An Object-Based Visual Selection Model with Bottom-up and Top-down Modulations

Alcides X. Benicasa
Department of Information Systems
Federal University of Sergipe
Itabaiana, Brazil
Email: alcides@ufs.icmc.usp.br

Marcos G. Quiles
Department of Science and Technology
Federal University of São Paulo
São Paulo, Brazil
Email: quiles@unifesp.br

Liang Zhao, Roseli A.F. Romero
Department of Computer Science
University of São Paulo
São Carlos, Brazil
Email: zhao.rafrance@icmc.usp.br

Abstract—Research in qualitative models of visual attention has mainly focused on the bottom-up guidance of early visual features. Here we propose a new model which combines both bottom-up and top-down modulation into the visual selection model. The proposed model is composed of five main components: a Visual Feature Extraction module, a LEGION network for image segmentation to deal with objects, a Multi-Layer Perceptron (MLP) network for object recognition; a Kohonen’s Self-Organizing Maps (SOM) combined with a network of integrate and fire neurons which creates our attribute-saliency map, and finally, an object selection module which highlights the most salient object in the scene. Experiments with synthetic and real images is conducted and the results demonstrate the effectiveness of the proposed approach.

Index Terms—bottom-up and top-down visual attention; image segmentation; recognition of objects;

I. INTRODUCTION

Every day we are faced with complex scenes, estimated to be on the order of $10^7 - 10^8$ bits per second at the optical nerve. Our visual system needs to be able to analyze and understand a large amount of visual information, excluding everything that is not important. In order to do so, our visual system must delivery attention to specific targets while ignoring others.

But what is a specific target? There are two distinct components that drive the human visual attention: first the top-down attention, where voluntary visual control guides the attention for specific features or known objects in the image, and second, the bottom-up attention, where the attention is involuntarily guided by visual features, such as colors, deep, etc. [6].

But how can we attend to objects before we recognize them? In bottom-up models, such as [1], [7], [5], [10], there is no type of labeling, since only the primitive features of the image are considered to identify the salient point or region, which is directly related to the unsupervised learning, whose goal is to find groups of similar objects. On the other hand, models implementing top-down mechanism, i.e. [9], [13], the objects to be identified in the image must be labeled previously, characteristic of the supervised learning, since the search occurs from prior information about the specific object. Top-down attention modulates competition between visual stimuli. This process involves a concept of working memory, where some information, such as the search for a particular object or location that is kept in memory driven the selection to a specific object. It can also influence the response of bottom-up clues [3].

Here, we propose a new object-based visual selection model with both bottom-up and top-down modulations. Our model is composed by the following modules: a Visual Feature Extraction module, which is responsible for extracting the early visual features, such as colors, orientation, etc.; a Locally Excitatory Globally Inhibitory Oscillator Networks (LEGION) network for image segmentation. It is worth noting that a segmentation module is mandatory when dealing with objects and real scenes; a Multi-Layer Perceptron (MLP) network for object recognition; a Kohonen’s Self-Organizing Maps (SOM) combined with a network of integrate and fire neurons which creates our attribute-saliency map, and finally, an object selection module which, based on the guidance from the attribute-saliency map, selects the most salient object in the scene.

This paper is organized as follows. In Section 2, a brief review of the early visual features extraction and the segmentation models presented. Section 3 presents the proposed model. Computer simulations are presented in Section 4. Finally, Section 5 draw some conclusions about this work.

II. BACKGROUND

In this section we review the features combination strategies and the segmentation mechanism used in our visual selection model.

A. Extraction of Early Visual Features

For starting the process of visual attention, it is necessary to extract the primitive information of the entrance image. According to [6], the first processing stage in any model of bottom-up attention is the computation of early visual features, where neurons at the earliest stages are tuned to simple visual attributes. We use in this work a simple features map to detect local spatial discontinuities in intensity contrast, color, orientation, location, and recognition, and are combined into a unique self-organizing map, presented in the following sections.

According to [6], [10], the saliency map is produced initially by a set of maps representing primary features, such as inten-
sity, 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, 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. We consider in this work only the conspicuity maps. The Figure 1 shows the conspicuity maps created from the model described.

Formally and according to [14], a input image $I$ is subsampled into a dyadic Gaussian pyramid by convolution with a linearly separable Gaussian filter and decimation by a factor of two. This process is repeated to obtain the next levels $\sigma = [0, \ldots, 8]$ of the pyramid. According to [2], the encoding process is equivalent to sampling the image with Laplacian operators of many scales. Thus, the code tends to enhance salient image features.

The intensity contrast is a spatial difference in light intensity in an image [6]. The intensity map $I$ for each level $(i)$ of the pyramid $I$ is defined by the following equation:

$$I_i = \frac{r_i + g_i + b_i}{3}$$

(1)

this operation is repeated for each level of the input pyramid to obtain an intensity pyramid with levels $i$.

The maps of spatial difference in colors, computed in the brain using red/green and blue/yellow, to each level of the image pyramid is furthermore decomposed into maps for red/green ($RG_i$) and blue/yellow ($BY_i$) opponencies:

$$RG_i = \frac{r_i - g_i}{\max(r_i, g_i, b_i)}$$

(2)

and

$$BY_i = \frac{b_i - \min(r_i, g_i)}{\max(r_i, g_i, b_i)}$$

(3)

Local orientation maps, $O_{io}$, are extracted by convolving $I_i$ with oriented Gabor filters for four orientations $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$, defined by the following equations:

$$O_{io} = |I_i \ast G_0(\theta)| + |I_i \ast G_{\pi/2}(\theta)|$$

(4)

After extracting the intensity $I_i$, color $RG_i$ and $BY_i$, and orientation maps $O_{io}$, feature maps are extracted by across-scale subtractions $\odot$ between different levels of the same feature. Next these maps are combined to form the conspicuity maps (Figure 1). For more details see [8], [15].

The maps presented are constructed from primitive image information. Our work presents an important characteristic related to a specific conspicuity map, the conspicuity map that it will represent the known object in the image. In order to delivery attention to object instead of isolated pixels, a segmentation mechanism is necessary [10]. The segmentation process transform the image into a set of segments which can be interpreted as primitive objects of the scene. Next section reviews the segmentation method used in this work.

### B. Image Segmentation

The scene segmentation is performed by the LEGION proposed in [16]. The basic unit of LEGION is a van der Pol oscillator defined as a feedback loop between an excitatory variable $x_i$ and an inhibitory variable $y_i$ [12], [11]:

$$\frac{dx_i}{dt} = 3x_i - x_i^3 + 2 - y_i + I_x + S_i + \rho$$

(5a)

$$\frac{dy_i}{dt} = \epsilon(\alpha(1 + \tanh(x_i/\beta)) - y_i)$$

(5b)

where $I_x$ is the external stimulation, $S_i$ defines the coupling term from neighboring oscillators and $\rho$ denotes the amplitude of Gaussian noise. $\epsilon$, $\alpha$ and $\beta$ are parameters of the oscillator.

The nullclines and the trajectories of the LEGION oscillator (Eq. (5)) is depicted in Figure 2, in which the $x$-nullcline is a cubic function and the $y$-nullcline is a sigmoid function. When the oscillator receives an external input greater the zero $I_x + S_i + \rho > 0$, it is possible to observe that the $x$ and the $y$ nullclines intersect at just one point. In this situation, the oscillator is said to be enabled and a cycle limit dynamic is observed as presented in Fig. 2(a). On the other hand, if the total external stimulation is lower than zero, the nullclines of Eq. (5) intersect at a stable fixed point (see Fig. 2(b)). Here, no periodic oscillations are observed albeit the oscillator can be induced to spike by neighbor oscillators.

The LEGION proposed in [11] is able to perform segmentation only in toy scenes. In order to perform image segmentation on real images, a lateral potential term was introduced in [16]. This term is used to discriminate between major segments and noisy segments of the scene. Generally speaking, this mechanism can be explained as follows. If oscillator $i$ is located in the center of a segment, i.e. an...
homogeneous region of the image, it is able to receive a large input from its neighbors, and is defined as a leader of the segment. On the contrary, if oscillator \( i \) represents an isolated pixel, it does not receive inputs from its neighbor and cannot become a leader. Thus, the segmentation process is conducted in order to permit that only segments with at least one leader are allowed to oscillate.

Oscillators in the network are organized in such a way that each oscillator represent an image pixel. Moreover, connections between oscillator are defined in a 8-connected neighborhood, except on border where no wraparound mechanism is taken into account.

Thus, the connection term \( S_i \) of Eq. (5a) is defined as follows:

\[
S_i = \sum_{k \in \Delta_i} W_{ik} H(x_k - \theta_x) - W_z H(z - \theta_z) \tag{6}
\]

where \( W_{ik} \) defines the connection weight from oscillator \( k \) to \( i \) and is defined by taking similarity between pixels \( i \) and \( k \). \( \Delta_i \) defines neighborhood of oscillator \( i \). \( H \) is the Heaviside function defined as \( H(k) = 1 \) if \( k \geq 0 \) and \( H(k) = 0 \) otherwise. When the Heaviside function return 1, it means that the neurons is pulsing. \( \theta_x \) and \( \theta_z \) are thresholds. The constant \( W_z \) in Eq. (6) defines the inhibition strength of the global inhibitor \( z \). The dynamics of \( z \) is modelled by the following equation:

\[
\frac{dz}{dt} = \phi \left( \sum_k H(x_k - \theta_x) - z \right) \tag{7}
\]

where \( \phi \) is a constant that defines how fast the global inhibitor responds to the stimulation received from the network. Wang & Terman [16], by analysing the LEGION dynamics explained above, proposed a computer algorithm that follows the major aspects of the numerical simulation of Equations (5)-(7). The algorithm perform segmentation by forming blocks of synchronized oscillators, each block corresponding to one segment. By synchronization we assume simultaneous jumping to the active phase (see Fig. 2). Segments are represented by oscillators pulsing in unison and the segmentation is achieved by the desynchronization between different blocks os synchronized oscillators (pulsing with different phases). A complete description of Wangs & Terman algorithm can be found in [16].

III. Model Description

The proposed approach to select salient objects is composed by the following modules: a Visual Feature Extraction module, a LEGION network for image segmentation, a Multi-Layer Perceptron (MLP) network for object recognition, a Kohonen’s Self-Organizing Maps (SOM) combined with a network of integrate and fire neurons which creates our attribute-saliency map, and finally, and object selection module which highlights the most salient object in the scene.

Figure 3 depicts a flowchart of our model. The computational flow can be described as follows. First, the scene is presented to the module responsible for extracting the early visual features of the scene and to the LEGION segmentation network. To implement both modules, we use the models described in Section II-A and II-B, respectively. The output from those modules feed the following modules: the MLP network for object recognition and the SOM network which creates the attribute-saliency map. These two last modules are presented in the following sections.

A. Object Recognition Module (MLP)

As described in previous sections, the our model takes both bottom-up and top-down modulations into account. Early
visual attributes, i.e. color contrast, etc, defines the bottom-up signal. On the other hand, information about previously memorized objects (top-down modulation) is responsible for guiding the selection process. Thus, in order to apply our model to select the salient objects of a given scene, the MLP network must be trained with a set of objects representing desired targets of the scene.

The set of objects used to train the MLP object recognition network is composed of a set of binary images extracted from the scene. In order to do so, the input scene is presented to the LEGION network which delivers a set of segmented objects. Those objects are manually labeled and used to set up the weight of the MLP network through the backpropagation algorithm. Figure 4 shows a couple of objects that can be recognized by the MLP network, and thus modulate the object selection process.

It is worth noting that in order to deliver attention to novel objects, a new training phase with those objects must be conducted.

Fig. 4. Samples of objects for training the object recognition module.

After the training process, the MLP is able to recognize a set of segments (objects). Thus, the overall dynamics of the system can be understood as follows. Every time a segment is highlighted into the LEGION network, or the segment is pulsing, it is directly presented to the MLP network. The output of the MLP indicates if the object is among those memorized by the recognition system or not. Whether the object is recognized by the MLP, the output value of the network is used for setting the attribute recognition parameter \( R_{i,j} \) (Figure 5), where \( i \neq j \) represent the spatial position of pixels inside each segment. Initially \( R_{i,j} = 0 \) for all neurons. At the end of this process, all the neurons related to objects that should receive attention (top-down modulation) will be assigned to a value of recognition \( (R_{i,j} = 1) \) that will modulate the attentional process.

Fig. 5. Segmentation and recognition value.

B. The SOM Network

According to Haykin [4], one of the main attributes of a SOM network is to map an input signal pattern of arbitrary dimension into a network of one or two dimensions. Moreover, this mapping process is performed in order that the proximity of samples in the original feature space are preserved into the network (topological property). Here, a SOM has been applied to samples of \( \eta \)-dimension: color channels \( RGB \), intensity contrast \( I \), spatial difference in colors \( RG \) and \( RY \), orientations \( O_{\theta} \) with \( \theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \), location \([i,j]\) in a two-dimensional plane of each pixel of the input image and, finally, the recognition value \( R_{i,j} \) of each pixel set by the recognition module (MLP).

For training the SOM, we take all pixels of the scene with their respectively associated values \([R,G,B,I,RG,BY,O_0,O_{45},O_{90},O_{135},i,j,R_{i,j}]\). Note that the attribute \( R_{i,j} \) was obtained as described in the former section. The attributes next to each other according to their features are mapped in next positions in the SOM.

The SOM network feeds the Attribute-Saliency Map described in the next section.

C. The Attribute-Saliency Map

The Attribute-Saliency Map is defined as a network composed of neurons with two types of connections: excitatory and inhibitory. Excitatory connections represent a cooperative mechanism responsible for synchronizing groups of neurons that represents closely patterns of the SOM network (neurons with similar weights). Moreover, the inhibitory connections are designed to inhibit SOM regions related to background objects of the scene, allowing the region of the SOM related to the most salient object of the scene to be selected.

When an object (neuron) pulse in an instant \( t \), defined by Equation (5), it signal is presented the SOM network. The winner neuron of the SOM, stimulate its associated neurons of the attribute-salience map, which has its state updated by the following equation:

\[
\dot{v}_i = -v_i + E_i - W_Y Y_i, \quad i = 1, \ldots, n \tag{8}
\]

where \( n \) is the number of neurons of the saliency map. Equation (8) represents an Integrate and Fire neuron. It is important to observe that the attribute-salience map is constituted by a number of neurons equal to that of the SOM. It means that each neuron of the SOM network has its respective neuron in the attribute-saliency map. The variable \( v_i \) represents some voltage-like state of oscillator \( i \), \( W_Y \) is the weight of inhibition from the coupling inhibitory \( Y_i \).

Let \( l \) be a neuron belonging to an active segment into the LEGION, and \( k \) its respective index, the pattern \( l_k = [R, G, B, I, RG, BY, O_0, O_{45}, O_{90}, O_{135}, i, j, R_{i,j}] \) is presented to the SOM. The similarity between the pattern representing pixel \( l_k \) to its respective winner SOM neuron is defined by the following equation:

\[
d(m,l) = \exp \left(-\frac{\sum_{j=1}^{13} W_j (m_j - l_j)^2}{1} \right) \tag{9}
\]

The excitatory coupling term \( E_i \) and the inhibitory coupling term \( Y_i \) are defined by the following equation:
in which $E_i$ will be updated if and only if the value of $E_i$
contain the maximum value of activation of the neuron $i$
(winner), $m$ represents each neuron of SOM, $j$ is the feature
index and $W_j$ defines the weight associated with each feature
$j$. Given that the information bottom-up and top-down are
clustered into a single map, adjusting the weights $W_j$ makes
it possible to direct the attention for desired features. So that,
if $W_j = 0$ for all the primitive information of the input image,
our model becomes top-down model, and if $W_j = 0$ for
information related to object recognition, our model becomes
a bottom-up model.

The inhibitory connections are determined based on the
contrast among attributes. Thus, if two neurons are fed by
similar attributes, that is, the contrast between them is small
or zero, the term $Y_i$ Equation (8) approaches to zero, be-
cause the function negative exponential of this equation, the
weight of inhibitory coupling assumes a high value. On the
other hand, when the signal of such neurons are defined by
different attributes, the weight of the inhibitory connection
between them is small or even zero. Thus, objects with similar
characteristics are mutually inhibited because of competition
generated by inhibitory connections. An object that has a high
contrast to the other is not inhibited and remains oscillating,
that is, represents the attributes of the object under focus of
attention of the system.

Generally, it is assumed that the salient object is the one
which presents the greatest contrast with the other objects
in the scene, named here the salient object. This assumption
is supported directly from biological experiments that have
shown that the contrast of objects that make up the scene is
considered more important than the absolute level of each of
the visual attributes of tasks in visual inspection [17], [18].

D. Object Selection

To identify in the input image the of salients objects, we
consider the term $v_i$ of Equation (8), which represents the
potential of saliency of the neuron $i$. The attributes on the
neuron with potential $v_i$ of higher value, it is equal the
attributes of the most salient area of the input image. To
identify the position of the salient object in the input image is
necessary only to verify the attributes $x$ and $y$ of the neuron $i$,
which indicates the spatial position of a pixel into the scene.
Consequently, the whole segment associated to that pixel is
highlighted. As the value of $v_i$ decreases, we can identify the
next salient objects of the image.

IV. COMPUTER SIMULATIONS

This section analyses the behavior of the proposed model
taken a set of 3 images into account. We run some experiments
considering the colors channels $R$, $G$ and $B$, intensity contrast
$I$, difference in colors $RG$ and $BY$, orientations $O_0$, $O_{45}$, $O_{90}$
and $O_{135}$, location in a two-dimensional plane $x$ and $y$ the
recognition value $R_{xy}$. 

$$E_i = Y_i = d(m_l, l_k), \quad h = 1, \ldots, n \quad (10)$$

Fig. 6. Simulation 1. (a)Input image (b)SOM (c)Salient attribute (d)Salient
object (e)Segmentation and Known Object (f)RG (g)BY (h)O_0 (i)O_{45} (j)O_{90} (k)O_{135}.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 7. Object selection on real scene. (a) Results from our model (b)
Location selected as most salient from Itti et al. model [7].

The Figure 6(a) presents an image containing apparently
three salient objects (the red car, the blue traffic signal and
the white strip) regarding to the contrast to other objects
(background). In (d), the position of the most salient object is
presented (see the white cross signal). The SOM generated
from the input image considering all the features is presented
in (b). The map of attribute-saliency presented in (c) depicts
the main region of attention in the feature space. For this input
image, the main attributes that guide attention to the most
salient objects are the maps (i) and (k), the other maps, due to
its homogeneity, do not guide attention. The same scene was
was applied as input to Itti’s model and the results are shown
in Figure 7. From the viewpoint bottom-up, we concluded that
the results obtained by our model are similar.

According to [17], some salient features are very easy to
find albeit, for some cases, as for example in Figure 8(a), the
variation in orientation of an object is not a guiding feature. In
the next simulation we used a synthetic image to demonstrate the overall behavior of the our model in this case. The Figure 8(a) presents an image with an object previously trained by MLP, which must be recognized. The Figure 4(d) presents the trained object. As one can see in Figure 8(b), the top-down modulation can overcome bottom-up saliency by selecting a known object instead of the (bottom-up) most salient, it is even clear in the absence of any bottom-up clue. Due the top-down characteristics of our model, it was possible to highlight the salient object (top-down).

Fig. 9. Simulation 3. (a)Input image (b)Segmentation (c)Salient object.

Depending on the input scene and the diversity of salient objects, it is possible to adjust the parameter $W_j$ in Equation (9), in order to assign weights to desired features. The Figure 9(a) presents the image of a MRI (magnetic resonance imaging) diagnosed with cancer. Here, it is interesting to observe that attention is also guided by specific features which could bias attention for possible lesions. Figure 9(a) presents an image with an lesion (Figure 4(c)) previously trained by MLP. For this simulation the parameter $W_j$ was adjusted to bias the map $R_{x,y}$.

V. Conclusions

In this work we have introduced a object-based model of visual selection modulated by bottom-up and top-down attention. By including on object recognition modules implemented using a MLP network, our model were able to select objects regarding to its visual features as also the previous knowledge of the system. Moreover, computer simulations demonstrated the capacity of the model LEGION for segmentation, allowing our model to work with real images. To validate the proposed model, simulations with three images were carried out. There, we have shown that the model can select objects even in the absence of bottom-up guidance. As future work, we intend to carry out further experiments in real scenario and also improve the top-down mechanism.

REFERENCES