In: Arbib M.A. (ed.), *The Handbook of Brain Theory and Neural Networks*, 2nd Ed., MIT Press, Cambridge MA, pp. 1215-1219, 2003.

Visual Scene Segmentation

DeLiang Wang

Department of Computer and Information Science
and Center for Cognitive Science
The Ohio State University, Columbus, OH 43210-1277, U.S.A.
dwang@cis.ohio-state.edu

INTRODUCTION

A remarkable achievement of the visual system is visual scene analysis, which involves two basic perceptual processes: the segmentation of a visual scene into a set of coherent patterns (objects) and the recognition of memorized ones. In this article, I focus on scene segmentation. Closely related to scene segmentation are figure-ground segregation, which emphasizes segmentation of one object from the rest of the scene (background), and perceptual organization.

Although humans perform scene segmentation with apparent ease, automatic scene segmentation is a very challenging problem. This is so despite dozens of years of intensive research in computer vision and image processing, where image segmentation is the commonly used term.

Objects appear in a natural scene as the grouping of similar sensory features and the segregation of dissimilar ones. Studies in visual perception, in particular Gestalt psychology, have uncovered a number of principles for perceptual organization. I briefly summarize some of the more important principles (see Chapter 6 of Palmer, 1999):

• *Proximity*. Features nearby in space and time tend to group into the same segment.

• *Similarity*. Features that have similar attributes, such as brightness, tend to group.

• *Connnectedness*. A uniform, connected region, for instance a blackboard, tends to form a single segment.

• *Memory*. Features that belong to the same memory pattern tend to group.

Compared to recognition, far fewer papers address scene segmentation in neural networks. A main reason is that the problem is particularly challenging for neural networks because they also need to address the binding problem. The binding problem refers to how the coherence of a pattern, generally as a large collection of features, is represented in a neural network. One proposal is the grandmother-cell representation, which claims that individual neurons can become so selective as to represent individual objects. Multiple objects in a visual scene would be represented by the coactivation of multiple cells. Another proposal, known as temporal correlation (von der Malsburg, 1981), encodes the binding by the correlation of temporal activities of feature-detecting cells. A special form of temporal correlation is *oscillatory correlation*, where basic units are neural oscillators (Terman and Wang, 1995). Oscillatory correlation is the underlying representation for a number of recent studies that have substantially advanced the capability of neural networks in scene segmentation.

I first review non-oscillatory approaches, and then turn to oscillatory approaches. Concluding remarks are given in the Discussion section.

NON-OSCILLATORY APPROACHES

Boltzmann Machine

In an early study, Sejnowski and Hinton (1987) introduced the use of a Boltzmann machine (see SIMULATED ANNEALING AND BOLTZMANN MACHINES) for figure-ground segregation. Their network consists of two types of binary units: figure units and edge units. Units in the network are locally and symmetrically connected with fixed excitatory and inhibitory weights. Such connections reflect local cooperation within a connected region and local competition between figure and background. There are two kinds of input to the network: bottom-up input that contains the location and orientation of edges (or line segments) and top-down input that corresponds to visual attention and provides a necessary bias for selecting a figure in an image. The desired output of the system is that the units corresponding to a figural object and its boundary are active while the rest of the network units are silent. Their simulation results on small synthetic images produce desired results. As pointed out by the authors, it is uncertain whether the Boltzmann machine approach is applicable to segmenting real images.

FBF Model

Grossberg and Wyse (1991) proposed the so-called FBF model for scene segmentation. The FBF model iterates between two subsystems: a feature contour system and a boundary contour system in the order of feature-boundary-feature, hence the acronym. The feature contour system detects local features using on-center/off-surround and off-center/on-surround filters and then performs diffusion within an image region, while the boundary contour system detects local edges and performs contour completion. Subsequently, a filling-in process spreads a region label until there is a boundary signal. The model has been tested using simple images, and its performance on real images is unclear given the recognized difficulty of contour completion in real images. Labeling by filling-in is rather cumbersome: too few labels may miss significant regions and too many generate duplicate segments. This labeling process, however, can be significantly improved by introducing top-down attention for selecting one region, as described by Sejnowski and Hinton (1987), and shift of attention for selecting multiple regions sequentially.

Classification-based Approach

Neural networks are well-established as pattern classifiers. Since scene segmentation, in some sense, may be viewed as a classification problem, neural networks have been used in many studies to do image segmentation as classification. A training stage precedes actual classification, and in the training stage multiple classes are formed, corresponding to multiple labels for the regions of interest. After training, the classifier is used to label each individual pixel in the image. Koh et al. (1995) proposed a hierarchical Kohonen map (see SELF-ORGANIZING MAPS; KOHONEN MAPS) for range image segmentation. At each level of the hierarchy, a Kohonon map is used to segment an image into a given number of regions. A hierarchy is used for two purposes. First, the requirement to know the number of segments *a priori* is alleviated. Second, the hierarchy embodies multiple scales of the input feature space. More recently, Alirezaie et al. (1997) used both the Kohonen map and a multilayer perceptron for segmenting two-dimensional (2-D) MRI (magnetic resonance imaging) images of the human brain. Both networks are trained to classify three kinds of tissue: white matter, gray matter, and cerebrospinal fluid. After training, the two networks are used to segment the same images. They report good performance for Kohonen maps and worse performance for multilayer perceptrons.

A fundamental limitation of all classification-based approaches to segmentation, neural networks or not, is that classification is based on local information only, whereas proper segmentation of a location depends on the image context of that location. This limitation is illustrated in Fig. 1, where twelve line segments are arranged in two different ways in Fig. 1a and Fig. 1b. Because the same set of line segments occurs in both images, local classification

produces the same segmentation result, while the two images are perceptually organized very differently. Of course one can train a system to classify a pixel together with its neighborhood. But fixed neighborhoods cannot capture the variability of image contexts.

OSCILLATORY APPROACHES

Early Simulations

Theoretical considerations and the discovery of coherent oscillations in the visual cortex in late eighties (see SYNCHRONIZATION OF NEURONAL RESPONSES AS A PUTATIVE BINDING MECHANISM) have triggered much interest in exploring oscillatory correlation to address scene segmentation and figure-ground segregation. Most of the early models employ harmonic oscillators and all-to-all connections to reach synchronization. These models are fundamentally limited in addressing the scene segmentation problem, because critical information about the topology of sensory features is lost.

Recognizing the limitation of all-to-all connectivity, Sporns et al. (1991) constructed a locally connected network for modeling perceptual organization. To achieve proper synchronization they use reentrant connections and dynamic weights that quickly adapt to presynaptic and postsynaptic stimulation. Schillen and König (1994) proposed a network that performs synchronization and desynchronization in multiple feature domains. Their network uses Wilson-Cowan oscillators, which model oscillations from an interacting population of excitatory and inhibitory neurons, and time delays between elements of neighboring oscillators to achieve both synchronization and desynchronization. To deal with multiple feature domains, the network is extended to include multiple modules and cross-module connections. Both of these studies have been tested using only synthetic stimuli.

LEGION Networks

Terman and Wang (1995) proposed and analyzed a class of locally excitatory globally inhibitory oscillator networks, called LEGION. Each oscillator i in a LEGION network is a relaxation oscillator:

$$\dot{x}_i = f(x_i) - y_i + I_i + S_i + \rho \tag{1a}$$

$$\dot{y}_i = \varepsilon(g(x_i) - y_i) \tag{1b}$$

Here $f(x) = 3x - x^3 + 2$ is a cubic, $g(x) = \alpha[1 + \tanh(x/\beta)]$ is a sigmoid (α and β are parameters). I_i denotes external input, and ρ intrinsic noise. The parameter ε is a small positive number, which yields two time scales that are the defining property of relaxation oscillations. When $I_i > 0$, (1) gives rise to a stable limit cycle, which alternates between a silent phase (small x values) and an active phrase (large x values). Due to ε, the oscillator activity changes slowly within either of the two phases but the alternation between the two phases takes place rapidly, referred to as jumping. S_i denotes the overall input from the network, and it contains a local excitatory term and a global inhibitory term. The excitatory term represents the excitatory input from a set of adjacent oscillators that connect to i. In a 2-D LEGION network, the set in the simplest case contains four immediate neighbors. This architecture is shown in Fig. 2a. The inhibitory term specifies the inhibition from a global inhibitor (see Fig. 2a).

Terman and Wang conducted an extensive analysis on LEGION networks, based on an earlier analysis by Somers and Kopell (1993) on two coupled relaxation oscillators. They showed that LEGION exhibits the mechanism of *selective gating* as follows. When an oscillator jumps to the active phase, its activity spreads to its neighboring oscillators, which further spread the activity to

their neighbors, and so on. This leads to synchronization in LEGION. In addition, oscillating groups inhibit each other through global inhibition so that at most one group can be in the active phase at a time. This leads to desynchronization. They proved the following theorem: LEGION networks achieve both synchronization and desynchronization in no greater than *m* cycles of oscillations, where *m* is the number of patterns in an input scene, so long as *m* does not exceed the segmentation capacity. The segmentation capacity refers to the maximum number of patterns that can be segmented by LEGION, corresponding to the ratio of the oscillation period to the duration of the active phase. The capacity is about 5 to 7 for typical parameter values.

The following simulation illustrates the selective gating mechanism. An input image with three caricature patterns, a rabbit, a duck and a flower, is simultaneously presented to a 30x30 LEGION network, as shown in Fig. 2b. The oscillators under stimulation become oscillatory, while those without stimulation cannot oscillate. Fig. 2c shows the temporal evolution of every stimulated oscillator. The activities of the oscillators representing each object are combined together in Fig. 2c. Although the oscillators in the network start with random phases, the synchronization within each pattern and the desynchronization between different patterns are clearly attained in just two oscillation periods.

Image Segmentation Using LEGION

Wang and Terman (1997) extended LEGION to distinguish between major image regions and noisy fragments; the latter are collected into the background. For gray-level images, each oscillator corresponds to one pixel, and two neighboring oscillators are connected with a weight proportional to pixel similarity. To speed up simulation with a large number of oscillators needed for processing real images, Wang and Terman also abstracted an algorithm that follows LEGION dynamics. To illustrate typical segmentation results, Fig. 3a shows a gray-level aerial image. Fig. 3b shows the result of segmentation by the algorithm. The image is segmented into 23 regions, each of which corresponds to a different intensity level in Fig. 3b, which indicates a distinct phase of oscillators. Note that the segmentation capacity is removed in the algorithm for computational efficiency. In the simulation, different segments pop out from the network sequentially, as similarly shown in Fig. 2c. As displayed in Fig. 3b, almost all major regions are successfully segmented. The black scattered areas in the figure represent the background.

A variety of real imagery has been successfully segmented by LEGION networks and their variants, including intensity images such as medical (Shareef et al., 1999) and satellite images, texture images, and image sequences (motion). Oscillatory correlation provides a unique way to scene labeling. As illustrated in Fig. 2c, segmentation is performed in *time*; each segment pops out at a distinct time from the network and different segments alternate in time. Once a segment is in the active phase, all of its features, but none of the ones from competing segments, are simultaneously available for later visual processing such as attention and recognition (see Wang and Liu, 2002).

Contour Extraction

Yen and Finkel (1998) used laterally coupled phase oscillators to extract salient contours. Excitatory and inhibitory connections in their network encode orientation relations between edge filters. The saliency of a contour is embodied by the total activity of a synchronized oscillator group that corresponds to the contour. Using integrate-and-fire oscillators, Horn and Opher (1999) studied the detection of borders between regions. Their network uses difference-of-gaussian coupling, and their simulations suggest that at local minima of total network activity, firing oscillators tend to correspond to edges that separate different regions.

DISCUSSION

The field of neural networks has seen major advances in visual scene segmentation in recent years. The temporal correlation hypothesis is a biologically plausible representation to deal with the binding problem. The selective gating mechanism provides a computational foundation for oscillatory correlation. These advances have finally enabled neural networks to analyze real scenes. I conclude with a brief discussion of two issues for future research.

1. Multi-cue interaction. There are many grouping cues responsible for perceptual organization. It is important to build systems that synergistically integrate multiple cues, not simply breaking them to independent modules. Is there, or should there be, a common segmentation mechanism at a deeper level?

2. Top-down analysis. Studies in human visual psychophysics demonstrate strong top-down influence on scene analysis. Sources of top-down information include recognition, goal and expectation, short-term memory, and attention. Few studies have seriously addressed these issues in the context of analyzing real scenes. Their importance deserves far more attention.

REFERENCES

Alirezaie, J., Jernigan, M. E., and Nahmias, C., 1997, Neural network-based segmentation of magnetic resonance images of the brain, IEEE Trans. Nuclear Sci., 44: 194-198.

Grossberg, S., and Wyse, L., 1991, A neural network architecture for figure-ground separation of connected scenic figures, Neural Net., 4: 723-742.

*Horn, D., and Opher, I., 1999, Collective excitation phenomena and their applications, in Pulsed Neural Networks, (W. Maass and C. M. Bishop, Eds.), Cambridge MA: MIT Press, pp. 297-320.

Koh, J., Suk, M., and Bhandarkar, S. M., 1995, A multilayer self-organizing feature map for range image segmentation, Neural Net., 8: 67-86.

*Palmer, S. E., 1999, Visual Science, Cambridge MA: MIT Press.

Schillen, T. B., and König, P., 1994, Binding by temporal structure in multiple feature domains of an oscillatory neuronal network, Biol. Cybern., 70: 397-405.

Sejnowski, T. J., and Hinton, G. E., 1987, Separating figure from ground with a Boltzmann machine, in Vision, Brain, and Cooperative Computation, (M. A. Arbib and A. R. Hanson, Eds.), Cambridge MA: MIT Press, pp. 703-724.

Shareef, N., Wang, D. L., and Yagel, R., 1999, Segmentation of medical images using LEGION, IEEE Trans. Med. Imaging, 18: 74-91.

Somers, D., and Kopell, N., 1993, Rapid synchrony through fast threshold modulation, Biol. Cybern., 68: 393-407.

Sporns, O., Tononi, G., and Edelman, G. M., 1991, Modeling perceptual grouping and figure-ground segregation by means of active re-entrant connections, Proc. Natl. Acad. Sci. USA, 88: 129-133.

Terman, D., and Wang, D. L., 1995, Global competition and local cooperation in a network of neural oscillators, Physica D, 81: 148-176.

von der Malsburg, C., 1981, The correlation theory of brain function, Internal Report 81-2, Max-Planck-Institute for Biophysical Chemistry.

Wang, D. L., and Liu, X., 2002, Scene analysis by integrating primitive segmentation and associative memory, IEEE Trans. Syst. Man Cybern. - Part B: Cybern., to appear.

Wang, D. L., and Terman, D., 1997, Image segmentation based on oscillatory correlation, Neural Comp., 9: 805-836 (for errata see Neural Comp., vol. 9, pp. 1623-1626, 1997).

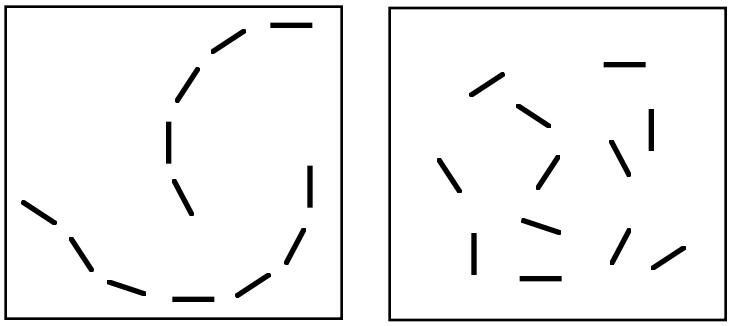
Yen, S.-C., and L. H. Finkel, 1998. Extraction of perceptually salient contours by striate cortical networks. Vis. Res., 38: 719-741.

FIGURE CAPTIONS

Figure 1. Perceptual organization of line segments. **a**. A spatial arrangement of 12 line segments. **b**. A different arrangement.

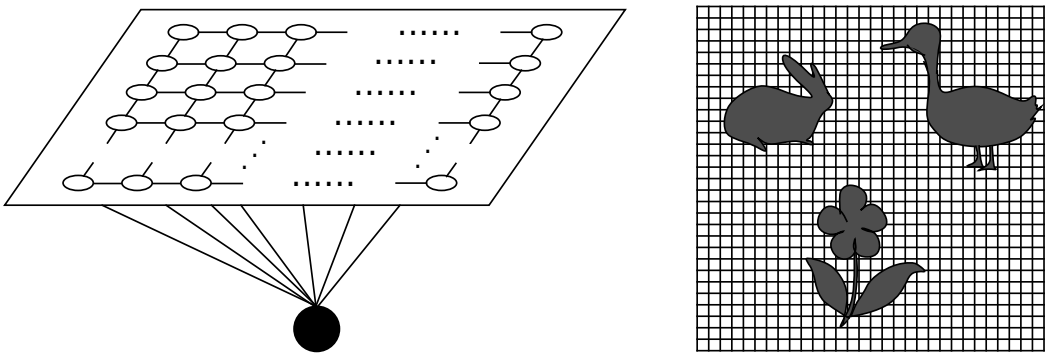
Figure 2. LEGION network. **a**. Architecture of 2-D LEGION, where white circles indicate oscillators and the black circle indicates the global inhibitor. **b**. An input image as sampled by a 30x30 LEGION network. **c**. Temporal activity of the network (adapted from D.L. Wang, *Cognitive Science*, vol. 20, p. 425, 1996), where the upper three traces show the combined temporal activities of the oscillator groups corresponding to the indicated patterns. The bottom trace shows the activity of the global inhibitor.

Figure 3. Image segmentation (from Wang and Terman, 1997). **a**. A gray-level image with 160x160 pixels. **b**. Segmentation result for **a**, where each segment is indicated by a distinct gray level.

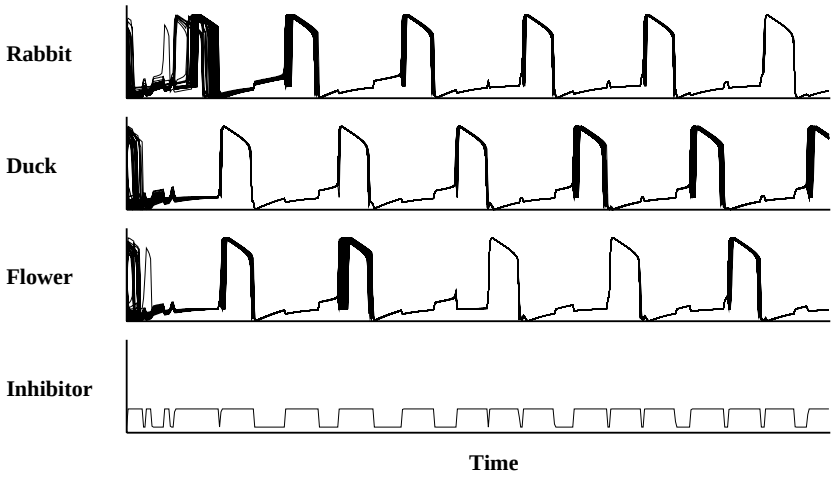


a **b**

Figure 1

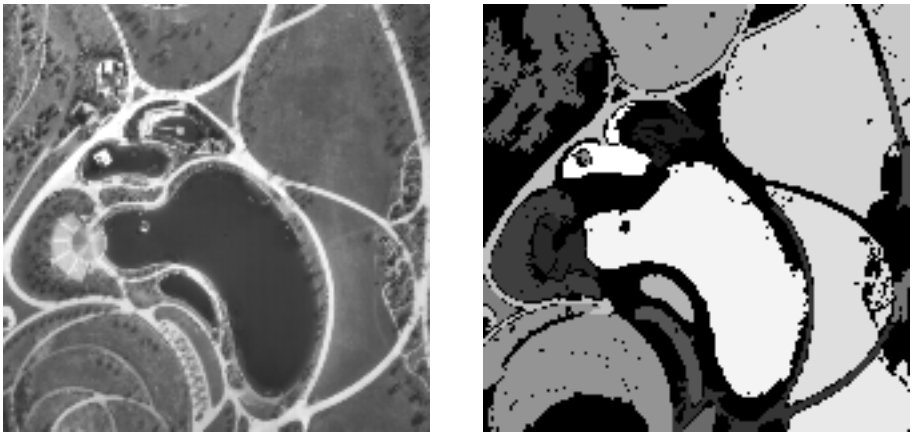


a b



c

Figure 2



a **b**

Figure 3