Neural Networks 24 (2011) 54–64

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Selecting salient objects in real scenes: An oscillatory correlation model

Marcos G. Quiles [a], DeLiang Wang [b,*], Liang Zhao [c], Roseli A.F. Romero [c], De-Shuang Huang [d]

[a] Department of Science and Technology, Federal University of São Paulo (Unifesp), São José dos Campos, SP, Brazil
[b] Department of Computer Science & Engineering and Center for Cognitive Science, The Ohio State University (OSU), Columbus, OH 43210, USA
[c] Department of Computer Science, Institute of Mathematics and Computer Science, University of São Paulo (USP), São Carlos, SP, Brazil
[d] The Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China

ARTICLE INFO

Article history:
Received 20 April 2010
Received in revised form 6 September 2010
Accepted 7 September 2010

Keywords:
Object selection
LEGION
Oscillatory correlation
Visual attention

ABSTRACT

Attention is a critical mechanism for visual scene analysis. By means of attention, it is possible to break down the analysis of a complex scene to the analysis of its parts through a selection process. Empirical studies demonstrate that attentional selection is conducted on visual objects as a whole. We present a neurocomputational model of object-based selection in the framework of oscillatory correlation. By segmenting an input scene and integrating the segments with their conspicuity obtained from a saliency map, the model selects salient objects rather than salient locations. The proposed system is composed of three modules: a saliency map providing saliency values of image locations, image segmentation for breaking the input scene into a set of objects, and object selection which allows one of the objects of the scene to be selected at a time. This object selection system has been applied to real gray-level and color images and the simulation results show the effectiveness of the system.

1. Introduction

The feeling of seeing everything around us is a mere illusion. At a given time, only a small part of the visual scene undergoes scrutiny and reaches the level of awareness. The perceptual mechanism of selecting a part of the visual input for conscious analysis is called selective visual attention, and it is a mechanism that is fundamentally important for the survival of an organism (Desimone & Duncan, 1995; Pashler, 1998; Yantis, 1998). Visual attention is thought to involve two aspects (Yantis, 1998). The first one is called bottom-up (or stimulus-driven) attention that is based on analyzing stimulus characteristics of the input scene. Bottom-up control is mostly associated with feature contrast among the items that compose the scene. For example, when a red item is presented among green ones, it pops out from the visual scene to the eye. The second aspect is top-down control (or goal-driven attention) that is influenced by the intention of the viewer, like looking for a specific thing.

Besides the stimulus-driven and goal-driven aspects of attentional control, an important component of visual attention is selection, concerning how to select a part of a visual scene for further analysis. The visual system can select spatial locations (location-based attention), visual features (feature-based attention), or objects (object-based attention) (for reviews see Egeth & Yantis, 1997; Yantis, 2000). Recent behavioral and neurophysiological evidence establishes that the selection of objects plays a central role in primate vision (Martinez, Ramanathan, Foxe, Javitt, & Hillyard, 2007; O'Craven, Downing, & Kanwisher, 1999; Richard, Lee, & Vecera, 2008; Roelfsema, Lamme, & Spekreijse, 1998; Shinn-Cunningham, 2008; Wang, Kristjansson, & Nakayama, 2005). It is believed that a preattentive process, in the form of perceptual organization, is performed unconsciously by the brain. This process is responsible for segmenting the visual scene into a set of objects which then act as wholes in the competition for attentional selection (Desimone & Duncan, 1995). Perceptual organization has been extensively studied in Gestalt psychology where it is emphasized that the visual world is perceived as an agglomeration of well-structured objects, not as an unorganized collection of pixels. Object formation is governed by Gestalt grouping rules such as connectedness, proximity, and similarity.

Due to the competitive nature of visual selection, most of the neural models are based on winner-take-all (WTA) networks (Itti & Koch, 2001a; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985). Through neural competition, a WTA network selects one neuron, the winner, in response to a given input (Arbib, 2003). In this way, a pixel or location, not an object, of the scene is selected. In Itti et al. (1998), when a neuron wins competition, a circle of a fixed radius surrounding the neuron is considered to be the region receiving attention (spotlight). Usually, these models make use of a two-dimensional saliency map that encodes the conspicuity

* Corresponding author. Tel.: +1 614 292 6827.
E-mail addresses: quiles@unifesp.br (M.G. Quiles), dwang@cse.ohio-state.edu (D. Wang), zhao@icmc.usp.br (L. Zhao), rafrance@icmc.usp.br (R.A.F. Romero), dshuang@iim.ac.cn (D.-S. Huang).

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.09.002

over the visual scene (Itti & Koch, 2001a; Koch & Ullman, 1985). The saliency map is used to direct the deployment of attention (Gottlieb, Kusunoki, & Goldberg, 1998; Itti & Koch, 2001a; Koch & Ullman, 1985). These visual selection models correspond to location-based theories of visual attention, but not object-based theories.

According to Sun and Fisher (2003), object selection has at least the following advantages:

- visual search is more efficient;
- selection of something instead of empty locations;
- it allows for hierarchical selection.

In order to develop a neural model of visual selection that is object-based, one has to address how to group the elements, or features, of a visual scene into a set of coherent objects. The problem of how sensory elements of a scene are combined together to form perceptual objects in the brain is known as the *binding problem* (Revounsuo & Newman, 1999; von der Malsburg, 1981).

Von der Malsburg proposed *temporal correlation theory* to address the binding problem (von der Malsburg, 1981). The theory asserts that objects are represented by the temporal correlation of the firing activities of spatially distributed neurons coding different object features. A natural way of encoding temporal correlation is using synchronization of neural oscillators where each oscillator encodes some feature of an object (Terman & Wang, 1995; von der Malsburg & Schneider, 1986; Wang, 2005; Yu & Slotine, 2009). This form of temporal correlation is called *oscillatory correlation* (Terman & Wang, 1995) whereby oscillators that encode different features of the same object are synchronized and those that encode different objects are desynchronized. Note that binding can occur at multiple levels, including the binding of local pixels to form an image region, which is addressed in this paper, and the binding of region-level features (e.g. shape) to form a high-level entity (e.g. house). The oscillatory correlation theory has been applied to various tasks of scene analysis, such as texture segmentation, motion analysis, and auditory scene segregation (see Wang, 2005, for an extensive review).

Although oscillation-based models for visual attention have been studied for years (Niebur, Koch, & Rosin, 1993), the first attempt to perform object selection using oscillatory correlation was made by Wang (1999). This study achieves size-based object selection based on LEGION (Locally Excitatory Globally Inhibitory Oscillator Network) and a slow inhibition mechanism. Given an input scene composed of several objects, this model selects the largest segment while all the others remain silent thanks to competition among the objects formed by LEGION segmentation. In terms of competition, when a segment becomes active, it sets the slow inhibitor with a value based on the size of the segment, allowing only the segments with larger sizes to overcome the slow inhibition. Thus, after a number of oscillation cycles, only the largest segment survives the competition and keeps oscillating. However, the model considers just object size in competition, which restricts its applicability as a general visual selection model. Size-based selection using oscillatory correlation was also considered by Kazanovich and Borisyuk (2002) where the frequency and amplitude of oscillators are used to perform selection. Their simulations showed that the model can perform consecutive selection of objects, though only synthetic images were used. That model was extended in Borisyuk and Kazanovich (2004) where a novelty detection mechanism using a short-term working memory was incorporated. Although this model aims to solve a more complex cognitive task, it only deals with toy images. A different object-based model for visual attention was proposed in Sun and Fisher (2003). Although this model performs object-based selection, it assumes that perceptual organization has already been done. Another model was proposed by Tiesinga (2005). This model is based on Hodgkin–Huxley type neurons to reproduce experimental results of stimulus competition in V4. The model can produce quantitative results of visual selection albeit the competition is restricted to only two stimuli.

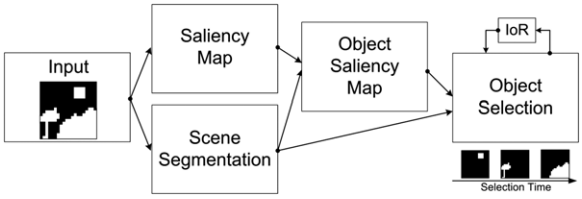

Fig. 1. Diagram of the proposed object selection model, which is composed a saliency map, a scene segmentation module (implemented by a LEGION network), an object-saliency map, and an object selection module that includes an inhibition-of-return (IoR) mechanism. Arrows indicate the computational flow of the system. The images shown below the selection module illustrates a sequence of the objects selected.

Recently, another object selection model was proposed by Chik, Borisyuk, and Kazanovich (2009). This model uses Hodgkin–Huxley neurons in a two-layer architecture. The first layer defines peripheral neurons representing feature detectors and the second layer is composed of two central neurons responsible for the formation of the focus of attention and also the shifting between the objects of the scene. Although this model offers a mechanism to select different objects in real scenes, it does not consider object-level saliency. A related model presents a more complete framework composed of three modules responsible for selective attention, contour extraction, and segmentation (Borisyuk, Kazanovich, Chik, Tikhanoff, & Cangelosi, 2009). Although the results are promising, the concept of object saliency is still missing.

Here we propose an object-based visual selection model with three major components. First, a saliency map is employed to calculate point-wise conspicuity over the input scene. This saliency map is intended to simulate feature- and location-based aspects of visual attention which is based on the contrast between local features, such as color, intensity, and orientation. Second, the LEGION network is used to segment the input image, and this network is intended to perform the task of perceptual organization in a biologically plausible manner. Third, an object-based selection network is proposed. This selection network chooses the most salient object using an *object-saliency map* created by integrating the results from the saliency map and LEGION segmentation. The object-saliency map extends the notion of saliency from a single location to an object. Moreover, based on an inhibition of the return mechanism, our selection network is able to shift from a previous selected object to the next. Fig. 1 shows a flowchart of our model. To our knowledge, this is the first model that can select objects from real scenes based on general object saliency.

We should clarify that, by an object, we mean an image region which roughly corresponds to a visual surface (Marr, 1982). Broadly speaking, an object in a three-dimensional environment includes multiple surfaces, and a complex object such as a car often needs to be defined in a hierarchical manner. This paper focuses on selecting salient regions from visual scenes.

This paper is organized as follows. In Section 2, an overview of the saliency map and LEGION segmentation is presented. Section 3 describes the selection model of the system. Evaluation results are presented in Section 4. Finally, Section 5 offers a few concluding remarks.

2. Background

In this section, we review the saliency map and the segmentation mechanism used in our visual selection model.

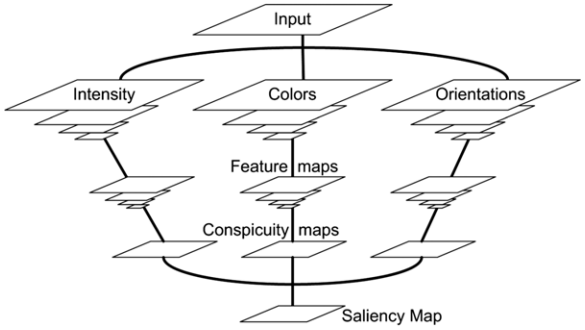

Fig. 2. Flowchart of a saliency map.

2.1. Saliency map

To compute the saliency, we use the saliency map proposed in Itti et al. (1998) and Koch and Ullman (1985). This saliency map mimics the properties of early vision in primates and is based on the idea that a unique map is used to control the deployment of attention (Gottlieb et al., 1998; Itti & Koch, 2001a; Koch & Ullman, 1985).

The saliency map is an explicit two-dimensional map responsible for encoding the saliency over all points of the visual scene. It focuses on the role of local feature contrast in guiding attention (Itti & Koch, 2001a; Itti et al., 1998). Despite its simple architecture based on feedforward feature-extraction mechanisms, this model has proved to have robust performance when dealing with complex scenes and it achieves some qualitative results matching human visual search (Itti & Koch, 2000).

Generally speaking, the saliency map is produced in the following way. First, a set of maps representing primary features, such as color and orientation, are extracted from the input scene. After that, in order to model the center–surround receptive fields, operations are performed over different spatial scales of those maps. This process, followed by a normalization operator (explained later), results in a new set of maps called *feature maps*. Next, feature maps are combined into a set of *conspicuity maps*. Finally, a linear combination of conspicuity maps results in the *saliency map*. A flowchart of this process is shown in Fig. 2.

Formally, given a static image Υ as input, a set of nine spatial-scale dyadic Gaussian pyramids is created by a convolution of a low-pass filter and downsampling of the filtered image by a factor of two (Burt & Adelson, 1983). A dyadic Gaussian pyramid represents a set of images in which the image at one level is a reduced version of the image at the previous level in both resolution and density. Here, a separable Gaussian kernel $[1, 5, 10, 10, 5, 1]/32$ is used. Note that to perform convolution near image borders, the missing pixels have their values set to the mean value of the present pixels. The result is a set of $\Upsilon(i),\ i \in \{0, 1, 2, \ldots, 8\}$, that corresponds to the nine levels from $\Upsilon(0)$ (original image) to $\Upsilon(8)$ (scale eight with a resolution that is $1/256$ of the input image). The Gaussian pyramid provides an efficient way to highlight features at different scales of a scene.

Each $\Upsilon(i)$ is composed of three channels defined as r, g, and $b \in [0, 1]$, which represent red, green, and blue values, respectively. The intensity map, I, for each level (i) of the pyramid is computed as

$$I(i) = \frac{r(i) + g(i) + b(i)}{3}. \tag{1}$$

From the r, g, and b channels we also extract the red–green (RG) and blue–yellow (BY) maps for each level. To extract these color opponencies, we use the definition proposed in Walther and Koch (2006) which gives better results than those in Itti et al. (1998). The RG and the BY maps are defined as follows:

$$RG(i) = \frac{r(i) - g(i)}{\max(r(i), g(i), b(i))} \tag{2}$$

and,

$$BY(i) = \frac{b(i) - \min(r(i), g(i))}{\max(r(i), g(i), b(i))}. \tag{3}$$

Moreover, in order to avoid the hue instability when the intensity level is low, RG and BY are set to zero when $\max(r, g, b) < 0.1$ (Cheng, Jiang, Sun, & Wang, 2001; Gonzalez & Woods, 2002).

Local orientation maps, R_θ, are extracted by convolving I with oriented Gabor filters for four orientations $\theta \in \{0°, 45°, 90°, 135°\}$:

$$R_\theta(i) = |I(i) * G_0(\theta)| + |I(i) * G_{\pi/2}(\theta)| \tag{4}$$

where $G(\theta)$ represents a Gabor kernel with orientation θ, and a subscript indicates the phase of a kernel.

After extracting the intensity (I), color (RG and BY), and orientation maps (R_θ), feature maps are extracted by across-scale subtractions ($\ominus$) between different levels of the same feature. This operation is performed in two steps. First the surround map (s) is rescaled to the size of the center map (c) by a linear interpolation of pixels. After that, a pointwise subtraction is applied. The operator $\ominus$ mimics the center–surround receptive fields in the visual cortex:

$$F_I(c, s) = |I(c) \ominus I(s)| \tag{5}$$

$$F_{RG}(c, s) = |RG(c) \ominus RG(s)| \tag{6}$$

$$F_{BY}(c, s) = |BY(c) \ominus BY(s)| \tag{7}$$

$$F_\theta(c, s) = |R_\theta(c) \ominus R_\theta(s)| \tag{8}$$

where $c \in \{2, 3, 4\}$ represents the levels of the center map and $s \in \{c + 3, c + 4\}$ represents the surround levels. Next, these maps are combined to form the conspicuity maps. The conspicuity map for intensity (C_I) is calculated as follows:

$$C_I = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(F_I(c, s)) \tag{9}$$

where $\oplus$ is an across-scale addition operator and $\mathcal{N}$ is a normalization operator responsible for enhancing the responses of those maps that have a few active locations (high values) and suppressing those with homogeneous activity (Itti & Koch, 2001b; Itti et al., 1998). The normalization operator first normalizes the values of the feature maps to the same range and then multiplies each map by the squared difference between the global maximum and the average of the local maxima for individual maps.

The conspicuity map for colors (C_H) is calculated using the following equation:

$$C_H = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[\mathcal{N}(F_{RG}(c, s)) + \mathcal{N}(F_{BY}(c, s))\right]. \tag{10}$$

Fig. 3 illustrates how the conspicuity map for colors is calculated for a given scene. The conspicuity map for orientation is generated in two steps. First, an intermediary conspicuity map for each orientation is calculated:

$$C_\theta = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(F_\theta(c, s)). \tag{11}$$

Second, these maps are combined into a unique conspicuity map representing all orientations:

$$C_R = \sum_{\theta \in \{0°, 45°, 90°, 135°\}} \mathcal{N}(C_\theta). \tag{12}$$

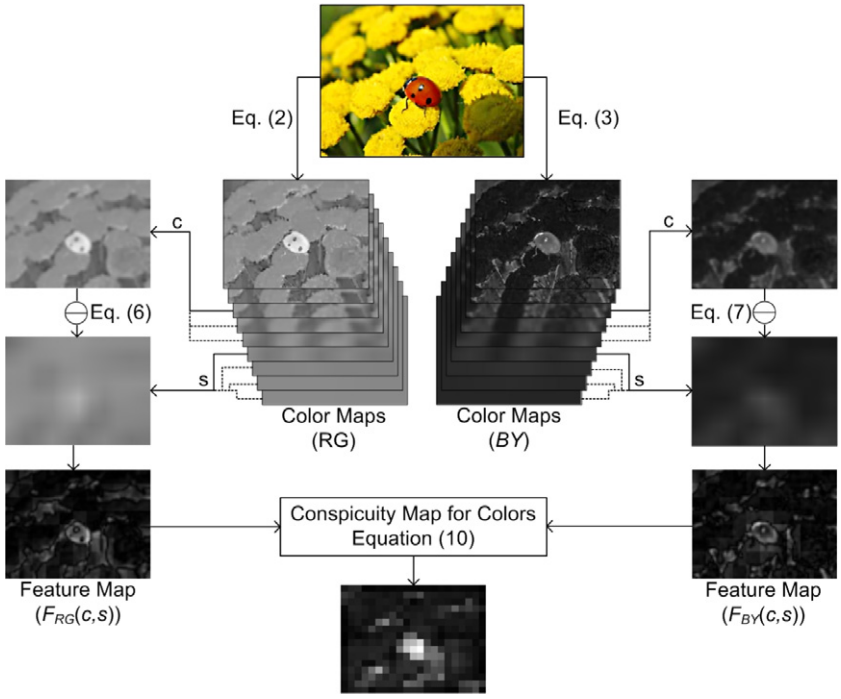

Fig. 3. (Color online) Flowchart for calculating the conspicuity map for colors (C_H).

Finally, the saliency map is computed by a linear combination of the conspicuity maps:

$$S^m = \frac{1}{3}\left[\mathcal{N}(C_I) + \mathcal{N}(C_H) + \mathcal{N}(C_R)\right]. \tag{13}$$

Normally, the saliency map is computed at scale four, which means a map size that is 1/16 of the input image size. The saliency map S^m is used to compute the object-saliency map described in Section 3.

2.2. Image segmentation

The scene segmentation model proposed in Wang and Terman (1997) is an extension of the LEGION model (Terman & Wang, 1995). The basic unit of LEGION is a relaxation oscillator defined as a feedback loop between an excitatory variable x_i and an inhibitory variable y_i (Terman & Wang, 1995):

$$\dot{x}_i = 3x_i - x_i^3 + 2 - y_i + I_i + S_i + \rho \tag{14a}$$

$$\dot{y}_i = \epsilon(\alpha(1 + \tanh(x_i/\beta)) - y_i) \tag{14b}$$

where I_i represents the external stimulation, S_i the input from neighboring oscillators in the network, and ρ denotes the amplitude of Gaussian noise. The parameter ϵ is a small positive number. If I_i is set to a constant and the terms S_i and ρ are removed, Eq. (14) becomes a typical relaxation oscillator (van der Pol, 1926). The noise term ρ not only serves to test the robustness of the model but also helps to segregate different input patterns (Terman & Wang, 1995).

Fig. 4 shows the nullclines and the trajectories of a single oscillator defined in Eq. (14), where the x-nullcline is a cubic function and the y-nullcline is a sigmoid function. If the total stimulation received by the oscillator, $I_i + S_i + \rho > 0$, the x and the y nullclines intersect at just one point at the middle branch of the cubic. In this case, the oscillator is said to be *enabled* and a stable cycle limit is observed (see Fig. 4(a)). The periodic orbit alternates between an *active phase* and a *silent phase*, which correspond to high and low x values, respectively (see Fig. 4(a)). The transition between the two phases occurs rapidly in comparison with the motion within each phase, thus referred to as *jumping*. The parameter α controls how much time the oscillator spends in these two phases. When the total input $I_i + S_i + \rho < 0$, the two nullclines of Eq. (14) intersect at a stable fixed point on the left branch of the cubic (see Fig. 4(b)). In this case, the oscillator does not produce a periodic orbit and no oscillation is observed. As the oscillator can be induced to oscillate by external stimulation, such a state is called *excitable*. The parameter β controls the steepness of the sigmoid which is normally set to a small value in order to make the sigmoid approach a step function (Terman & Wang, 1995).

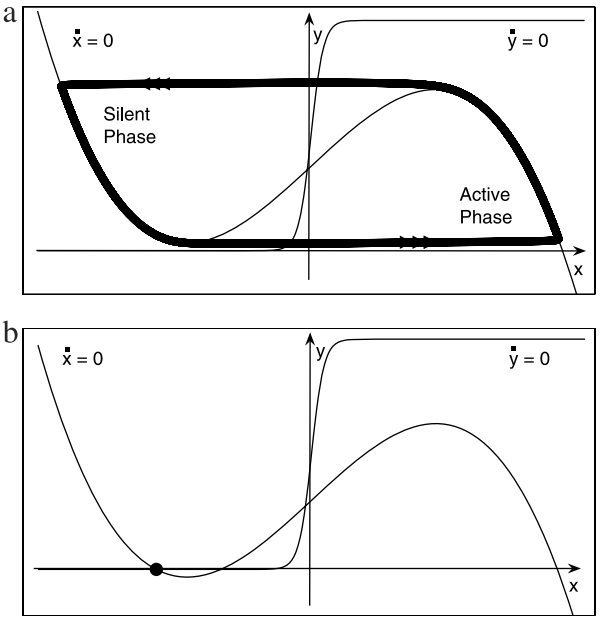

Fig. 4. Dynamics of a single relaxation oscillator. (a) Behavior of an enabled oscillator. A limit cycle trajectory is represented by a bold curve and the arrows indicate the motion direction. (b) Behavior of an excitable oscillator. In this case, a stable fixed point is observed indicated by the dot.

$\mathcal{I}_i$ represents the total external stimulation received by oscillator i. In the original LEGION model (Terman & Wang, 1995), $\mathcal{I}_i$ was a constant. To perform image segmentation on real images, a lateral potential term was later introduced to distinguish between major regions and noisy fragments (Wang & Terman, 1997). This mechanism can be explained as follows. If oscillator i lies in the center of a homogeneous image region, it is able to receive a large input from its neighbors; in this case it is defined as a *leader*. On the other hand, if it corresponds to an isolated fragment of the image, it does not receive a large input from its neighborhood and hence cannot become a leader. Based on this idea, only blocks which have at least one leader are allowed to oscillate.

To perform the segmentation task, a two-dimensional LEGION network is used. Here, the oscillators are typically connected with their eight immediate neighbors, except on the borders where no wraparound is applied.

For this network, the connection term S_i of Eq. (14a) is defined as follows:

$$S_i = \sum_{k \in N(i)} W_{ik} H(x_k - \theta_x) - W_z H(z - \theta_z) \tag{15}$$

where W_{ik} defines the dynamic connection weight from oscillator k to i and $N(i)$ represents a set of oscillators that comprises the neighborhood of i (Wang & Terman, 1997). H represents the Heaviside function defined as $H(v) = 1$ if $v \geq 0$ and $H(v) = 0$ otherwise. The dynamic connection weights W_{ik} are formed on the basis of the permanent connection weights following dynamic normalization which ensures that each oscillator gets equal weights from its neighbors. Dynamic weights are rapidly formed on the basis of the correlation between presynaptic and postsynaptic activity (for details see Wang and Terman (1997)). θ_x and θ_z are thresholds.

W_z in Eq. (15) defines the inhibition weight associated with the global inhibitor z. The dynamics of z is defined as

$$\dot{z} = \phi \left(\sum_k H(x_k - \theta_x) - z \right), \tag{16}$$

where ϕ is a parameter that controls how fast the global inhibitor reacts to the stimulation received from the oscillators. Note that z approaches the number of oscillators in the active phase, and will be used to represent the size of each synchronized oscillator block (segment).

Based on the LEGION dynamics described above, Wang and Terman (1997) developed a computer algorithm for image segmentation that follows the main aspects observed on the numerical simulations of the Eqs. (14)–(16). In particular, segmentation is the process of forming blocks of synchronized oscillators, each block corresponding to one segment. Here synchronization means simultaneous jumping to the active phase (see Fig. 4). Different blocks are desynchronized, i.e. they do not stay in the active phase at the same time. Detailed description of this algorithm can be found in Wang and Terman (1997).

3. Model description

In Fig. 1 we have shown a flowchart of our model that is composed of three modules: image segmentation, saliency map, and object selection. The computational flow can be described as follows. First, an input image feeds the image segmentation and saliency map modules. Second, the segmentation result and the saliency map generated by these modules are combined to build an object-saliency map that feeds the object selection module. Third, the object selection module selects the most salient object and suppresses all the others. Finally, the inhibition of return (IoR) mechanism is included in the object selection module that inhibits the previously selected object in order to allow the next most salient object to be selected. This process is repeated until all objects have been selected or when the input image is withdrawn. Fig. 5 shows the interaction between the segmentation and selection networks along with the object saliency map.

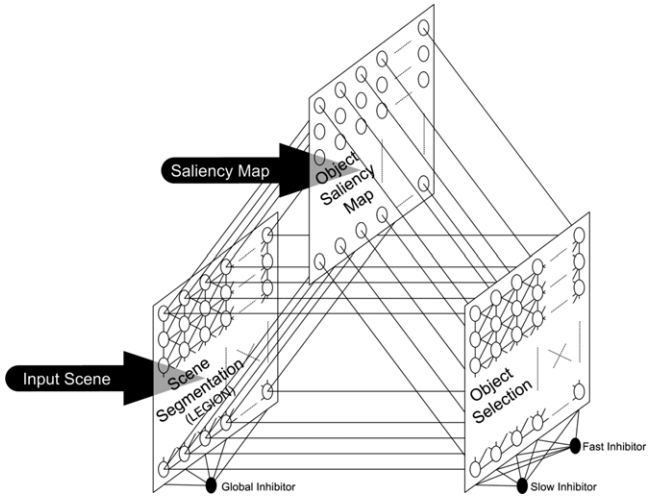

Fig. 5. Interaction between modules. Empty circles represent pixel locations in the object-saliency map, and oscillators in the segmentation and selection networks. The black circles indicate inhibitors: the global inhibitor (z) in the segmentation network and the slow (z_s) and the fast (z) inhibitors in the selection network. The connections between modules are one-to-one correspondence.

The following sections describe how the object-saliency map is created and how object selection works.

3.1. Object-saliency map

The object-saliency map, S^o, is responsible for providing the level of saliency of each object in the input scene. This map differs from the saliency map presented in Section 2.1 in that it represents the saliency of each object instead of each pixel. First, in order to create a one-to-one correspondence between the saliency map and the LEGION network, the saliency map is rescaled to the input image size by means of linear interpolation. After that, for each segment produced by the LEGION, its average saliency is calculated from all the corresponding points in S^m (Eq. (13)):

$$\bar{S}_i^o = \frac{\sum_{j \in O(i)} S_j^m}{|O(i)|}, \tag{17}$$

where $\bar{S}_i^o$ is the average saliency of the segment that contains pixel i; $O(i)$ is the set of all pixels grouped with pixel i in the same segment via oscillator synchronization; S_j^m is value of the saliency map at pixel j (Eq. (13)); and $|O(i)|$ is the size of $O(i)$. After calculating the saliency for all segments, the object size is incorporated into the saliency value by the following equation:

$$S_i^o = \bar{S}_i^o \sqrt[5]{\frac{|O(i)|}{|O_M|}}, \tag{18}$$

where the fifth-root function, chosen empirically, is used to moderate the saliency of relatively small segments. $|O_M|$ is the size of the largest segment in the input image. S^o defines the object-saliency map which is used as input to the object selection network.

As described above, to calculate the object-saliency map we utilize the results generated by the previous stages. Thus, the segmentation process must be concluded before selection can happen. As pointed in Wang and Terman (1997), the segmentation

module (LEGION) takes no longer than $M + 1$ cycles to segment the input image, where M is the number of major segments, or segments that contain at least one leader as described in Section 2.2. It is worth noting that the number of segments is unknown in advance. However, as mentioned before, in order to deal with real images containing large numbers of pixels, the segmentation process is performed by an efficient approximation algorithm proposed in Wang and Terman (1997). An interesting property of this algorithm is that the segmentation process is completed when every leader has jumped to the active phase once (see Section 2.2). In this way, we can generate the object-saliency map and perform visual selection after the segmentation process is completed.

3.2. Object selection

The object selection network is an extension of the LEGION model following the ideas developed in Wang (1999). The architecture of this network is shown in Fig. 5 in which a fast and a slow inhibitor are responsible for desynchronizing the objects and selecting the one of them, respectively.

This network follows the dynamics described in Section 2.2. The main differences between our network for object selection and LEGION for image segmentation are the presence of the slow inhibitor, the introduction of the IoR mechanisms, and how the external stimulation is defined.

In our selection network, each oscillator is connected to its eight nearest neighbors as follows. If two neighboring oscillators have their corresponding oscillators in the segmentation network synchronized, they are connected. On the other hand, if the corresponding oscillators in the segmentation network do not belong to the same object (i.e. desynchronized), the connection between the two oscillators in the object selection module is set to zero. Such connectivity can be readily set up using dynamic weights that quickly increase their strengths when presynaptic and postsynaptic oscillators are both active (Terman & Wang, 1995; von der Malsburg, 1981). Thus, the objects formed in the LEGION are directly transported to the object selection network.

The external stimulation $\mathcal{I}_i$ is defined as follows:

$$\mathcal{I}_i = V_i H(S_i^o - Cz_s) H(r_i - \theta_z), \tag{19}$$

where V_i is set to a high value if the corresponding oscillator i in the segmentation module is enabled. Otherwise, V_i is set to a low value. In this way, oscillators in the object selection network corresponding to a segment in LEGION assume high values of V, whereas oscillators representing noisy fragments (the background) have a low V value. S_i^o is the object-saliency value from Eq. (18). C is a parameter that controls the number of objects that can be selected at a time (Wang, 1999). z_s models the slow inhibitor and r_i represents the IoR component.

The dynamics of the slow inhibitor is defined as

$$\dot{z}_s = \psi \left[\sum_k \frac{S_k^o H(x_k - \theta_x)}{|O(k)|} - z_s \right]^+ - \mu \epsilon z_s \tag{20}$$

where the function $[v]^+ = v$ if $v \geq 0$ and 0 otherwise. The parameters ψ and μ are on the order of 1. The slow inhibitor is characterized by a fast rise and a slow decay owing to the small value of the relaxation parameter ϵ in the second term. The selection process is produced by the Heaviside function and the slow inhibitor which allows to become active just the oscillators with $S_i^o \geq Cz_s$.Thus, by setting a proper value of C as defined in Wang (1999), only the object with the highest value of S^o is allowed to oscillate, i.e. to be selected.

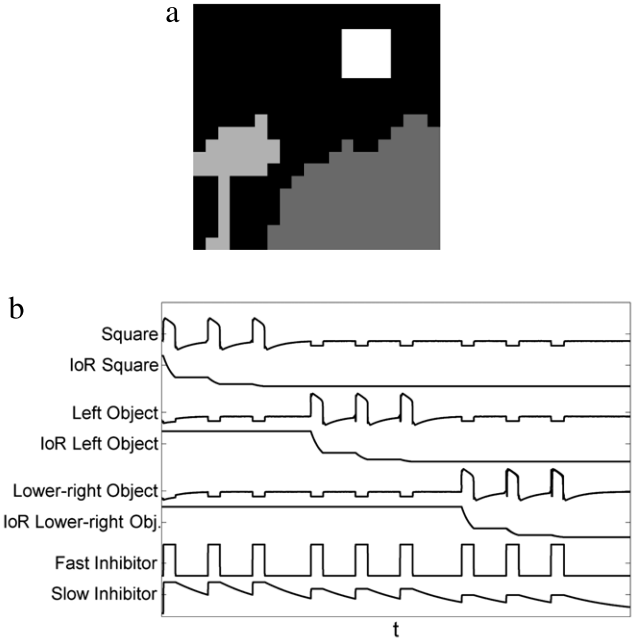

Fig. 6. Illustration of the object selection process. The selection network is integrated using the fourth-order Runge–Kutta method. (a) Object-saliency map showing three objects: a square, a left object and a lower-right object. (b) Activity of each oscillator block and its corresponding IoR, plus the activity traces of the fast and the slow inhibitor.

The variable r_i in Eq. (19) models the IoR component of each oscillator described by the following equation:

$$\dot{r}_i = -\omega r_i H(x_i - \theta_x). \tag{21}$$

Initially, for each oscillator i, r_i is set to 1. Every time an oscillator jumps to the active phase, its r_i value is reduced following Eq. (21). After a number of cycles controlled by parameter ω, r_i approaches zero. Thus, the second Heaviside function of Eq. (19) returns zero and the oscillator is inhibited. Due to the presence of the IoR, the selection network is allowed to select the next most salient object, which resembles attentional shifts in visual perception (Itti & Koch, 2001a).

The object-saliency value is also used to set the initial state of each oscillator. Once we have the saliency of all the objects, we can use these values to determine which object oscillates so as to avoid the time-consuming competition for selection. To achieve this behavior, the initial value of y_i (Eq. (14b)) is set according to its object-saliency value in the following way:

$$y_i = 2\alpha(1 - S_i^o) + V_i. \tag{22}$$

Based on Eq. (22), the oscillators of the selection network representing the object with the highest saliency have their initial y_i values set in the silent phase close to the left knee of the cubic nullcline and the oscillators with low saliency far from the left knee in the silent phase (see Fig. 4). In the special case where two or more objects have the same object saliency, the selection network chooses all of them, which will oscillate desynchronously until they are inhibited by the IoR.

Fig. 6 shows an illustration of the selection process performed by the object selection network. Consider Fig. 6(a) to be an object-saliency map described in Section 3.2. This map feeds the object selection network. There, the square object, corresponding to the brightest region, represents the most salient object while the lower-right object, the darkest one, represents the least salient object. The saliency value of each object serves two functions. First, it is used as input in Eq. (19) to decide which object is allowed to pulse. Second, it defines the initial values of y_i in Eq. (22). As

we can see in Fig. 6(b), the square is the first to be selected while the others remain silent. As the oscillators representing the square keep pulsing the IoR takes effect, and after some time determined by ω (Eq. (20)), the oscillators are inhibited allowing the next most salient object to become selected, in this example, the left object. This process continues until all the objects have been selected once.

The overall behavior of our model can be understood as follows. An input image feeds the saliency map and the segmentation module as illustrated in Fig. 1. The saliency map calculates the saliency of all pixels. This process incorporates the role of local feature contrast in guiding attention. In parallel, the LEGION segregates the input image into a set of segments. The LEGION network is able to achieve rapid synchronization among oscillators activated by the same object and desynchronization between different blocks of oscillators representing different segments (Terman & Wang, 1995; Wang & Terman, 1997). After obtaining the saliency map and the segmentation result, the object saliency map is generated. Eq. (18) incorporates the size of an object into the object-saliency map. This map feeds the object selection module, which becomes the original LEGION model if we eliminate the two Heaviside functions in Eq. (19) (Terman & Wang, 1995). In this equation, the first Heaviside plays the role of object selection and the second the IoR. If the first Heaviside returns 0, i.e. the object saliency value that feeds the oscillator does not exceed the level of the slow inhibitor, the oscillator is excitable and can be recruited to oscillate by one of its neighbors based on the term S_i in Eq. (14a). However, considering that the oscillators within a block are not connected to oscillators from another block, and the object saliency value for the whole block is the same, if the first Heaviside of an oscillator is 0, the Heaviside of the whole block is also zero. Thus, the object is inhibited. On the other hand, if the object saliency value that feeds a block of oscillators exceeds slow inhibition, the oscillators are allowed to oscillate and the object represented by them is selected. At the same time, the slow inhibitor assumes a new value through Eq. (20) which represents the object saliency of the currently active segment. As a result, other objects with smaller object saliency values are prevented from being selected.

Once a block is oscillating, the IoR mechanism takes effect and each oscillator i within that block has its r_i reduced by Eq. (21). After a few cycles defined by ω, r_i approaches zero. Thus, the second Heaviside of Eq. (19) returns 0, which represents the inhibition of oscillator i and consequently the inhibition of the whole segment. Following the inhibition of this object, the slow inhibitor has its value decreased by Eq. (20) and the next most salient object is selected as shown in Fig. 6.

4. Simulation results

In this section, computer simulation results are presented. Before presenting the results, we first describe the parameters used in the modules. In the saliency map module (Section 2.1), we apply the same parameter values used in Itti et al. (1998), except for the definitions of the color opponencies and the Gaussian kernel as mentioned in Section 2.1. Image segmentation is performed by the algorithm presented in Wang and Terman (1997). In this algorithm, the coupling strength W_{ij} between two neighbor oscillators is set up according to their similarity using the following rule. For gray level images,

$$W_{ij} = I_M/(1 + |I_i - I_j|). \tag{23}$$

For color images,

$$W_{ij} = I_M/\left(1 + \sum_{h\in\{r,g,b\}} |h_i - h_j|\right) \tag{24}$$

where I_M is the maximum value of the channels I, r, g, and b. In our simulations, this value is set to 255. I_i is the gray level of pixel i. h_i represents the color channel (r, g, and b) of a color pixel i. The parameter W_z in Eq. (15) defines the strength of the global inhibitor. When W_z is set to a high value, it is more difficult to group pixels into a single object, which consequently leads to more and smaller regions. In a way, W_z provides a control on the scale of analysis which is not addressed in this study. W_z is adjusted for each input image in order to produce a reasonable segmentation result (Wang & Terman, 1997) and its value will be given when describing the simulations.

The object selection network presented in Section 3.2 is integrated by using the fast numerical method of singular limit which allows for simulating large networks of relaxation oscillators (Linsay & Wang, 1998). The following parameter values are used for integrating the selection network by the singular limit method: $\alpha = 6.5$, $W_z = 0.7$, and $\mu = 0.125$. All the other parameters are not necessary when solving the equations using this method. $C = 1.65$ is used for all the experiments. Note that selection results are not very sensitive to these parameter values.

First, two gray level images are used as an input. Fig. 7(a) shows the first input figure. Fig. 7(b) presents the saliency map from Fig. 7(a) where brighter pixels indicate higher saliency points. Here, by using $W_z = 20$ the LEGION network produces 17 segments as shown in Fig. 7(c). Based on the results from the saliency map (Fig. 7(b)) and LEGION (Fig. 7(c)), the object-saliency map is shown in Fig. 7(d). In this figure, a brighter object indicates a higher saliency one. This map feeds the object selection network which first chooses the most salient object shown in Fig. 7(e), representing a lake in the central part of the scene. After that, due to the IoR mechanism described in Section 3.2, the oscillators representing the first selected object are inhibited allowing the system to select the second most salient object which is shown in Fig. 7(f). In all the simulations presented in this paper, only the first and the second selected objects are shown to illustrate the selection process. The next simulation, presented in Fig. 8, is performed on an MRI (magnetic resonance imaging) image of the human head. As in Figs. 7, 8(a) shows the input image and Fig. 8(b) the saliency map. For this image, $W_z = 20$ and the LEGION network produces 21 segments as shown in Fig. 8(c). From the object-saliency map in Fig. 8(d), one can see that the cortex is the most salient object, thus, the first object to be selected as presented in Fig. 8(e). The second object selected by the network is shown in Fig. 8(f), corresponding to the brainstem.

Next, we present results on color images in Figs. 9–12, following the same format as in Figs. 7 and 8. For all of them, $W_z = 20$. In Fig. 9(a), due to the high contrast of the beetle with its background composed of mostly yellow and green things, the beetle seems to be the first object to pop out from the scene for a human observer. This percept agrees with the result from our object-saliency map in Fig. 9(d), where the segment corresponding to the beetle is the brightest. As we can see in Fig. 9(e), the first object to be selected is indeed the beetle.

Fig. 10 presents a simulation of a scene where the most salient object appears to be a boat to a human observer. Again, due to its high contrast with background objects, the boat is selected by our system as the first object (see Fig. 10(e)). Part of an orange tree is shown in Fig. 11(a). For this input image, our model selects the two oranges as the first and the second object emerging from the competition, and the selected objects are shown in Fig. 11(e) and (f), respectively. Fig. 12 presents a scene of a person in Central Park, New York. For this color image, the first selected object is the upper body of the person shown in Fig. 12(e) and the second selected object corresponds to the left part of the park scene shown in Fig. 12(f).

Other simulations with gray and color images have been conducted, and results with similar quality to that of the above simulation results have been obtained.

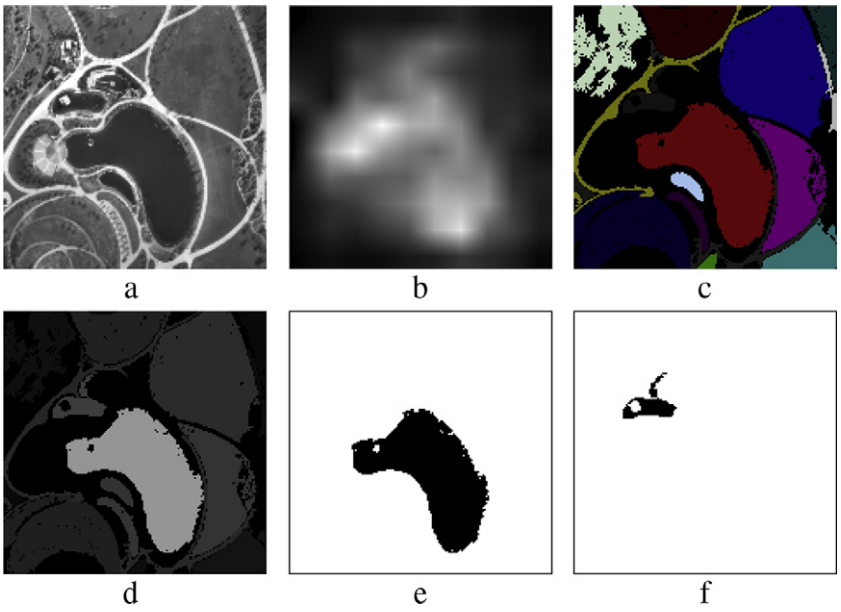

Fig. 7. (Color online) Object selection result for a gray level image. (a) Input image which is an aerial image with 160 × 160 pixels. (b) Saliency map. (c) Result of LEGION segmentation, where each segment is represented by a distinct color. (d) Object-saliency map. (e) First object selected. (f) Second object selected.

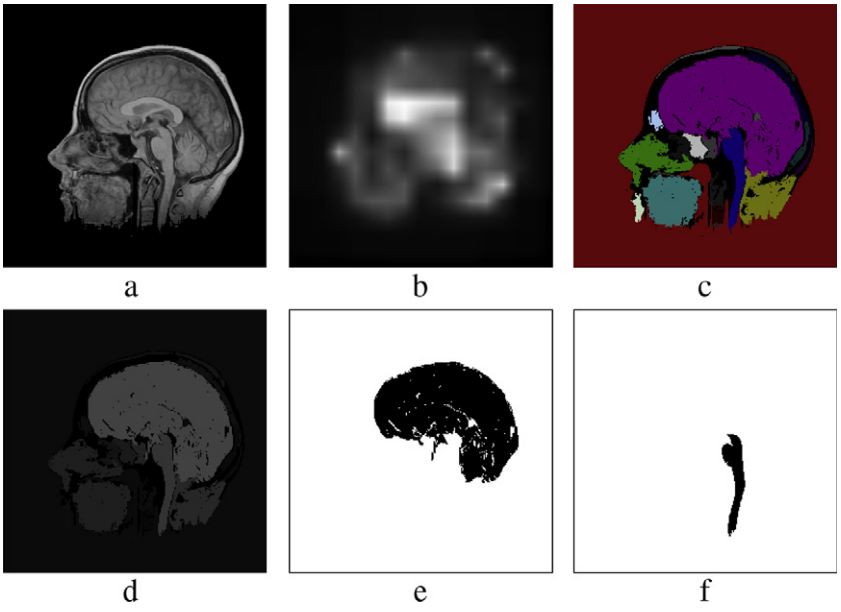

Fig. 8. (Color online) Object selection result for a gray level image. (a) The input image which is an MRI image with 257 × 257 pixels. (b) Saliency map. (c) Result of LEGION segmentation. (d) Object-saliency map. (e) First object selected. (f) Second object selected.

5. Concluding remarks

Object-based attention has received empirical support (Martinez et al., 2007; O'Craven et al., 1999; Richard et al., 2008; Roelfsema et al., 1998; Shinn-Cunningham, 2008; Wang et al., 2005). In this paper, we have presented a novel object selection model based on oscillatory correlation theory. This model integrates several modules. A saliency map, which calculates the saliency values of all the locations of the input scene, a LEGION network for segmenting the scene into a set of segments or objects, and an object selection network for selecting the most salient object of the scene. Modeling visual attention with an oscillator network is motivated by physiological studies suggesting that synchronous activity plays a fundamental role in solving the binding problem and visual attention (Fries, Reynolds, Rorie, & Desimone, 2001; Jermakowicz & Casagrande, 2007; Singer & Gray, 1995). In contrast to previous computational models of location-based visual attention, our model, due to the use of an image segmentation network, is able to deal with objects directly. By integrating the saliency map, the segmentation module, and the IoR mechanism, our selection network can select a set of objects sequentially according to their saliency. The selection of objects based on their intrinsic saliency proposed here also contrasts to other recent oscillatory models for selection (Borisyuk et al., 2009; Chik et al., 2009).

Our model has several limitations that need be addressed in future work. The proposed system only addresses bottom-up aspects of attentional selection, and top-down guidance of attention is not modeled. Top-down analysis could be modeled by including a working memory and an associative memory, as investigated in previous work (Borisyuk & Kazanovich, 2004; Wang & Liu, 2002). Incorporation of other visual features, such as motion

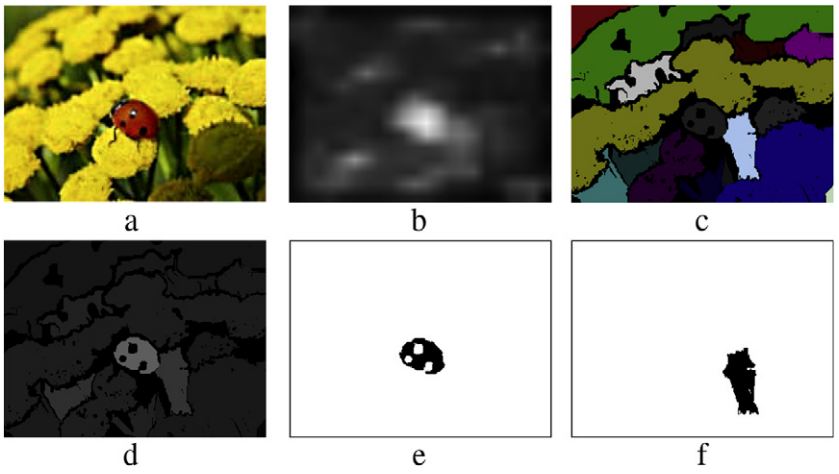

Fig. 9. (Color online) Object selection result for a color image. (a) Input image with 351 × 256 pixels. (b) Saliency map. (c) Result of LEGION segmentation. (d) Object-saliency map. (e) First object selected. (f) Second object selected.

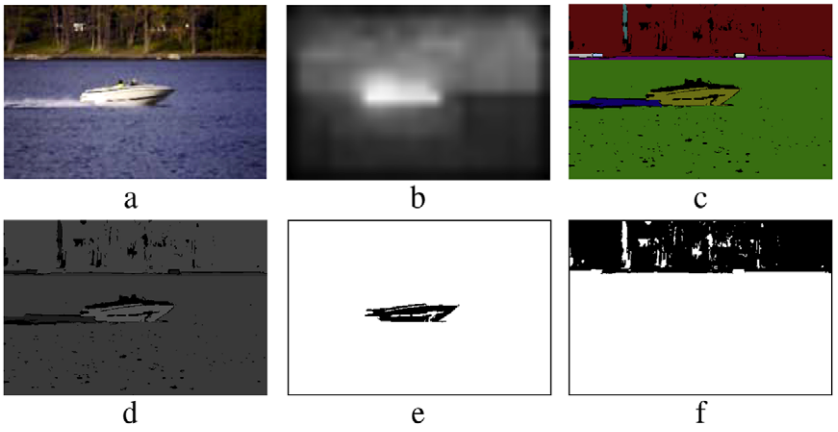

Fig. 10. (Color online) Object selection result for a color image. (a) Input image with 385 × 256 pixels. (b) Saliency map. (c) Result of LEGION segmentation. (d) Object-saliency map. (e) First object selected. (f) Second object selected.

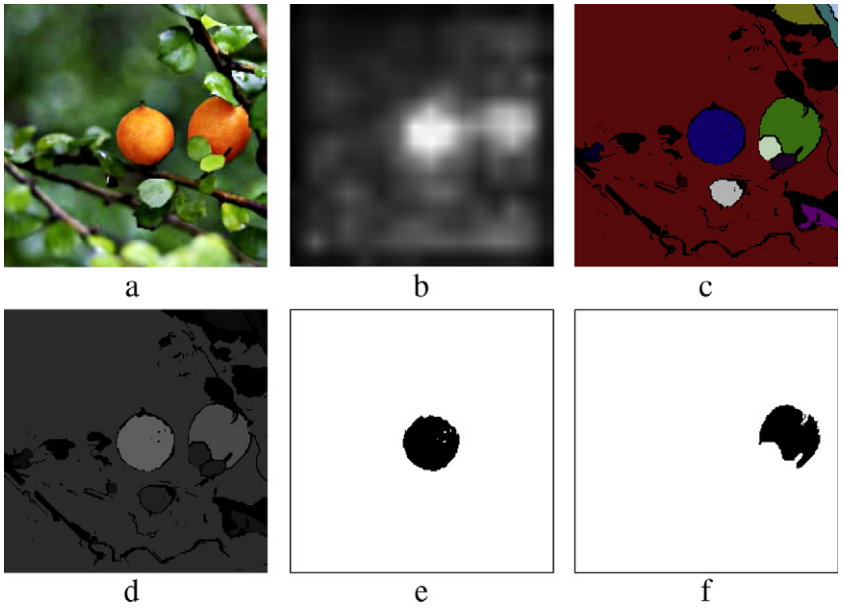

Fig. 11. (Color online) Object selection result for a color image. (a) Input image with 256 × 256 pixels. (b) Saliency map. (c) Result of LEGION segmentation. (d) Object-saliency map. (e) First object selected. (f) Second object selected.

and object contour, among others, could further enhance the performance of the system (see Wang, 2005). Finally, it should also be stated that even though the architecture of our model is motivated by experimental studies of visual attention, our model

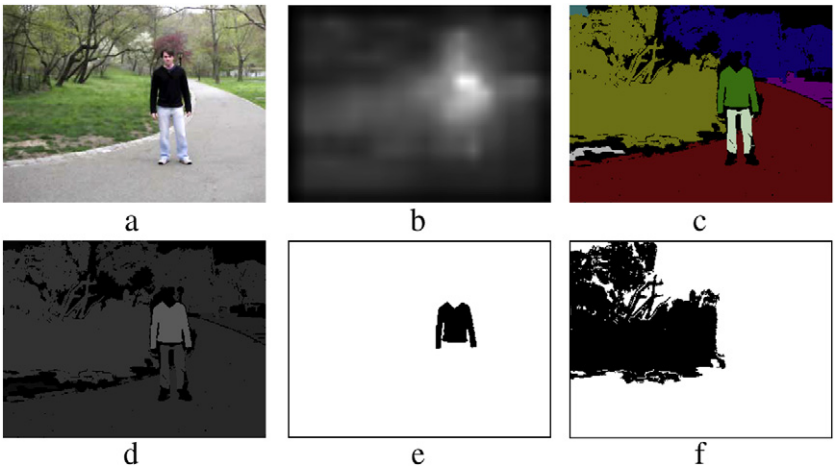

Fig. 12. (Color online) Object selection result for a color image. (a) Input image with 341 × 256 pixels. (b) Saliency map. (c) Result of LEGION segmentation. (d) Object-saliency map. (e) First object selected. (f) Second object selected.

does not simulate psychophysical data in a quantitative way as its purpose is to perform selection of objects in real scenes. From the psychological standpoint, many aspects of the model are gross simplifications. For example, our model does not allow an object to be selected more than once. Also, the time course of shifting from one object to another is not addressed although there is potential consistency between gamma-band oscillations (about 40-Hz) (Singer & Gray, 1995) and the rate of attentional shifts of about 50–100 ms (Pashler, 1998; Saarinen & Julesz, 1991) (see Fig. 6). Neurocomputational models have been developed to simulate perceptual data of visual attention (see Corchs & Deco, 2001, among others).

Acknowledgements

This work was undertaken while M.G.Q. was a visiting scholar in the Perception and Neurodynamics Lab at The Ohio State University. M.G.Q. was supported by the So Paulo State Research Foundation (FAPESP). D.L.W. was supported in part by an NGI University Research Initiatives grant and the K.C. Wong Education Foundation (Hong Kong). L.Z. and R.A.F.R. were supported by the Brazilian National Research Council (CNPq).

References

Arbib, M. A. (Ed.) (2003). *Handbook of brain theory and neural networks* (2nd ed.). Cambridge, MA: MIT Press.
Borisyuk, R., & Kazanovich, Y. (2004). Oscillatory model of attention-guided object selection and novelty detection. *Neural Networks*, *17*, 899–915.
Borisyuk, R., Kazanovich, Y., Chik, D., Tikhanoff, V., & Cangelosi, A. (2009). A neural model of selective attention and object segmentation in the visual scene: an approach based on partial synchronization and star-like architecture of connections. *Neural Networks*, *22*, 707–719.
Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, *31*, 532–540.
Cheng, H. D., Jiang, X. H., Sun, Y., & Wang, J. (2001). Color image segmentation: advances and prospects. *Pattern Recognition*, *34*, 2259–2281.
Chik, D., Borisyuk, R., & Kazanovich, Y. (2009). Selective attention model with spiking elements. *Neural Networks*, *22*, 890–900.
Corchs, S., & Deco, G. (2001). A neurodynamical model for selective visual attention using oscillators. *Neural Networks*, *14*, 981–990.
Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.
Egeth, H. E., & Yantis, S. (1997). Visual attention: control, representation, and time course. *Annual Review of Psychology*, *48*, 269–297.
Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, *291*, 1560–1563.
Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing* (2nd ed.). New Jersey: Prentice Hall.
Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, *391*, 481–484.
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.
Itti, L., & Koch, C. (2001a). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*, 194–203.
Itti, L., & Koch, C. (2001b). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*, 161–169.
Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 1254–1259.
Jermakowicz, W. J., & Casagrande, V. A. (2007). Neural networks a century after Cajal. *Brain Research Reviews*, *55*, 264–284.
Kazanovich, Y., & Borisyuk, R. (2002). Object selection by an oscillatory neural network. *BioSystems*, *67*, 103–111.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.
Linsay, P. S., & Wang, D. L. (1998). Fast numerical integration of relaxation oscillator networks based on singular limit solution. *IEEE Transactions on Neural Networks*, *9*, 523–532.
Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.
Martinez, A., Ramanathan, D., Foxe, J., Javitt, D., & Hillyard, S. (2007). The role of spatial attention in the selection of real and illusory objects. *The Journal of Neuroscience*, *27*, 7963–7973.
Niebur, E., Koch, C., & Rosin, C. (1993). An oscillation-based model for the neuronal basis of attention. *Vision Research*, *33*, 2789–2802.
O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, *401*, 584–587.
Pashler, H. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.
Revounsuo, A., & Newman, J. (1999). Binding and consciousness. *Consciousness and Cognition*, *8*, 123–127.
Richard, A. M., Lee, H., & Vecera, S. P. (2008). Attentional spreading in object-based attention. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 842–853.
Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*, 376–381.
Saarinen, J., & Julesz, B. (1991). The speed of attentional shifts in the visual field. *Proceedings of the National Academy of Sciences of the United States of America*, *88*, 1812–1814.
Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*, 182–186.
Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, *18*, 555–586.
Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence*, *146*, 77–123.
Terman, D., & Wang, D. L. (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D*, *81*, 148–176.
Tiesinga, P. H. E. (2005). Stimulus competition by inhibitory interference. *Neural Computation*, *17*, 2421–2453.
van der Pol, B. (1926). On 'relaxation oscillations'. *Philosophical Magazine*, *2*, 978–992.
von der Malsburg, C. (1981). The correlation theory of brain function. *Technical report internal report 81-2*. Max-Planck institute for biophysical chemistry. Göttingen, Germany.
von der Malsburg, C., & Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics*, *54*, 29–40.
Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407.
Wang, D. L. (1999). Object selection based on oscillatory correlation. *Neural Networks*, *12*, 579–592.
Wang, D. L. (2005). The time dimension for scene analysis. *IEEE Transactions on Neural Networks*, *16*, 1401–1426.

Wang, D. L., Kristjansson, A., & Nakayama, K. (2005). Efficient visual search without top–down or bottom–up guidance. *Perception & Psychophysics*, *67*, 239–253.
Wang, D. L., & Liu, X. (2002). Scene analysis by integrating primitive segmentation and associative memory. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics), 32*, 254–268.
Wang, D. L., & Terman, D. (1997). Image segmentation based on oscillatory correlation. *Neural Computation*, *9*, 805–836.
Yantis, S. (1998). *Control of visual attention*, *Attention* (pp. 223–256). London: Psychology Press, (Chapter).
Yantis, S. (2000). *Goal-directed and stimulus-driven determinants of attentional control*: *Vol. 18*. *Attention and performance xviii* (pp. 73–103). Cambridge: MIT Press, (Chapter).
Yu, G., & Slotine, J.-J. (2009). Visual grouping by neural oscillator networks. *IEEE Transactions on Neural Networks*, *20*, 1871–1884.