Computational Scene Analysis

DeLiang Wang

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

Summary. A remarkable achievement of the perceptual system is its scene analysis capability, which involves two basic perceptual processes: the segmentation of a scene into a set of coherent patterns (objects) and the recognition of memorized ones. Although the perceptual system performs scene analysis with apparent ease, computational scene analysis remains a tremendous challenge as foreseen by Frank Rosenblatt. This chapter discusses scene analysis in the field of computational intelligence, particularly visual and auditory scene analysis. The chapter first addresses the question of the goal of computational scene analysis. A main reason why scene analysis is difficult in computational intelligence is the binding problem, which refers to how a collection of features comprising an object in a scene is represented in a neural network. In this context, temporal correlation theory is introduced as a biologically plausible representation for addressing the binding problem. The LEGION network lays a computational foundation for oscillatory correlation, which is a special form of temporal correlation. Recent results on visual and auditory scene analysis are described in the oscillatory correlation framework, with emphasis on real-world scenes. Also discussed are the issues of attention, feature-based versus model-based analysis, and representation versus learning. Finally, the chapter points out that the time dimension and David Marr's framework for understanding perception are essential for computational scene analysis.

1 Introduction

Human intelligence can be broadly divided into three aspects: Perception, reasoning, and action. The first is mainly concerned with analyzing the information in the environment gathered by the five senses, and the last is primarily concerned with acting on the environment. In other words, perception and action are about input and output, respectively, from the viewpoint of the intelligent agent (i.e. a human being). Reasoning involves higher cognitive functions such as memory, planning, language understanding, and decision

making, and is at the core of traditional artificial intelligence [49]. Reasoning also serves to connect perception and action, and the three aspects interact with one another to form the whole of intelligence.

This chapter is about perception - we are concerned with how to analyze the perceptual input, particularly in the visual and auditory domains. Because perception seeks to describe the physical world, or scenes with objects located in physical space, perceptual analysis is also known as scene analysis. To differentiate scene analysis by humans and by machines, we term the latter *computational scene analysis*[1]. In this chapter I focus on the analysis of a scene into its constituent objects and their spatial positions, not the recognition of memorized objects. Pattern recognition has been much studied in computational intelligence, and is treated extensively elsewhere in this collection.

Although humans, and nonhuman animals, perform scene analysis with apparent ease, computational scene analysis remains an extremely challenging problem despite decades of research in fields such as computer vision and speech processing. The difficulty was recognized by Frank Rosenblatt in his 1962 classic book, "Principles of neurodynamics" [47]. In the last chapter, he summarized a list of challenges facing perceptrons at the time, and two problems in the list "represent the most baffling impediments to the advance of perceptron theory" (p. 580). The two problems are figure-ground separation and the recognition of topological relations. The field of neural networks has since made great strides, particularly in understanding supervised learning procedures for training multilayer and recurrent networks [2, 48]. However, progress has been slow in addressing Rosenblatt's two chief problems, largely validating his foresight.

Rosenblatt's first problem concerns how to separate a figure from its background in a scene, and is closely related to the problem of scene segregation: To decompose a scene into its comprising objects. The second problem concerns how to compute spatial relations between objects in a scene. Since the second problem presupposes a solution to the first, figure-ground separation is a more fundamental issue. Both are central problems of computational scene analysis.

In the next section I discuss the goal of computational scene analysis. Section 3 is devoted to a key problem in scene analysis - the binding problem, which concerns how sensory elements are organized into percepts in the brain. Section 4 describes oscillatory correlation theory as a biologically plausible representation to address the binding problem. The section also reviews the LEGION[2] network that achieves rapid synchronization and desynchronization, hence providing a computational foundation for the oscillatory correlation theory. The following two sections describe visual and auditory scene

---

[1] This is consistent with the use of the term Computational Intelligence.

[2] LEGION stands for Locally Excitatory Globally Inhibitory Oscillator Network [68].

analysis separately. In Section 7, I discuss a number of challenging issues facing computational scene analysis. Finally, Section 8 concludes the chapter.

Note that this chapter does not attempt to survey the large body of literature on computational scene analysis. Rather, it highlights a few topics that I consider to be most relevant to this book.

2 What is the Goal of Computational Scene Analysis?

In his monumental book on computational vision, Marr makes a compelling case that understanding perceptual information processing requires three different levels of description. The first level of description, called computational theory, is mainly concerned with the goal of computation. The second level, called representation and algorithm, is concerned with the representation of the input and the output, and the algorithm that transforms from the input representation to the output representation. The third level, called hardware implementation, is concerned with how to physically realize the representation and the algorithm.

So, what is the goal of computational scene analysis? Before addressing this question, let us ask the question of what purpose perception serves. Answers to this question have been attempted by philosophers and psychologists for ages. From the information processing perspective, Gibson [21] considers perception as the way of seeking and gathering information about the environment from the sensory input. On visual perception, Marr [30] considers that its purpose is to produce a visual description of the environment for the viewer. On auditory scene analysis, Bregman states that its goal is to produce separate streams from the auditory input, where each stream represents a sound source in the acoustic environment [6]. It is worth emphasizing that the above views suggest that perception is a private process of the perceiver even though the physical environment may be common to different perceivers.

In this context, we may state that *the goal of computational scene analysis is to produce a computational description of the objects and their spatial locations in a physical scene from sensory input.* The term 'object' here is used in a modality-neutral way: An object may refer to an image, a sound, a smell, and so on. In the visual domain, sensory input comprises two retinal images, and in the auditory domain it comprises two eardrum vibrations. Thus, the goal of visual scene analysis is to extract visual objects and their locations from one or two images. Likewise, the goal of auditory scene analysis is to extract streams from one or two audio recordings.

The above goal of computational scene analysis is strongly related to the goal of human scene analysis. In particular, we assume the input format to be similar in both cases. This assumption makes the problem well defined and has an important consequence: It makes the research in computational scene analysis perceptually relevant. In other words, progress in computational scene analysis may shed light on perceptual and neural mechanisms.

This restricted scope also differentiates computational scene analysis from engineering problem solving, where a variety and a number of sensors may be used.

With common sensory input, we further propose that computational scene analysis should aim to achieve human level performance. Moreover, we do not consider the problem solved until a machine system achieves human level performance in *all* perceptual environments. That is, computational scene analysis should aim for the versatile functions of human perception, rather than its utilities in restricted domains.

3 Binding Problem and Temporal Correlation Theory

The ability to group sensory elements of a scene into coherent objects, often known as perceptual organization or perceptual grouping [40], is a fundamental part of perception. Perceptual organization takes place so rapidly and effortlessly that it is often taken for granted by us the perceivers. The difficulty of this task was not fully appreciated until effort in computational scene analysis started in earnest. How perceptual organization is achieved in the brain remains a mystery.

Early processing in the perceptual system clearly involves detection of local features, such as color, orientation, and motion in the visual system, and frequency and onset in the auditory system. Hence, a closely related question to perceptual organization is how the responses of feature-detecting neurons are bound together in the brain to form a perceived scene? This is the well-known *binding problem*. At the core of the binding problem is that sensory input contains multiple objects simultaneously and, as a result, the issue of which features should bind with which others must be resolved in objection formation. I illustrate the situation with two objects - a triangle and a square - at two different locations: The triangle is at the top and the square is at the bottom. This layout, shown in Figure 1, was discussed by Rosenblatt [47] and used as an instance of the binding problem by von der Malsburg [60]. Given feature detectors that respond to triangle, square, top, and bottom, how can the nervous system bind the locations and the shapes so as to perceive that the triangle is at the top and the square is at the bottom (correctly), rather than the square is on top and the triangle is on bottom (incorrectly)? We should note that object-level attributes, such as shape and size, are undefined before the more fundamental problem of figure-ground separation is solved. Hence, I will refer to the binding of local features to form a perceived object, or a percept, when discussing the binding problem.

How does the brain solve the binding problem? Concerned with shape recognition in the context of multiple objects, Milner [32] suggested that different objects could be separated in time, leading to synchronization of firing activity within the neurons activated by the same object. Later von der Malsburg [59] proposed a correlation theory to address the binding problem. The

Input   Feature Detectors
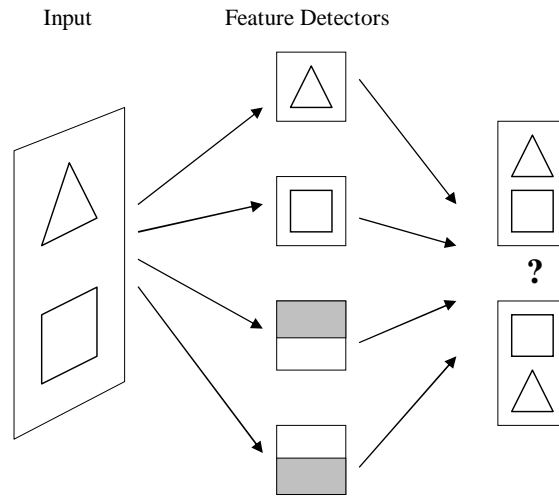
Fig. 1. Illustration of the binding problem. The input consists of a triangle and a square. There are four feature detectors for triangle, square, top, and bottom. The binding problem concerns whether the triangle is on top (and the square at bottom) or the square is on top (and the triangle at bottom).

correlation theory asserts that the temporal structure of a neuronal signal provides the neural basis for correlation, which in turn serves to bind neuronal responses. In a subsequent paper, von der Malsburg and Schneider [61] demonstrated the temporal correlation theory in a neural model for segregating two auditory stimuli based on their distinct onset times - an example of auditory scene analysis that I will come back to in Section 6. This paper proposed, for the first time, to use neural oscillators to solve a figure-ground separation task, whereby correlation is realized by synchrony and desynchrony among neural oscillations. Note that the temporal correlation theory is a theory of representation, concerned with how different objects are represented in a neural network, not a computational algorithm; that is, the theory does not address how multiple objects in the input scene are transformed into multiple cell assemblies with different time structures. This is a key computational issue I will address in the next section.

   The main alternative to the temporal correlation theory is the hierarchical coding hypothesis, which asserts that binding occurs through individual neurons that are arranged in some cortical hierarchy so that neurons higher in the hierarchy respond to larger and more specialized parts of an object. Eventually, individual objects are coded by individual neurons, and for this reason hierarchical coding is also known as the cardinal cell (or grandmother cell) representation [3]. Gray [23] presented biological evidence for and against the hierarchical representation. From the computational standpoint, the hier-

archical coding hypothesis has major drawbacks, including the need to encode a prohibitively large number of scenarios by cells [59, 65].

It should be clear from the above discussion that the figure-ground separation problem is essentially the same as the binding problem. A layered perceptron may be viewed as a computational implementation of the hierarchical coding hypothesis, and the problems challenging Rosenblatt's perceptrons underline the limitations of hierarchical coding.

4 Oscillatory Correlation Theory

A special form of temporal correlation - oscillatory correlation [52] - has been studied extensively. In oscillatory correlation, feature detectors are represented by oscillators and binding is represented by synchrony within an assembly of oscillators and desynchrony between different assemblies. The notion of oscillatory correlation is directly supported by the substantial evidence of coherent oscillations in the brain. In addition, the activity of a neuron or a local group of neurons can be accurately modeled by an oscillator. It is worth pointing out here that a mathematical oscillator need not always produce periodic behavior; indeed an oscillator in response to a time varying input often exhibits a variety of aperiodic responses.

Like the temporal correlation theory, the oscillatory correlation theory is a representation theory, not a computational mechanism. A computational mechanism for the oscillatory correlation theory needs to exhibit three key features [65]. First, the mechanism can synchronize a locally coupled assembly of oscillators. Second, it can desynchronize different assemblies of oscillators that are activated by multiple, simultaneously present objects. Third, both synchrony and desynchrony must occur rapidly in order to deal with the changing environment.

The first neural network that successfully met the above requirements is the LEGION mechanism proposed in 1995 by Terman and Wang [52, 62]. LEGION builds on relaxation oscillators characterized by two time scales [58]. Formally, a relaxation oscillator, i, is defined as a pair of an excitatory unit x_i and an inhibitory unit y_i [52]:

$$\dot{x}_i = 3x_i - x_i^3 + 2 - y_i + I_i + S_i + \rho \tag{1a}$$

$$\dot{y}_i = \varepsilon(\alpha(1 + \tanh(x_i/\beta)) - y_i) \tag{1b}$$

In the above equation, I_i denotes the external stimulation to the oscillator and S_i the input from the rest of the network, to be specified below. ρ denotes the amplitude of intrinsic noise (e.g. Gaussian noise) which assists the process of desynchronization, and α and β are parameters. ε is a small positive parameter, and it is this parameter that induces the two time scales with y on a slower one.

Figure 2 illustrates the oscillator defined in (1). As shown in Fig. 2A, the x-nullcline (i.e. $\dot{x} = 0$) is a cubic function and the y-nullcline is a sigmoid
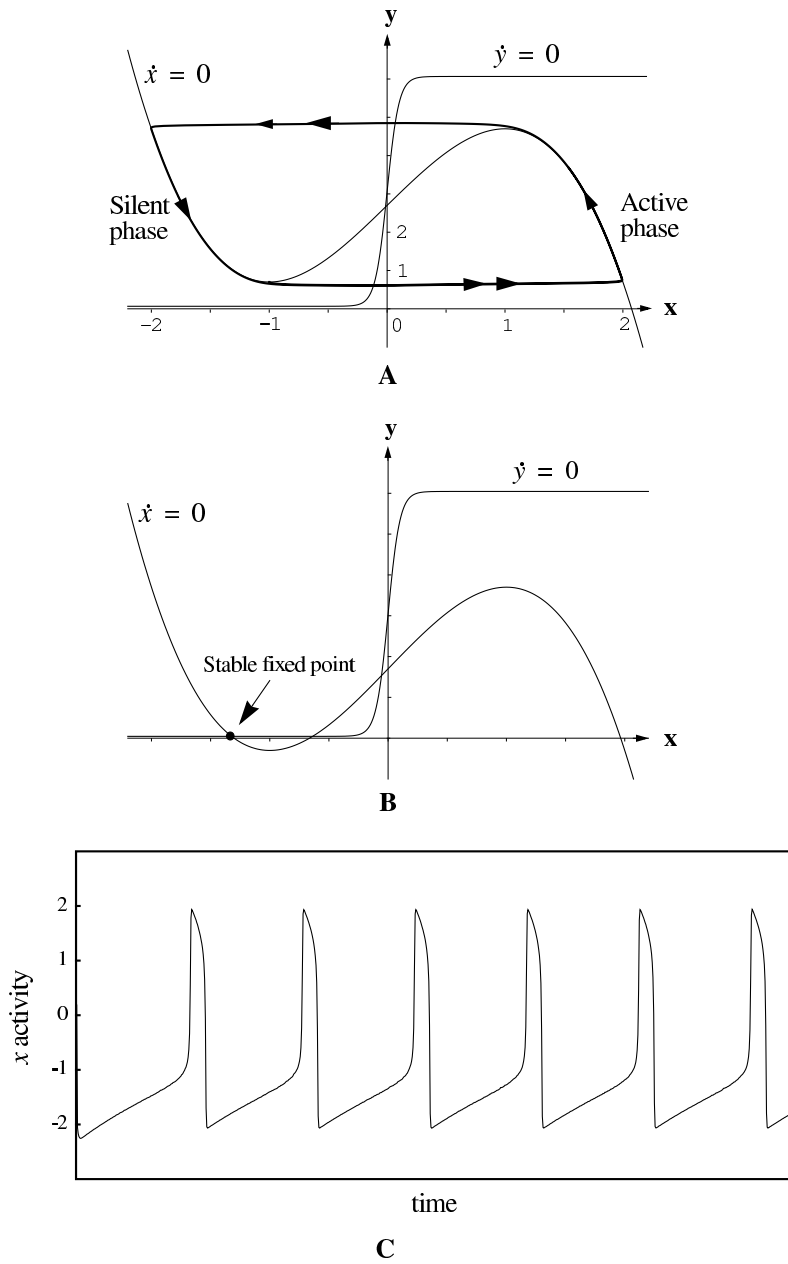
Fig. 2. Behavior of a relaxation oscillator. A. Enabled state of the oscillator. This state produces a limit cycle shown as the bold curve. The direction of the trajectory is indicated by the arrows, and jumps are indicated by double arrows. B. Excitable state of the oscillator. This state produces a stable fixed point. C. Temporal activity of the oscillator in the enabled state. The curve shows the x activity.

function. When $I > 0$, the two nullclines intersect at a single point on the middle branch of the cubic, and the oscillator produces a stable limit cycle shown in Figure 2A. In this case, the oscillator is referred to as *enabled*, and the limit cycle alternates between an *active phase* with relatively high x values and a *silent phase* with relatively low x values. Within each of the two phases the oscillator exhibits slow-varying behavior. However, the transition between the two phases occurs rapidly, called *jumping*. The role of α is to determine the relative times the oscillator spends in the two phases - a larger α produces a relatively shorter active phase. The situation when $I < 0$ is shown in Fig. 2B. In this case, the two nullclines intersect at a stable fixed point on the left branch of the cubic, and no oscillation occurs - the oscillator is referred to as *excitable*. Whether the state of an oscillator is enabled or excitable depends solely on external stimulation; in other words, oscillation is stimulus dependent. The x activity of an enabled state is given in Fig. 2C, and it resembles a spike train. Indeed, relaxation oscillators have been widely used as models of single neurons, where x is interpreted as the membrane potential of a neuron and y the activation state of ion channels [19, 35, 36]. A relaxation oscillation may also be interpreted as an oscillating burst of neuronal spikes, where x corresponds to the envelope of the burst.

In a LEGION network an oscillator is excitatorily coupled with other oscillators in its neighborhood, and excites a global inhibitor which then inhibits every oscillator in the network. Specifically, S_i in (1a) is defined as

$$S_i = \sum_{k \in N(i)} W_{ik} H(x_k - \theta_x) - W_z H(z - \theta_z) \qquad (2)$$

where $N(i)$ denotes a set of neighbors of i, and W_{ik} the connection weight from oscillator k to i. H stands for the Heaviside step function, and θ_x and θ_z are thresholds. W_z is the weight of inhibition from the global inhibitor z, which is defined as

$$\dot{z} = \phi(\sigma_\infty - z) \qquad (3)$$

where ϕ is a parameter. $\sigma_\infty = 1$ if at least one oscillator is in the active phase and $\sigma_\infty = 0$ otherwise. From (3) it is easy to see that $z \to 1$ when σ_∞ equals 1.

On the basis of the earlier analysis by Somers and Kopell [51] on two coupled relaxation oscillators, Terman and Wang [52] conducted an extensive analysis on LEGION networks. 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, their neighbors, and so on. This leads to fast synchronization within the oscillator assembly that contains the oscillator. In addition, the oscillator activates the global inhibitor which prevents the oscillators of different assemblies from jumping up. This leads to desynchronization among different oscillator assemblies. They proved the following theorem: A LEGION network can reach both synchronization within each assembly and desynchronization between different assemblies, and
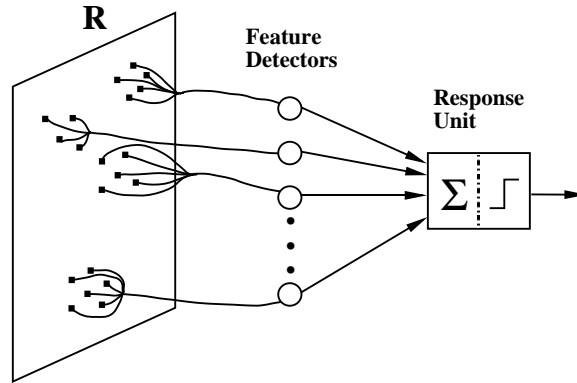
Fig. 3. Diagram of a perceptron. R denotes the input layer, which projects to a layer of feature detectors. The response unit takes a weighted sum of the responses of all the detectors, and outputs 1 if the sum passes a certain threshold and 0 otherwise.

does so in no greater than m cycles of oscillations, where m is the number of the oscillator assemblies. In other words, both synchronization and desynchronization are achieved rapidly.

The selective gating mechanism of LEGION successfully meets the three computational requirements stated at the beginning of this section. Subsequent research has shown that rapid synchronization and desynchronization can also be achieved using other types of oscillators, such as Wilson-Cowan and spike (integrate-and-fire) oscillators, although conclusions are typically drawn from numerical simulations. See [65] for a broad review on this topic.

As a concrete application of LEGION dynamics, I describe a solution to a classic problem in neural computation - the connectedness problem [64]. The connectedness problem, first described by Minsky and Papert in 1969, is the centerpiece of their consequential critique on perceptrons [33]. The connectedness predicate is innocuously simple: To classify whether an input pattern is connected or not. To appreciate the significance of this predicate, I need to give some context on perceptron theory. Rosenblatt's perceptrons [46, 47] are classification networks. A typical perceptron, illustrated in Figure 3, computes a predicate. It consists of a binary input layer R, which symbolizes retina, a layer of binary feature detectors, and a response unit that signals the result of a binary classification. A feature detector is activated if and only if all the pixels within the area of R sensed by the detector are black. The response unit outputs 1 if a weighted sum of all the feature detectors exceeds a threshold, and outputs 0 otherwise.

Minsky and Papert [33] define the *order* of a predicate as the smallest number of pixels in R that must be sensed by some feature detector in order to compute the predicate. With this notion, they prove that the order of the connectedness predicate increases at least as fast as $|R|^{1/2}$. That is, the order of this predicate is unbounded. What does this result mean? It means that,

to compute a predicate of an unbounded order requires feature detectors with too large receptive fields (relative to R) and too many of detectors to be computationally feasible [33]. It is important to understand that the result is not about *computability*, or *whether* a perceptron exists to solve the problem. With a finite size of R, the number of connected patterns is finite, and we can simply find a perceptron to solve the problem, in which each connected pattern is sensed by a single feature detector. However, the number of connected patterns grows exponentially except for one-dimensional R [65], and this way of computing the connectedness predicate is computationally intractable. Hence, their result is about the *scalability* or *computational complexity*.

Thanks to recurrent connectivity and oscillatory dynamics, LEGION solves the connectedness problem in general form [64]. To explain the solution, Figure 4 shows the response of a two-dimensional (2-D) LEGION network to two binary images: The first one is a connected figure showing a cup (Fig. 4A) and the second one is a disconnected figure showing the word 'CUP' (Fig. 4D). To ensure that the network has no binding preference, we randomize oscillator phases at the beginning. The random initial conditions are illustrated in Fig. 4B, where the diameter of a circle represents the x activity of the corresponding oscillator. A snapshot of the network activity shortly after the beginning is shown in Fig. 4C where the oscillator assembly representing the cup is synchronized and other oscillators are in the excitable state. The response of the same network to 'CUP' is depicted in Figures 4E-G at different times. In this case, the network forms three assemblies corresponding to each of the three letters. Figure H shows the temporal activity of all the enabled oscillators for the connected cup image, where excitable oscillators are omitted. The upper panel of Fig. 4H shows the combined activity of the assembly representing the cup, and the middle panel shows the activity of the global inhibitor. Despite randomized phases to begin with, the assembly reaches synchrony in the first oscillation cycle. The temporal response to the disconnected 'CUP' is shown in Fig. 4I, where synchrony within each of the three assemblies and desynchrony between them are both achieved in the first two cycles. As illustrated in Figs. 4H and 4I, every time an assembly jumps to the active phase the global inhibitor is triggered. Thus, how many assemblies - or put differently how many connected patterns in the input image - is revealed by the ratio of the response frequency of the global inhibitor to the oscillation frequency of an enabled oscillator. A ratio of 1 indicates there is one pattern in the input figure, and thus the figure is connected. A ratio greater than 1 indicates there are more than one pattern in the input figure, and thus the figure is disconnected. Hence the solution to the predicate is given by a simple test of whether the ratio exceeds a threshold θ, chosen in the range $2 > \theta > 1$. The bottom traces of Fig. 4H and Fig. 4I show the ratios, where θ is chosen to be 1.6. As shown in the figure, the connectedness predicate is correctly computed beyond a beginning period that corresponds to the process of assembly formation.
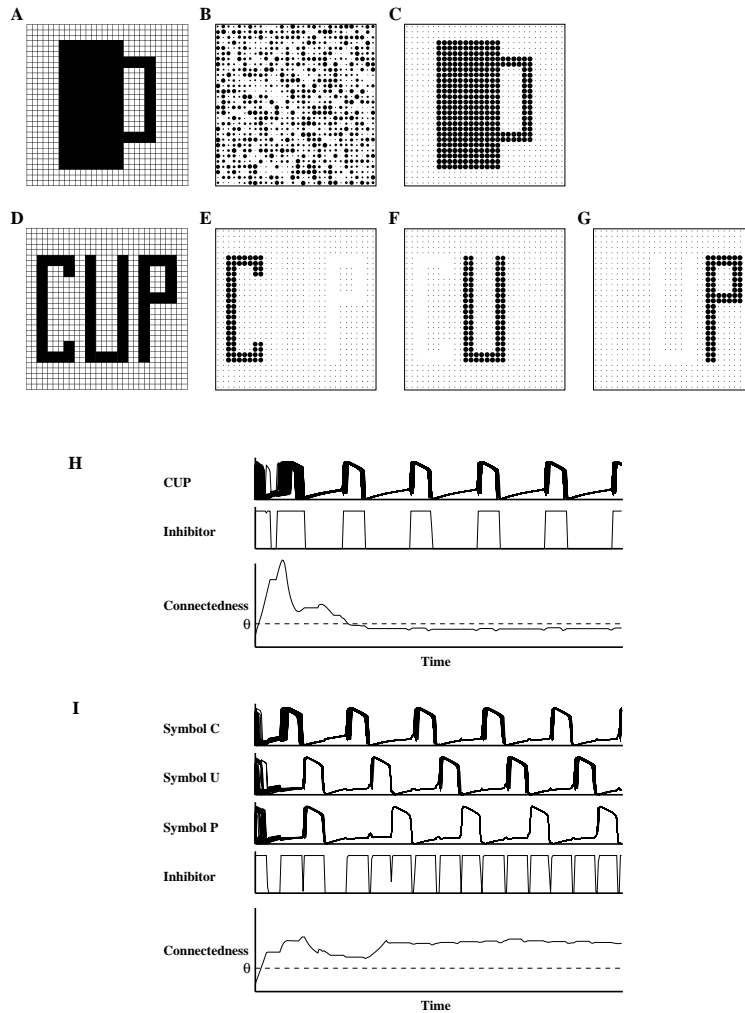
Fig. 4. Oscillatory correlation solution to the connectedness problem (from [64]). A. An input image with 30x30 binary pixels showing a connected cup figure. B. A snapshot from corresponding LEGION network showing the initial conditions of the network. C. A subsequent snapshot of the network activity. D. Another input image depicting three connected patterns forming the word 'CUP'. E.-G. Snapshots of the LEGION network at three different times. H. The upper trace shows the temporal activity of the oscillator assembly representing the connected cup image, the middle trace the activity of the global inhibitor, and the bottom trace the ratio of the global inhibitor's frequency to that of enabled oscillators. The threshold is indicated by the dash line. I. The upper three traces show the temporal activities for the three assemblies representing the three connected patterns in the disconnected 'CUP' image, the next-to-bottom trace the activity of the global inhibitor, and the bottom one the ratio of the global inhibitor's frequency to that of enabled oscillators along with.

The oscillatory correlation theory provides a general framework to address the computational scene analysis problem. The following two sections deal with visual and auditory scene analysis respectively.

5 Visual Scene Analysis

For computational scene analysis, some measure of similarity between features is necessary. What determines if local sensory elements should be grouped into the same object or separated apart? This is the main subject of Gestalt psychology [27, 71]. Major Gestalt grouping principles are summarized below [40]:

- *Proximity.* Sensory elements that are located closely in space tend to be grouped.
- *Similarity.* Elements with similar attributes, such as color, depth, or texture, tend to group.
- *Common fate.* Elements that move together, or show common motion, likely belong to the same object. Common fate is an instance of similarity in a sense, and it is listed separately to emphasize the importance of visual dynamics in perceptual organization.
- *Good continuity.* Elements that form smooth continuations of each other tend to be bound together.
- *Connectedness* and *common region.* Connected elements or elements inside the same connected region have the tendency to group.
- *Familiarity.* Elements that belong to the same memorized pattern tend to group.

To apply the above grouping principles requires a process of feature extraction, which may be a complex operation for certain features such as motion and texture. With extracted features, oscillatory correlation represents a general approach to visual scene analysis. In this approach, an oscillator typically corresponds to a spatial location and connection weights between neighboring oscillators are determined by feature extraction. The oscillator network then evolves autonomously. After assembly formation takes place, different assemblies representing different objects will pop out from the network at different times. It is *segmentation in time* that is unique of this approach to scene analysis. As a result, such segmentation gives rise to the notion of a *segmentation capacity* [69] - at least for networks of relaxation oscillators with a non-instantaneous active phase - that refers to a limited number of oscillator assemblies that may be formed. The segmentation capacity corresponds to the integer ratio of the oscillation period to the duration of the active phase for relaxation oscillators.

The first application of the oscillatory correlation approach to real image segmentation was made by Wang and Terman [69]. Their segmentation system is based on LEGION dynamics. Unlike synthetic images, real images are often
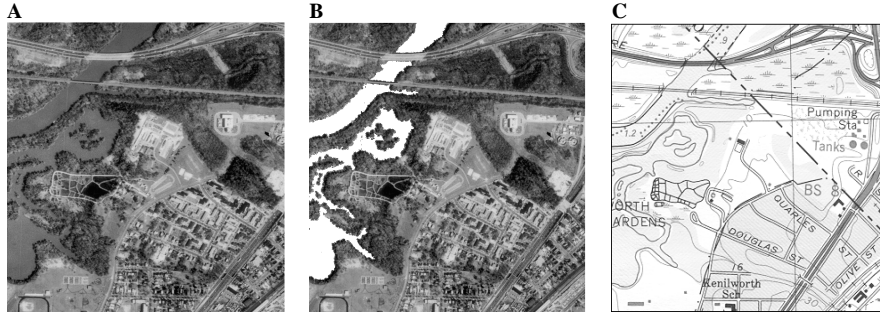
Fig. 5. Extraction of hydrographic regions (from [10]). A. Input satellite image consisting of 640x606 pixels. B. Extraction result, where segmented waterbodies are indicated by white. C. Corresponding 1:24,000 topographic map.

noisy. Image noise may result in many fragments and deteriorate the result of oscillatory correlation. To address the problem of fragmentation, a lateral potential is introduced for each oscillator in order to distinguish between major assemblies and noisy fragments. A major assembly should contain at least one oscillator that lies at the center of a sizable homogeneous region. Such an oscillator, called a leader, has a high lateral potential because it receives a large amount of lateral excitation from its neighborhood. On the other hand, a fragment does not have a leader. All fragments, forming a background, will cease oscillating after a few periods. Another issue that has to be addressed is computing time required for integrating a large oscillator network. To alleviate the computational burden, Wang and Terman abstracted an algorithm from oscillatory dynamics that retains key features of LEGION dynamics, such as jumping and rapid spread of excitation and inhibition. The abstracted algorithm has the option to eliminate the segmentation capacity in order to segment a large number of regions. In the Wang-Terman system, each oscillator is mutually connected with its 8-nearest neighbors, and the connection weight between oscillators i and j is set proportional to $1/(1 + |I_i + I_j|)$, where I_i and I_j represent the corresponding pixel intensities. W_z in (2) is a key parameter that controls the granularity of segmentation, whereby smaller values of W_z produce fewer and larger segments.

In a subsequent study, Chen et al. [10] suggested the idea of weight adaptation to perform feature-preserving smoothing before segmentation. In addition, they proposed to add a logarithmic normalization factor in excitatory coupling (cf. (2)):

$$S_i = \frac{\sum_{k \in N(i)} W_{ik} H(x_k - \theta_x)}{log(\sum_{k \in N(i)} H(x_k - \theta_x) + 1)} - W_z H(z - \theta_z) \qquad (4)$$

The resulting algorithm produces robust segmentation results. An example is given in Figure 5, where the task is to extract hydrographic objects from satellite images from the United States Geological Survey (USGS). Figure

A   **B**   **C**
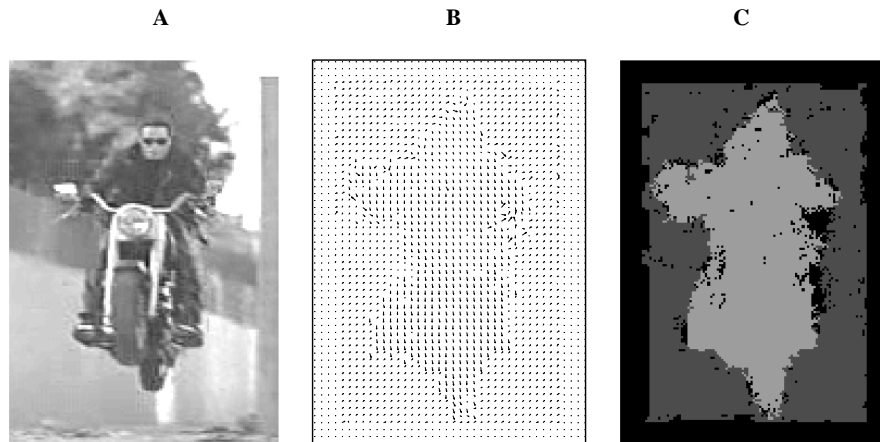
Fig. 6. Motion segmentation (from [8]). A. A frame of a motion sequence. B. Estimated optic flow. C. Result of segmentation.

5A gives the original image containing water bodies, and Figure 5B shows the corresponding extraction results, where the extracted waterbodies are displayed as white and overlaid on the original image. For reference, Figure 5C provides the corresponding map from the USGS. A careful comparison should reveal that the extracted waterbodies match the image better than the map, since the latter is often not up to date.

Cesmeli and Wang [8] applied LEGION to motion-based segmentation that considers motion as well as intensity for analyzing image sequences (see also [75]). In their system, two pathways perform an initial optic flow estimation and intensity-based segmentation in parallel. A subsequent network combines the two to refine local motion estimates. Motion analysis and intensity analysis complement each other since the former tends to be reliable for inhomogeneous, textured regions while the latter is most effective for homogeneous regions. The use of LEGION for segmentation allows for multiple motions at the same location, as in the case of motion transparency. The resulting system significantly reduces erroneous motion estimates and improves boundary localization. A typical example is given in Figure 6. A frame of a motion sequence is shown in Fig. 6A, where a motorcycle rider jumps to a dry canal with his motorcycle while the camera is tracking him. Due to the camera motion, the rider and his motorcycle have a downward motion with a small rightward component and the image background has an upright diagonal motion. Figure 6B shows the estimated optic flow after integrating motion and brightness analyses, and it is largely correct. The rider with his motorcycle is then segmented from the image background as depicted in Figure 6C. Their oscillator model has been favorably compared with a number of motion segregation algorithms including the one by Black and Anandan [5] based on robust statistics.

A large number of studies have applied the oscillatory correlation approach to visual scene analysis tasks, including segmentation of range and texture images, extraction of object contours, and selection of salient objects. See [65] for a recent review on this subject.

6 Auditory Scene Analysis

What grouping principles govern auditory scene analysis? Bregman systematically addresses this question in a comprehensive book [6]. According to Bregman, grouping principles analogous to Gestalt laws revealed in the visual domain are responsible for the segregation of auditory input into streams. Displaying the acoustic input in a 2-D time-frequency (T-F) representation such as a spectrogram, major grouping principles for auditory scene analysis (ASA) are given below [6, 13]:

- *Proximity in frequency and time.* Two tones that are close in frequency or time tend to be grouped into the same stream (an auditory object).
- *Periodicity.* Harmonically related frequency components tend to grouped.
- *Onset and offset.* Frequency components that onset or offset at the same time tend to be organized into the same stream.
- *Amplitude and frequency modulation.* Frequency components that have common temporal modulation tend to be grouped together. This principle applies to both amplitude modulation and frequency modulation.
- *Continuous/smooth transition.* Tones that form a continuous, or discontinuous but smooth, trajectory tend to be fused.
- *Familiarity.* Sound components belonging to the same learned pattern, such as a syllable, have the tendency to group.

Auditory scene analysis takes place in two stages in the brain according to Bregman [6]. The first stage, known as the segmentation stage [66], decomposes the acoustic mixture reaching the ears into a collection of time-frequency segments, each corresponding to a contiguous region in a T-F representation. The second stage groups the segments into streams.

The first study on auditory segregation using oscillatory correlation was made by von der Malsburg and Schneider [61]. As discussed in Section 3, their segregation is based on common onsets. However, their model relies on global connectivity to achieve synchronization among the oscillators that are stimulated at the same time. Desynchronization is obtained with a global inhibitor. Subsequently Wang [63] studied stream segregation by employing a 2-D LEGION network, where one dimension represents time and another one represents frequency. With appropriate connectivity, the LEGION network exhibits a set of psychophysical phenomena, such as dependency of stream segregation on spectrotemporal proximity and competition among different perceptual organizations (see [9, 38] for recent extensions).

Wang and Brown [66] studied a more complex problem: Segregation of voiced speech from its acoustic interference. After feature extraction using a model of auditory periphery that comprises cochlear filtering and mechanical to neural transduction, they compute a number of mid-level representations from peripheral responses, including a correlogram (autocorrelation) and a summary correlogram. The core of their model is an oscillator network with two layers performing auditory segmentation and grouping, respectively. The two-layer structure is designed to embody Bregman's two-stage notion. Auditory segmentation is based on cross-channel correlation in the frequency domain and temporal continuity in the time domain. Specifically, the first layer is a LEGION network where each oscillator is connected with its four nearest neighbors in time and frequency. The connection weight along the frequency axis is set to one if the corresponding cross-channel correlation exceeds a certain threshold and zero otherwise. The connection weight along the time axis is set to one uniformly. In response to an input mixture, the segmentation layer produces oscillator assemblies, representing regions of acoustic energy such as harmonics or formants. The second layer groups the segments that emerge from the first layer. Specifically, this layer contains lateral connections with both excitation and inhibition but no global inhibitor. Grouping in this layer is based on the dominant pitch extracted from the summary correlogram within each time frame. The extracted dominant pitch is used to divide the oscillators of the frame into two classes: One is consistent with the pitch frequency and the other is not. Then excitatory connections are formed between the oscillators of the same class and inhibitory connections are formed between the two classes. This pattern of connectivity within the grouping layer promotes synchronization among a group of segments that have common periodicity.

Figure 7 gives an example of segregating a mixture of a male utterance and telephone ringing. Figure 7A displays the peripheral response to the mixture. The 2-D response is generated by a filterbank with 128 channels over 150 time frames. Figure 7B shows a snapshot of the grouping layer, where active oscillators, indicated by white pixels, primarily correspond to the speech utterance. Figure 7C shows another snapshot of the grouping layer taken at a different time. At this time, active oscillators mostly correspond to the background, i.e. the telephone ringing.

Wrigley and Brown [72] recently proposed an oscillator network to model auditory selective attention. Their model first performs peripheral processing and then auditory segmentation. A unique part of model is an interactive loop between an oscillator layer that performs stream segregation and a leaky integrator that simulates the attentional process. The weights of the connections between the oscillator layer and the leaky integrator are subject to modulation by the attentional interest of the model. Through this interaction, the attentional leaky integrator selects one dominant stream from the stream segregation layer. Their network successfully simulates a number of auditory grouping phenomena, including two-tone streaming with distracted attention
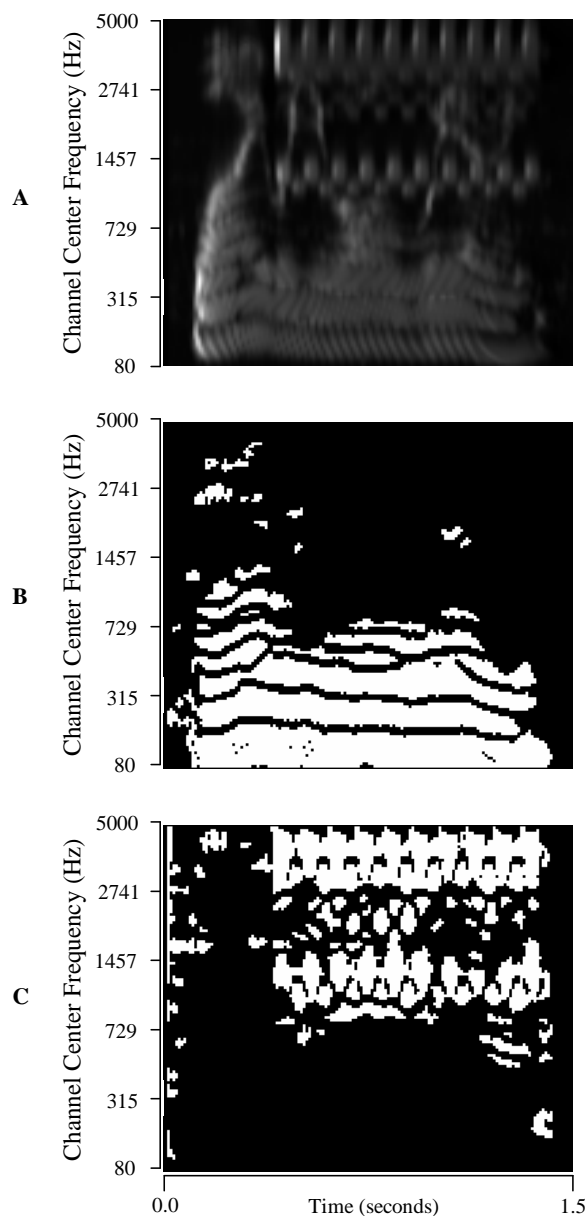
Fig. 7. Segregation of voiced speech from telephone ringing (from [66]). A. Peripheral response to an auditory stimulus consisting of a male utterance mixed with telephone ringing. A bank of 128 filters having center frequencies ranging from 80 Hz to 5 kHz is employed in peripheral processing. B. A snapshot of the grouping layer. Here, white pixels denote active oscillators that represent the segregated speech stream. C. Another snapshot showing the segregated background.

and sequential capturing. At a conceptual level, a major difference between this model and Wang's model [63] concerns whether attention can be directed to more than one stream: In the Wrigley and Brown model only one stream may be attended to at a time whereas in Wang's model attention may be divided by more than one stream. This issue will be revisited in Sect. 7.1.

7 Challenging Issues

7.1 Attention

The importance of attention for scene analysis can hardly be overstated. In a way, to perceive is to attend.

The issues of binding and attention are related. It has been frequently suggested that selective attention plays the role of binding. According to the popular feature integration theory of Treisman and Gelade [57], the visual system first analyzes a scene in parallel by separate retinotopic feature maps and focal attention then integrates the analyses within different feature maps to produce a coherent percept. In other words, attention provides a 'spotlight' on the location map to bind and select an object [55]. Arguing from the neurobiological perspective, Reynolds and Desimone [43] also suggested that attention provides a solution to the binding problem. An alternative view - object-based theories of attention [41, 42] - claims that selective attention operates on the result of binding. So the key question is whether attention precedes or succeeds binding.

A visual object can have an arbitrary shape and size. This consideration creates the following dilemma for the feature integration theory. On the one hand, it is a location-based theory of attention that binds at the same location individual analyses from different feature maps. On the other hand, to select an object, attention spotlight must also have arbitrary shape and size, adapting to a specific object and thus object-based. Without a binding process, what produces such an adaptive spotlight? This is an intrinsic difficulty if focal attention, rather than perceptual organization, is to bind features across different locations. The difficulty is illustrated by the finding of Field et al. [18] that a path of curvilinearly aligned (snake-like) orientation elements embedded in a background of randomly oriented elements can be readily detected by observers, whereas other paths cannot. This is illustrated in Fig. 8, which shows a snake-like pattern (left panel) in a cluttered background (right panel). Note that there are virtually an infinite number of snake patterns that can be constructed from orientation elements. Grouping seems to be required to yield organized patterns for attentional spotlight.

This difficulty, however, does not occur in object-based theories, in which binding provides multiple segments for focal attention to perform sequential analysis. Though sometimes difficult to tear object-based attention apart from
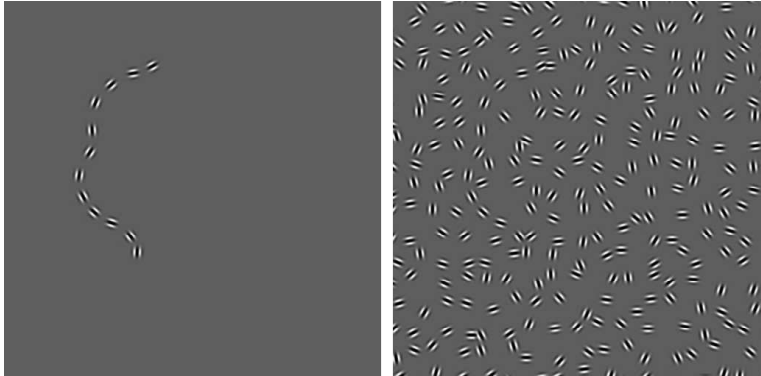
Fig. 8. Detection of snake-like patterns (from [18] with permission). Human observers can easily detect the occurrence of a snake-like pattern - shown on the left - that is embedded in a background of random orientation elements shown on the right. The snake pattern consists of 12 aligned orientation elements.

location-based attention, since the former implicitly provides the information for the latter, psychophysical and neurobiological studies show increasing support for the object-based view [15, 31, 37]. For example, a recent study demonstrates that the visual search for a target item in the presence of many distractors is very efficient if the distractors can be organized into a small number of groups on the basis of feature similarity, suggesting that visual attention examines organized groups rather than individual items [67]. Indeed, the Field et al. results have been successfully simulated by the oscillation model of Yen and Finkel [74].

The notion of a segmentation capacity (see Sect. 5) is a basic characteristic of the oscillatory correlation theory. A limited capacity naturally arises from relaxation oscillations because of their non-instantaneous active phase. On the other hand, networks of spiking neurons [7] or chaotic maps [76] do not exhibit a limited capacity. Although such a capacity is sometimes treated as a computational weakness [14, 70, 76], capacity limitation is a fundamental property of attention. Also it has been argued that a limited capacity is advantageous for information processing (e.g. [25, 29]).

Assuming a limited capacity, a related question is: Can we attend to more than one object at a time? A direct answer was offered by Cowan [12] after reviewing a large body of literature. His answer is that the attentional capacity is about four. Furthermore, the attentional capacity underlies the well-documented capacity in short-term memory. How to reconcile between a capacity limit of more than one and the phenomenological impression that we can focus on only one thing at a time? A capacity limit represents an upper bound on the number of items held by attention, and it does not necessarily mean that the attention span is constantly full. It may be possible for a subject to selectively attend to one thing in order to extract information from it.

Even in the case of selective attention, however, unselected items still receive some analysis. In the classic experiment of Cherry [11], for example, listeners can detect the change of the speaker gender from the 'unattended' ear.

Another important question regarding attention is what to attend when faced with a myriad of stimuli? This decision can be a matter of survival for an animal. Attention can be either goal-driven or stimulus-driven [73]. When the perceiver seeks to find something, e.g. when it looks for a pen in an office, its attention is goal-driven (also called active attention). In contrast, when the perceiver's attention is captured by some salient stimulus in the input scene, such as a red pen on a gray desk, attention is said to be stimulus-driven (or passive attention). It is important to realize that these two modes of attention likely occur simultaneously in a given act of attention. Goal-driven attention is controlled by the perceiver's intentions at the moment. Stimulus-driven attention, on the other hand, can be studied by varying stimulus properties of the input scene. Perceptual studies [42, 73] suggest that stimuli that differ from the rest of the scene in one or more feature dimensions, e.g. color, depth, and motion for vision, tend to capture attention. In other words, salient objects draw attention. The saliency of a stimulus has two aspects. The first is the difference between the stimulus and its surround and the second is the homogeneity of the surround [16]; a stimulus is highly salient when it is different from its surrounding stimuli that are similar to each other. Visual feature dimensions include luminance, color, orientation, motion, and depth. Auditory feature dimensions include loudness, pitch, temporal modulation, and location. In addition to feature saliency, abrupt changes to the scene tend to capture attention [73], including the onset of a new stimulus in the scene and the abrupt change in a feature dimension of an existing stimulus. In other words, novel objects draw attention.

7.2 Feature-based Analysis versus Model-based Analysis

Scene analysis can be performed on the basis of the features of the objects in the input scene or the models of the objects in the memory. Feature-based versus model-based analysis is also metaphorically characterized as bottom-up versus top-down analysis. Familiarity has been acknowledged as an organizing principle in scene analysis so the issue is not whether memorized objects influence scene analysis. What's at issue is how much model-based analysis contributes to scene analysis, or whether binding should be part of a recognition process.

According to some, binding is a byproduct of recognition, which is typically coupled with some selection mechanism that brings the pattern of interest into focus, and there is really no binding problem so to speak [44]. For example, Fukushima and Imagawa [20] proposed a model that performs recognition and segmentation simultaneously by employing a search controller that selects a small area of the input image for processing. Their model is based on

Fukushima's neocognitron model for pattern recognition, which is a hierarchical multilayer network. The neocognitron model is a prominent example of the hierarchical coding approach to the binding problem. The model contains a cascade of layers with both forward and backward connections. The forward path performs pattern recognition that is robust to a range of variations in position and size, and the last layer stores learned patterns. When a scene of multiple patterns is presented, a rough area selection is performed based on feature density of the input, and further competition in the last layer leads to a winner. The winning unit of the last layer, through backward connections, reinforces the pattern of the input image that is consistent with the stored template. This, in a way, segments that part of the input image from its background. After a while, the network switches to another area of high feature density and continues the analysis process. Their model has been evaluated on binary images of connected characters. Olshausen et al. [39] proposed a model that also combines pattern recognition and a model of selective attention. Their attention model is implemented by a shifting circuit that routes information in a hierarchical network while preserving spatial relations between visual features, and recognition is based on a Hopfield model of associative memory. The location and size of an attention blob are determined by competition in a feature saliency map, producing potential regions of interest on an image. This model is highlighted by Shadlen and Movshon [50] as an alternative to the temporal correlation theory. The model is evaluated on binary images with well-separated patterns. A later model along a similar line was proposed by Riesenhuber and Poggio [44], and it uses a hierarchical architecture similar to the neocognitron. Their model has been tested on two-object scenes: One is a stored pattern and another is a distractor.

In my opinion, model-based analysis has clear limits. Perceiving an object, e.g. a frog, with all its vivid details such as location, shape, color, orientation, and size, is more than simply recognizing that the object is a frog [24, 56]. Indeed, if feature analysis played no role in scene analysis, camouflage would not have emerged from animal evolution as a universal strategy of blending with the environment. This point is illustrated in Figure 9 which shows two frogs in a pond. It is effortful to spot a frog in its natural habitat even for an experienced observer. Also, perception operates on both familiar and unfamiliar objects, and model-based analysis is not applicable to the latter objects. Besides these and other conceptual difficulties with the hierarchical coding discussed in Section 3, it is unclear how the above model-based systems can be extended to analyze scenes where complex objects are arranged in arbitrary ways. As mentioned in Sect. 7.1, the number of possible snake patterns (see Fig. 8) seems too large to search in a top-down manner.

7.3 Learning versus Representation

Learning - both supervised and unsupervised - is central to neural networks (and computational intelligence in general). The development of neural net-

Fig. 9. A natural image that contains two frogs in their natural habitat.

works can be characterized, to a large extent, by the development of learning algorithms. Nowadays, much activity in neural networks is popularly called machine learning. There is also increasing interest in the research community to apply machine learning techniques to scene analysis. Some even argue that data-driven learning can do away with the need to search for appropriate representations for computational scene analysis.

While certain regularities of input data can be discovered by a learning system, the applicability of learning-based approaches to computational scene analysis is likely bounded. As pointed out by Konen and von der Malsburg [28], such approaches tend to be plagued by the problem of combinatorial explosion when dealing with realistically complex scenes. It is perhaps revealing to consider the connectedness problem in this context. The failure of perceptrons to solve this problem is rooted in the lack of a proper representation, not the lack of a powerful learning method. According to Minsky and Papert [34], "no machine can learn to recognize X unless it possesses, at least potentially, some scheme for representing X." (p. xiii). Indeed, modern multilayer perceptrons with the powerful backpropagation algorithm fare no better on the connectedness problem [65]. No learning machine, to my knowledge, has succeeded in solving this problem. The cause, as discussed in Sect. 4, is computational complexity - learning the connectedness predicate would require far too many training samples and too much learning time. The success of LEGION stems from the oscillatory correlation representation and the network structure.

The brain of a newborn possesses genetic knowledge resulting from millions of years of evolution. It is relevant to note that success is quite limited in teaching chimpanzees, the closest relatives of humans, basic language or arithmetic despite considerable effort by animal behaviorists. While in theory all is learnable, including genetic knowledge, evolution operates at much greater time scales. Even if evolutionary computing someday succeeded in uncovering the computational principles of evolution, the challenge would be insurmountable of simulating billions of years of environmental change that resulted in the flourishing of life on the earth. Furthermore, even if major evolutionary processes were totally reproducible on a computer there would still be no assurance that the result will a human rather than an amoeba.

Computational complexity should be of principal concern in computational intelligence. The essence of intelligence is the efficiency of information processing. Although stunning progress in computer speed and memory has enabled the execution of very complex algorithms, we should keep in mind that a slower algorithm will always be slower no matter how fast the computer is.

For those who are concerned with biological plausibility, the speed of human scene analysis has strong implications on the kind of processing employed. For visual analysis, it has been empirically shown that object identification in a visual scene takes less than 150 ms [4, 54]. Interestingly, the visual system categorizes novel scenes just as fast as highly familiar ones [17]. After estimation of processing time at various stages of the visual pathway, Thorpe and Fabre-Thorpe [53] conclude that such analysis must be based primarily on feedforward processing as there is little time left for iterative feedback. Coincidentally, a comprehensive analysis on noise robustness, time course, and language context led Allen [1] to essentially the same conclusion, that is, human speech analysis is primarily a bottom-up process. These results challenge the biological validity of the contemporary emphasis on statistical, model-based approaches [26] that typically boil down to a time-consuming search in a large probability space.

A major distinction between perception and reasoning is that the process of perception is rapid and automatic, whereas the process of reasoning is consciously deliberative and generally slow. Even when faced with certain ambiguous figures that permit multiple interpretations, such as the famous Necker cube shown in Fig. 10, perception seems to quickly dwell on one of the plausible interpretations and would require a slow, conscious effort to switch to a competing interpretation. This is not to say that perception never involves conscious deliberations. We do, on occasion, debate in our head how to make of an input scene. But such an experience is more of the exception rather than the rule. From an ecological point of view, perception needs to figure out what is out there quickly as the scene changes constantly due to both environmental change and locomotion. The speed of perception is critical to the survival of an organism.

The emphasis on representations contrasts that on learning. A representation is a formal system that encodes certain types of information. Marr's pio-
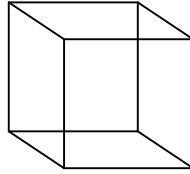
Fig. 10. Necker cube. This figure can be seen as a cube that is viewed either from above or from below.

neering study on computational vision exemplifies representational approaches to scene analysis [30]. In my view, the Marrian framework for computational perception provides the most promising roadmap for understanding scene analysis. Appropriate representations embody key insights and constraints in an information processing domain. How information is represented can profoundly affect how information is processed. For instance, the cepstral representation[3] separates voice excitation from vocal tract filtering [22], and the discovery of this representation pays a huge dividend to speech processing tasks including automatic speech recognition where cepstral features are an indispensable part of any state-of-the-art system.

Since a good representation often captures the current state of scientific understanding on human perception, it does not seem to make sense to let a computer program 'discover' it through machine learning. For example, cochlear processing of the acoustic information is well understood and amounts to an elaborate filterbank. Why not codify such understanding, which is nontrivial to figure out from scratch, in a representation of auditory periphery?

The above discussion makes it plain that the investigation of computational scene analysis can be characterized in large part as the pursuit of appropriate representations. This vision implies that the research in computational scene analysis is an interdisciplinary enterprise, as psychophysics as well as cognitive neuroscience contributes to uncovering effective representations.

So what is the role of learning in computational scene analysis? A representation provides a framework, a skeleton, but it is by no means all that is needed to solve the computational scene analysis problem. Learning plays an important role within a representational framework to adjust parameter values and precisely model the distribution of input data in relation to the system. A recent study by Roman et al. [45] offers an example in this regard. Their binaural system for speech segregation builds on an auditory peripheral model and the notion of binary time-frequency masks for speech separation, and computes the binaural cues of interaural time difference and interaural intensity difference which are known to be employed by the auditory system. However, their use of supervised training is responsible for high-quality estimates of binary masks, which in turn lead to good segregation results. In other words, effective learning can substantially enhance system performance.

---

[3] A cepstrum corresponds to the logarithm of the magnitude spectrum.

8 Concluding Remarks

In this chapter I have made an effort to define the goal of computational scene analysis explicitly. The challenges facing Rosenblatt's perceptrons are fundamentally related to the binding problem. Temporal correlation provides a biologically plausible framework to address the binding problem. Advances in understanding oscillatory dynamics lead to the development of the oscillatory correlation approach to computational scene analysis with promising results.

Perhaps the most important characteristic of natural intelligence compared to artificial intelligence is its versatility. Natural intelligence ranges from sensation, perceptual organization, language, motor control, to decision making and long-term planning. The substrate for all these functions is a brain - an immense network of neurons - whose structure is largely fixed after development. I have argued recently that time provides a necessary dimension to the versatility of brain function [65]. Temporal structure is shared by neuronal responses in all parts of the brain, and the time dimension is flexible and infinitely extensible.

Computational scene analysis is an extremely challenging problem. The bewildering complexity of perception makes it necessary to adopt a compass to guide the way forward and avoid many pitfalls along the way. I strongly recommend Marr's posthumous book to anyone who is attempted by the challenge.

Acknowledgments

References

[1] Allen JB (2005) Articulation and intelligibility. Morgan & Claypool
[2] Arbib MA ed (2003) Handbook of brain theory and neural networks. 2nd ed, MIT Press, Cambridge MA
[3] Barlow HB (1972) Single units and cognition: A neurone doctrine for perceptual psychology. Percept 1:371-394
[4] Biederman I (1987) Recognition-by-component: A theory of human image understanding. Psychol Rev 94:115-147
[5] Black MJ, Anandan P (1996) The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. CVGIP: Image Understanding 63:75-104
[6] Bregman AS (1990) Auditory scene analysis. MIT Press, Cambridge MA

[7] Campbell SR, Wang DL, Jayaprakash C (1999) Synchrony and desynchrony in integrate-and-fire oscillators. Neural Comp 11:1595-1619

[8] Cesmeli E, Wang DL (2000) Motion segmentation based on motion/brightness integration and oscillatory correlation. IEEE Trans Neural Net 11:935-947

[9] Chang P (2004) Exploration of behavioral, physiological, and computational approaches to auditory scene analysis. MS Thesis, The Ohio State University Department of Computer Science and Engineering (available at http://www.cse.ohio-state.edu/pnl/theses.html)

[10] Chen K, Wang DL, Liu X (2000) Weight adaptation and oscillatory correlation for image segmentation. IEEE Trans Neural Net 11:1106-1123

[11] Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. J Acoust Soc Am 25:975-979

[12] Cowan N (2001) The magic number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci 24:87-185

[13] Darwin CJ (1997) Auditory grouping. Trends Cogn Sci 1:327-333

[14] Domijan D (2004) Recurrent network with large representational capacity. Neural Comp 16:1917-1942

[15] Driver J, Baylis GC (1998) Attention and visual object recognition. In: Parasuraman R (ed) The attentive brain. MIT Press Cambridge MA, pp.299-326

[16] Duncan J, Humphreys GW (1989) Visual search and stimulus similarity. Psychol Rev, 96:433-458

[17] Fabre-Thorpe M, Delorme A, Marlot C, Thorpe S (2001) A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. J Cog Neurosci 13:1-10

[18] Field DJ, Hayes A, Hess RF (1993) Contour integration by the human visual system: Evidence for a local "association field". Vis Res 33:173-193

[19] FitzHugh R (1961) Impulses and physiological states in models of nerve membrane. Biophys J 1:445-466

[20] Fukushima K, Imagawa T (1993) Recognition and segmentation of connected characters with selective attention. Neural Net 6:33-41

[21] Gibson JJ (1966) The senses considered as perceptual systems. Greenwood Press, Westport CT

[22] Gold B, Morgan N (2000) Speech and audio signal processing. Wiley & Sons, New York

[23] Gray CM (1999) The temporal correlation hypothesis of visual feature integration: still alive and well. Neuron 24:31-47

[24] Kahneman D, Treisman A, Gibbs B (1992) The reviewing of object files: object-specific integration of information. Cognit Psychol 24:175-219

[25] Kareev Y (1995) Through a narrow window: Working memory capacity and the detection of covariation. Cognition 56:263-269

[26] Knill DC, Richards W eds (1996) Perception as Bayesian inference. Cambridge University Press, New York

[27] Koffka K (1935) Principles of Gestalt psychology. Harcourt, New York

[28] Konen W, von der Malsburg C (1993) Learning to generalize from single examples in the dynamic link architecture. Neural Comp 5:719-735

[29] MacGregor JN (1987) Short-term memory capacity: Limitation or optimization? Psychol Rev 94:107-108

[30] Marr D (1982) Vision. Freeman, New York

[31] Mattingley JB, Davis G, Driver J (1997) Preattentive filling-in of visual surfaces in parietal extinction. Science 275:671-674

[32] Milner, PM (1974) A model for visual shape recognition. Psychol Rev 81(6):521-535

[33] Minsky ML, Papert SA (1969) Perceptrons. MIT Press, Cambridge MA

[34] Minsky ML, Papert SA (1988) Perceptrons (Expanded ed). MIT Press, Cambridge MA

[35] Morris C, Lecar H (1981) Voltage oscillations in the barnacle giant muscle fiber. Biophys J 35:193-213

[36] Nagumo J, Arimoto S, Yoshizawa S (1962) An active pulse transmission line simulating nerve axon. Proc IRE 50:2061-2070

[37] Nakayama K, He ZJ, Shimojo S (1995) Visual surface representation: A critical link between lower-level and higher-level vision. In: Kosslyn SM, Osherson DN (eds) An invitation to cognitive science. MIT Press, Cambridge MA, pp. 1-70

[38] Norris M (2003) Assessment and extension of Wang's oscillatory model of auditory stream segregation. PhD Dissertation, University of Queensland School of Information Technology and Electrical Engineering

[39] Olshausen BA, Anderson CH, Van Essen DC (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. J Neurosci 13:4700-4719

[40] Palmer SE (1999) Vision science. MIT Press, Cambridge MA

[41] Parasuraman R ed (1998) The attentive brain. MIT Press, Cambridge MA

[42] Pashler HE (1998) The psychology of attention. MIT Press, Cambridge MA

[43] Reynolds JH, Desimone R (1999) The role of neural mechanisms of attention in solving the binding problem. Neuron 24:19-29

[44] Riesenhuber M, Poggio T (1999) Are cortical models really bound by the "binding problem"? Neuron 24:87-93

[45] Roman N, Wang DL, Brown GJ (2003) Speech segregation based on sound localization. J Acoust Soc Am 114:2236-2252

[46] Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386-408

[47] Rosenblatt F (1962) Principles of neural dynamics. Spartan, New York

[48] Rumelhart DE, McClelland JL eds (1986) Parallel distributed processing 1: Foundations. MIT Press, Cambridge MA

[49] Russell S, Norvig P (2003) Artificial intelligence: A modern approach. 2nd ed Prentice Hall, Upper Saddle River, NJ

[50] Shadlen MN, Movshon JA (1999) Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron 24:67-77.

[51] Somers D, Kopell N (1993) Rapid synchrony through fast threshold modulation. Biol Cybern, 68:393-407

[52] Terman D, Wang DL (1995) Global competition and local cooperation in a network of neural oscillators, Physica D 81:148-176

[53] Thorpe S, Fabre-Thorpe M (2003) Fast visual processing. In: Arbib MA (ed) Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge MA, pp. 441-444

[54] Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520-522

[55] Treisman A (1986) Features and objects in visual processing. Sci Am, November, Reprinted in The perceptual world, Rock I (ed). Freeman and Company, New York, pp. 97-110

[56] Treisman A (1999) Solutions to the binding problem: progress through controversy and convergence. Neuron 24:105-110

[57] Treisman A, Gelade G (1980) A feature-integration theory of attention. Cognit Psychol 12:97-136

[58] van der Pol B (1926) On "relaxation oscillations". Phil Mag 2(11):978-992

[59] von der Malsburg C (1981) The correlation theory of brain function. Internal Report 81-2, Max-Planck-Institute for Biophysical Chemistry, Reprinted in Models of neural networks II, Domany E, van Hemmen JL, Schulten K, eds (1994) Springer, Berlin

[60] von der Malsburg C (1999) The what and why of binding: the modeler's perspective. Neuron 24:95-104

[61] von der Malsburg C, Schneider W (1986) A neural cocktail-party processor. Biol Cybern 54:29-40

[62] Wang DL (1995) Emergent synchrony in locally coupled neural oscillators. IEEE Trans Neural Net 6(4):941-948

[63] Wang DL (1996) Primitive auditory segregation based on oscillatory correlation. Cognit Sci 20:409-456

[64] Wang DL (2000) On connectedness: a solution based on oscillatory correlation. Neural Comp 12:131-139

[65] Wang DL (2005) The time dimension for scene analysis. IEEE Trans Neural Net 16:1401-1426

[66] Wang DL, Brown GJ (1999) Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans Neural Net 10:684-697

[67] Wang DL, Kristjansson A, Nakayama K (2005) Efficient visual search without top-down or bottom-up guidance. Percept Psychophys 67:239-253

[68] Wang DL, Terman D (1995) Locally excitatory globally inhibitory oscillator networks. IEEE Trans Neural Net 6(1):283-286

[69] Wang DL, Terman D (1997) Image segmentation based on oscillatory correlation. Neural Comp 9:805-836 (for errata see Neural Comp 9:1623-1626)

[70] Wersing H, Steil JJ, Ritter H (2001) A competitive-layer model for feature binding and sensory segmentation. Neural Comp 13:357-388

[71] Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt, II. Psychol Forsch 4:301-350

[72] Wrigley SN, Brown GJ (2004) A computational model of auditory selective attention. IEEE Trans Neural Net 15:1151-1163

[73] Yantis S (1998) Control of visual attention. In: Pashler H (ed) Attention. Psychology Press, London, pp. 223-256

[74] Yen SC, Finkel LH (1998) Extraction of perceptually salient contours by striate cortical networks. Vis Res 38:719-741

[75] Zhang X, Minai AA (2004) Temporally sequenced intelligent block-matching and motion-segmentation using locally coupled networks. IEEE Trans Neural Net 15:1202-1214

[76] Zhao L, Macau EEN (2001) A network of dynamically coupled chaotic maps for scene segmentation. IEEE Trans Neural Net 12:1375-1385