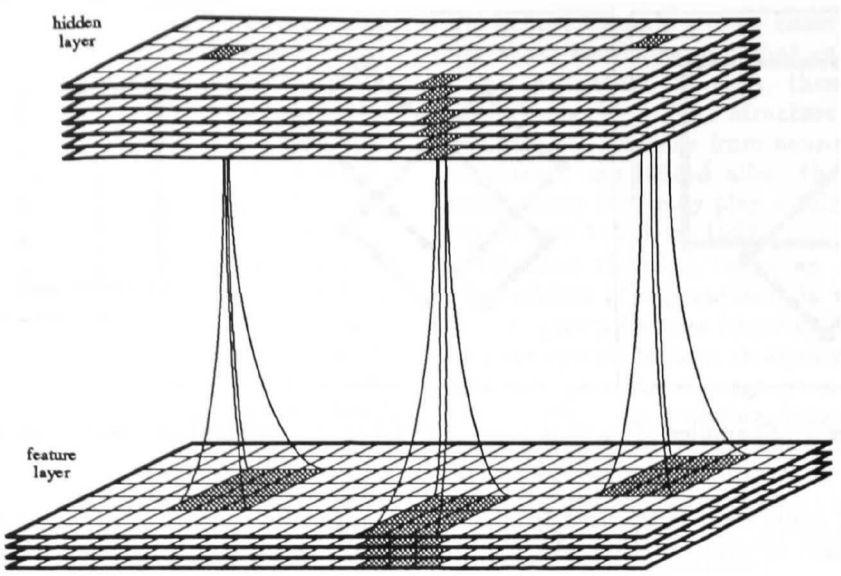

Figure 3: The architecture of MAGIC. The lower (input) layer contains the feature units; the upper layer contains the hidden units. Each layer is arranged in a spatiotopic array with a number of different feature types at each position in the array. Each plane in the feature layer corresponds to a different feature type. The grayed hidden units are reciprocally connected to all features in the corresponding grayed region of the feature layer. The lines between layers represent projections in both directions.

value in polar form, (a_i, p_i), where a_i is the amplitude and p_i is the phase. Similarly, the response of each hidden unit j, y_j, has components (b_j, q_j). The weight connecting unit i to unit j, w_{ji}, is also complex valued, having components (ρ_{ji}, θ_{ji}). The activation rule we propose is a generalization of the dot product to the complex domain. For a particular time step t,

$$net_j(t+1) = \mathbf{x}(t) \cdot \mathbf{w}_j = \sum_i x_i(t) w_{ji}^*$$

where net_j is the net input to hidden unit j and the asterisk denotes the complex conjugate. The net input is passed through a squashing nonlinearity that maps the amplitude of the response from the range $0 \rightarrow \infty$ to $0 \rightarrow 1$ but leaves the phase unaffected. The flow of activation from the hidden layer to the feature layer follows the same dynamics as the flow from the feature layer to the hidden layer. Note that updates are sequential by layer: the feature units activate the hidden units, which then activate the feature units.

In MAGIC, the weight matrix is Hermitian, i.e., $w_{ji} = w_{ij}^*$. This form of weight symmetry ensures that MAGIC will converge to a fixed point (Zemel, Williams, & Mozer, 1992).

Learning Algorithm

During training, we would like the hidden units to learn to detect configurations of features that reliably indicate phase relationships among the features. For instance, if the contours in the scene contain extended horizontal lines, one hidden unit might learn to respond to a collinear arrangement of horizontal segments. Because the unit's response depends on the phase pattern as well as the activity pattern, it will be strongest if the segments all have the same phase value.

The algorithm we have used is a generalization of back propagation. It involves running the network for a fixed number of iterations and, for each iteration, using back propagation to adjust the weights so that the feature phase pattern better matches a target phase pattern. Each training trial proceeds as follows:

1. A training example is generated at random. This involves selecting two contours and instantiating them in a scene. The features of one contour have *target* phase 0° and the features of the other contour have target phase 180°.
2. The training example is presented to MAGIC by setting the initial amplitude of a feature unit to 0.1 if its corresponding scene feature is present, or clamping it at 0.0 otherwise. The phases of the feature units are set to random values in the range 0° to 360°.
3. Activity is allowed to flow from the feature units to the hidden units and back to the feature units.
4. The new phase pattern over the feature units is compared to the target phase pattern (see step 1), and a measure of error is computed. This measure attempts to minimize the difference between the target and actual phases, and to maximize the confidence in the response. The error measure factors out any constant difference between the target and actual phases. See Mozer et al. (1992) for details.
5. Using a generalization of back propagation to complex valued units, error gradients are computed for

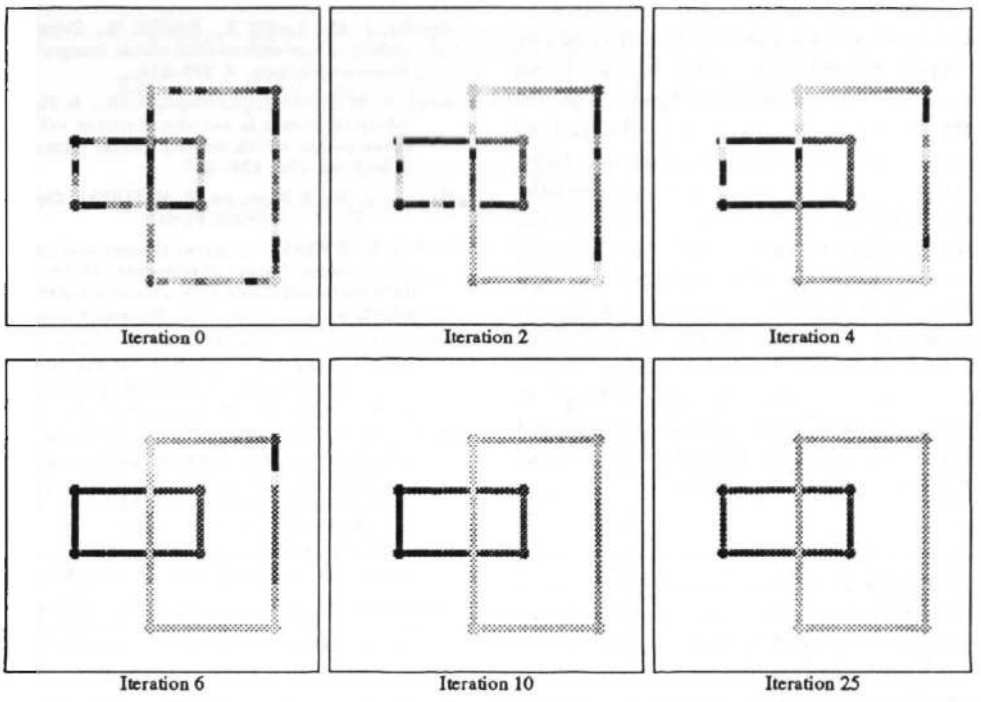

Figure 4: An example of MAGIC segmenting a scene. The "iteration" refers to the number of times activity has flowed from the feature units to the hidden units and back. The phase value of a feature is represented by a gray level. The cyclic phase continuum can only be approximated by a linear gray level continuum, but the basic information is conveyed nonetheless.

the feature-to-hidden and hidden-to-feature weights.

6. Steps 3–5 are repeated for a maximum of 30 iterations. The trial is terminated if the error increases on five consecutive iterations.
7. Weights are updated by an amount proportional to the average error gradient over iterations. Weight constraints are enforced to ensure that $w_{ji} = w^*_{ij}$ and that hidden units of the same "type" responding to different regions of the scene have the same weights.

Simulation Results

We trained a network with 20 hidden units per pool on examples like those shown in Figure 2. Each hidden unit attempts to detect and reinstantiate activity patterns that match its weights. One clear and prevalent pattern in the weights is a collinear arrangement of segments of a given orientation, all having the same phase value. When a hidden unit having weights of this form responds to a patch of the feature array, it tries to align the phases of the patch with the phases of its weight vector. By synchronizing the phases of features, it acts to group the features. Thus, one can interpret the weight vectors as the rules by which features are grouped.

Whereas traditional grouping principles indicate the conditions under which features should be bound together as part of the same object, the grouping principles learned by MAGIC also indicate when features should be segregated into different objects. For example, the weights of the vertical and horizontal segments are generally 180° out of phase with the diagonal segments. This allows MAGIC to segregate the vertical and horizontal features of a rectangle from the diagonal features of a diamond (see Figure 2, left panel). We had anticipated that the weights to each hidden unit would contain two phase values at most because each scene patch contains at most two objects. However, some units make use of three or more phases, suggesting that the hidden unit is performing several distinct functions. As is the usual case with hidden unit weights, these patterns are difficult to interpret.

Figure 4 presents an example of the network segmenting a scene. The scene contains two rectangles. The top left panel shows the features of the rectangles and their initial random phases. The succeeding panels show the network's response during the relaxation process. The lower right panel shows the network response at equilibrium. Features of each object have been assigned a uniform phase, and the two objects are 180° out of phase. The task here may appear simple, but it is quite challenging due to the illusory rectangle generated by the overlapping rectangles.

Empirical Tests of the Model

We are currently conducting psychological experiments to examine the role of feature grouping in human visual

processing. Our experiments include the following:

- Previous studies have shown that judgements of two features of a single object (e.g., size, texture) can be made without loss of accuracy or speed whereas a cost is incurred when the features to be judged are drawn from two different objects (Duncan, 1984; Vecera and Farah, 1992). Based on this rationale, we might expect subjects to identify two elements of a single contour (similar to those used with MAGIC) more rapidly and accurately than elements of disparate contours. This paradigm provides a means of determining whether people group in the same way as MAGIC and what the limitations of grouping are. For instance, we are currently conducting experiments to examine whether a contour is processed as a single entity even when its features are spatially distant and it is partially occluded by a second contour.
- The bottom-up and interactive segmentation models presented in Figure 1 make divergent predictions about the recognition process. In the bottom-up model, segmentation is guided by low-level cues and is not influenced by object knowledge per se. Hence, familiarity should not influence segmentation performance. This is a challenge to test empirically because of the difficulty in measuring segmentation performance directly. The paradigm we are considering involves a search for unfamiliar targets embedded in—and difficult to segment from—a background of distractors. The familiarity of the distractors is manipulated. The interactive model suggests that targets should be easier to identify among familiar distractors. The bottom-up model predicts no effect of distractor familiarity.
- If indeed there is a distinct stage of information processing at which segmentation occurs, then it might be possible to find a neurological patient who has an impairment in segmentation. There are now two such reports in the literature in which patients are unable to bind individual features from disparate locations simultaneously (Grailet et al., 1990; Riddoch and Humphreys, 1987). We are currently studying the feature binding abilities of a visually agnosic subject, CK, and believe that he too has an impairment at this stage of processing.

References

Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition, 13(2)*, 111–122.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*, 501–517.

Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, W., Munk, M., & Reitboek, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics, 60*, 121–130.

Enns, J. T., & Rensink, R. A. (1992). A model for the rapid interpretation of line drawings in early vision. In D. Brogan (Ed.), *Visual search II*. London: Taylor and Francis. In press.

Farah, M. J. (1990). *Visual agnosia*. Cambridge, MA: MIT Press/Bradford Books.

Grailet, J. M., Seron, X., Bruyer, R., Coyette, F., & Frederix, M. (1990). Case report of a visual integrative agnosia. *Cognitive Neuropsychology, 7*, 275–310.

Gray, C. M., Koenig, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit intercolumnar synchronisation which reflects global stimulus properties. *Nature (London), 338*, 334–337.

Hanson, A. R., & Riseman, E. M. (1978). *Computer vision systems*. New York: Academic Press.

Hinton, G. E. (1981). A parallel computation that assigns canonical object-based frames of reference. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence* (pp. 683–685). Los Altos, CA: Morgan Kaufmann.

Hinton, G. E., & Lang, K. (1985). Shape recognition and illusory conjunctions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (pp. 252–259). Los Angeles, California: (null).

Hinton, G. E., Williams, C. K. I., & Revow, M. D. (1992). Adaptive elastic models for hand-printed character recognition. In J. E. Moody, S. J. Hanson, & R. P. Lippman (Eds.), *Advances in neural information processing systems IV*. San Mateo, CA: Morgan Kaufmann.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*. In Press.

Kahneman, D., & Henik, A. (1981). Perceptual organization and attention. In M. Kubovy & J. R. Pomerants (Eds.), *Perceptual organization* (pp. 181–211). Hillsdale, NJ: Erlbaum.

Keeler, J. D., & Rumelhart, D. E. (1992). Self-organising segmentation and recognition neural network. In J. E. Moody, S. J. Hanson, & R. P. Lippman (Eds.), *Advances in neural information processing systems IV*. San Mateo, CA: Morgan Kaufmann.

Lowe, D. G., & Binford, T. O. (1982). Segmentation and aggregation: An approach to figure-ground phenomena. *Proceedings of the DARPA IUS Workshop* (pp. 168–178). Palo Alto, CA.

Löwel, S., & Singer, W. (1992). Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Nature, 255*, 209–211.

Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach*. Cambridge, MA: MIT Press/Bradford Books.

Mozer, M. C., Zemel, R. S., Behrmann, M., & Williams, C. K. I. (1992). Learning to segment images using dynamic feature binding. *Neural Computation*. In Press.

Riddoch, M. J., & Humphreys, G. W. (1987). A case of integrative visual agnosia. *Brain, 110*, 1431–1462.

Rock, I., & Palmer, S. E. (1990). The legacy of Gestalt psychology. *Scientific American, 263*, 84–90.

Rumelhart, D. E. (1992). Script handwritten word recognition in a neural network. Colloquium presented at the Institute of Cognitive Science, University of Colorado.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*, 29–56.

Strong, G. W., & Whitehead, B. A. (1989). A solution to the tag-assignment problem for neural networks. *Behavioral and Brain Sciences, 12*, 381–433.

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

Vecera, S., & Farah, M. J. (1992). Visual attention can select from spatially invariant object representations. Submitted for publication.

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

von der Malsburg, C., & Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics, 54*, 29–40.

Waltz, D. A. (1975). Generating semantic descriptions from drawings of scenes with shadows. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 19–92). New York: McGraw-Hill.

Zemel, R. S., Williams, C. K. I., & Mozer, M. C. (1992). Adaptive networks of directional units. Submitted for publication.