Universidade de São Paulo

Biblioteca Digital da Produção Intelectual - BDPI

Departamento de Ciências de Computação - ICMC/SCC

Comunicações em Eventos - ICMC/SCC

2014-10

An object-based visual selection model combining physical features and memory

Brazilian Conference on Intelligent Systems, 3th, 2014, São Carlos.

<http://www.producao.usp.br/handle/BDPI/48565>

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo

An Object-Based Visual Selection Model Combining Physical Features and Memory

Alcides X. Benicasa*, Marcos G. Quiles†, Thiago C. Silva‡, Liang Zhao§ and Roseli A.F. Romero‡

* Federal University of Sergipe, Itabaiana, SE, Brazil

Email: alcides@ufs.br

†Federal University of São Paulo, São José dos Campos, SP, Brazil

Email: quiles@unifesp.br

‡University of São Paulo, São Carlos, SP, Brazil

Email: {thiagoch,rafrance}@icmc.usp.br

§University of São Paulo, Ribeirão Preto, SP, Brazil

Email: zhao@usp.br

Abstract—In this paper, a new visual selection model is proposed, which combines both early visual features and object-based visual selection modulations. This model integrates three main mechanisms. The first is responsible for the segmentation of the scene allowing the identification of objects. In the second one, the average of saliency of each object is calculated for each feature considered in this work, which provides the modulation of the visual attention for one or more features. Finally, the third mechanism is responsible for building the object-saliency map, which highlights the salient objects in the scene. It will be shown that top-down modulation can overcome bottom-up saliency by selecting a known object instead of the most salient (bottom-up) and is even clear in the absence of any bottom-up clue. Several experiments with synthetic and real images are conducted and the obtained results demonstrate the effectiveness of the proposed approach for visual attention.

Keywords—bottom-up and top-down visual attention; object-based attention; recognition of objects.

I. INTRODUCTION

Every day we are faced with complex scenes, and our visual system needs to be able to analyze and understand a large amount of visual information while ignoring unimportant things. In order to do so, our visual system must deliver proper attention to specific objects while ignoring others. Biologically, an object with features that contrast to the background pops out and draws attention automatically [1]. The information that defines the contrast among objects is related to both primitive features of the scene and previous knowledge about specific targets (memory). This information is related to two distinct components that driving the human visual attention: the first one is the bottom-up attention, where the visual attention is involuntarily guided by visual features, such as colors, deep, etc. and the second is the top-down attention, in which the voluntary visual control guides the attention towards specific features or known objects in the scene [2]. It is worth noting that when there are objects with similar primitive features, top-down modulation turns out to be a dominate attribute to select one of the objects as the target.

Bottom-up models, as in [3], [4], [5], [6], do not consider the role of a working memory in the visual selection. In those models only the primitive features of the image are used to identify the salient point or the region. The selection is directly related to the unsupervised learning, whose goal is to find groups of similar objects according to their features without any type of label or supervision. On the other hand, the associative memory considered in top-down models might be associated to some form of supervision [7]. In this case, the objects to be identified in the scene must be previously labeled, which is an inherent characteristic from supervised learning methods, being the search guided by the prior external information about a specific object. This processing involves a concept of working memory, in which some information, such as a particular object or a location is kept in memory and modulates the selection process.

In this work, a prior external information about an object is provide for combining the low level and high level data classifications [8]. The low level classification is implemented by a traditional classification technique. The high level classification exploits the complex topological properties of the underlying network constructed from the input data. It will be shown that the combined classification approach is robust against variations in pattern recognition. It is worth noting that top-down attention can also influence the response of bottom-up clues. According to [9], [10], bottom-up attention alerts us to salient items in the scene, whereas top-down attention modulates bottom-up signals when we need to look for specific something.

Here, a new object-based visual selection model with both bottom-up and top-down modulations is proposed. More specifically, this 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 Network, named LEGION [11], for image segmentation. The segmentation module is mandatory when dealing with objects and real scenes; a Network-Based High Level Data Classi-

Fig. 1. A flowchart of object-based selection model proposed.

fication, named HLC [8], for object recognition; a Network of Integrate and Fire Objects, which creates object-saliency maps, and finally, an object selection module, that selects all the salient objects in the scene, based on the guidance from the object-saliency map. We provide qualitative and quantitative comparisons of the proposed model against an ground truth fixation maps [12] and four state-of-the-art proposed methods [3], [13], [14], [15] for object-based salient region detection.

II. PROPOSED 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 network-based high level data classifier for object recognition, a network of integrate and fire neurons, which creates the object-saliency map and, finally, an object selection module, which highlights the most salient objects in the scene.

Figure 1 depicts a flowchart of proposed model. Firstly, the scene is presented to the module responsible for extracting the early visual features of the scene and to LEGION segmentation network. Those modules were implemented following the specification described in [2] and [16], respectively. The output from those modules feed the following modules: the network for object recognition and the network of integrate and fire neurons, which creates the object-saliency map. These two last modules are presented in the following.

As described in previous sections, the proposed model takes both bottom-up and top-down modulations into account.

Early visual attributes, i.e. intensity contrast, color contrast and orientation contrast, define 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 the proposed model to select the salient objects of a given scene, the Network-Base High Level Classification (HLC) [8], must be trained with a set of objects representing desired targets of the scene.

After the training process, HLC network is able to recognize a set of segments (objects). Thus, the overall dynamics of the system can be understood as it follows. Each time a segment is highlighted into LEGION network, or the segment is pulsing, it is directly presented to HLC network. The output of the HLC indicates whether or not the object is among those memorized by the recognition system. If the object is recognized by HLC network, the output value of the network is used for setting the attribute recognition parameter $R_{i,j}$, where i and 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 ($R_{i,j} = [0, 1]$), that will modulate the attentional process. It is worth noting that segments with value of recognition below of a threshold, θ_r , value will not be considered as recognized. Hence, the value of recognition $R_{i,j}$ is defined by:

$$R_{i,j} = \begin{cases} 1, & \text{if } R_{i,j} \geq \theta_r \\ 0, & \text{otherwise} \end{cases}, \quad (1)$$

Next, for each pixel of the image, the following descriptors were extracted: intensity contrast I , spatial difference in colors RG and RY , orientations O_θ 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 the recognition value $R_{i,j}$ of each pixel set by the HLC recognition module.

Let l be a neuron belonging to an active segment into LEGION model, and k its respective index of the features, denoted here as $l_k = [I, RG, BY, O_\theta, i, j, R_{i,j}]$. Once the segmentation process is completed and the value of the saliency of all pixels that belong to the input image are properly calculated, the average of saliency of each feature k of the segment j can be defined as:

$$S_k^j = \frac{1}{n^j} \sum_{i \in n^j} l_{k,i}^j, \quad (2)$$

where n^j symbolizes the number of neurons in segment j and $l_{k,i}^j$ is the value of the saliency map at neuron i belonging to the feature k at segment j .

According to [17], [18], [19], another important feature that might guide the deployment of attention is the size of the object, which, in this work, is represented by n^j , i.e. the object size is incorporated into the saliency value S_k^j by the number of neurons in segment j . Therefore, the vector of the features is redefined as $l_k = [I, RG, BY, O_\theta, i, j, R_{i,j}, S]$. The range of any feature has been adopted to stay within the interval $[0, 1]$ by a normalization procedure defined as:

$$l'_k = \frac{l_k - \min_{l_k}}{\max_{l_k} - \min_{l_k}}, \quad (3)$$

where $\max_{l_k}$ and $\min_{l_k}$ represent the maximum value and the minimum value of the feature k , respectively.

The saliency value of an object is considered one of the major components of this work, as only from these values it is possible to select objects. The saliency value of an object is calculated based on bottom-up features, top-down features, or even both. Thus, the absence of salient features may automatically delete the regions or objects in the scene. Biologically, this behavior is termed asymmetrical search according to [18] and [20]. The asymmetry given to the presence or absence of a particular feature represents an important information for the development of visual attention models. Thus, here we propose a normalization of the saliency values of objects by the real contrasts values of saliency, described as follows:

$$\mathcal{N}_s(S_k^j) = \frac{1}{n^s} \sum_{i \in n^s} |D(S_k^j) - D(S_k^i)|, \quad (4)$$

$$D(S_k) = \frac{1}{n^s} \sum_{i \in n^s} |S_k - S_k^i|, \quad (5)$$

where n^s is the number of segments generated by the LEGION and $D(S_k)$ is the mean similarity between the saliency value of the feature k of an object j in relation to other objects. In general, according to Equation 5, objects with values near

the mean saliency, should have low contrast values, on the contrary, other objects are defined with high contrast values.

The Object-Saliency Map is usually defined as a network composed of neurons ([6], [5], [21], [7]). In this work, we will not consider each neuron individually, but groups of neurons that represent objects, defined by Equation (2). Therefore, the object-saliency map is defined as a network composed of objects, with two types of connections: excitatory and inhibitory. Excitatory connections represent a cooperative mechanism responsible for synchronizing groups of objects that represents closely patterns of similarities (objects with similar features). Moreover, the inhibitory connections are designed to inhibit objects related to background objects of the scene, allowing the object related to the most salient object of the scene to be selected.

When an object pulses in the LEGION, its signal is compared with all the other selected objects, whose states are updated by:

$$\dot{v}_i = -v_i + E_i - W_Y Y_i, \quad i = 1, \dots, n^j \quad (6)$$

where n^j is the number of pulsating objects. The variable v_i represents some voltage-like state of segment i , E_i is the excitatory coupling term and W_Y is the weight of inhibition from the coupling inhibitory Y_i . Equation (6) represents an integrate and fire neuron. However, in this case a neuron represents a whole object. This is possible due to the average of saliency of each feature of each object (see Equation(2)).

Let S_k^j be an object belonging to an active segment into LEGION model, and k its respective index of the features. The excitatory coupling term E_i and the inhibitory coupling term Y_i are defined as it follows:

$$E_i = Y_i = d(S_k^i, S_k^j), \quad (7)$$

in which E_i will be updated if and only if the value of E_i contains the maximum value of activation of the object i , S_k^i represents each pulsating object and k is the feature index. The similarity between the features representing the object S_k^j to the other objects is defined by:

$$d(m, l) = \exp \left(-\sqrt{\sum_{k=1}^8 W_k (m_k - l_k)^2} \right), \quad (8)$$

where W_k defines the weight associated with each feature k . Adjusting the weights W_k makes possible to direct the attention for desired features. So that, if $W_k = 0$ for all the primitive information of the input image, the proposed model becomes top-down model, and if $W_k = 0$ for information related to object recognition, it becomes a bottom-up model.

The inhibitory connections are determined based on the contrast among features. Thus, if two objects are fed by similar feature, that is, the contrast between them is small or zero, the term Y_i , in Equation (6), approaches to zero, because the weight of inhibitory coupling assumes a high value. On the other hand, when the signal of such objects is defined

Fig. 2. Qualitative analysis of OSM in real scenes in comparison to the fixation map [12] and saliency maps: [3], [13], [14], and [15], respectively shown from left to right. Benchmark images publicly available by [20] (1-4) and [12] (5-7).

by different features, the weight of the inhibitory connection among them is small or even zero. Thus, an object that has a high contrast in relation to the others is not inhibited and remains oscillating, that is, represents the features of the object under focus of attention of the system.

Roughly speaking, it is assumed that the salient object is the one which presents the greatest contrast in relation to other objects in the scene. This assumption is supported directly from biological experiments that have shown that the contrast is more important than the absolute level of each of the visual attributes of tasks in visual inspection [17], [22].

III. COMPUTER SIMULATIONS

According to [12], many visual attention models have been proposed to predict the locations of the scene in which a human would direct its attention, however, for each new model, the attention mechanism is evaluated using new images, which makes it difficult to compare the results. Thus, to minimize this limitation while performing qualitative and quantitative analysis, here we use fixation maps (FM) generated by tracking the eye movements of human observers for a variety of images. This publicly available data sets is maintained by several researchers (see [12] for a review). From this huge data set composed of several scenarios, we

select sixty-one real scenes from the benchmark proposed in [20] (36 scenes) and [12] (25 scenes). We also analyze the model behavior regarding to a variety of orthodox synthetic images derived from the psychophysical literature, specifically, we use the benchmark proposed in [20] (15 synthetic images). For comparative purposes, in addition to maps fixations and saliency map proposed by [3], we also take state-of-art models into account [12]. Therefore, the analysis presented in this section is based on the comparison of the proposed object-salient map (OSM) in relation to the fixation maps (FM) and the saliency maps generated according to the models proposed by [3], [13], [14], and [15], categorized respectively as: cognitive model, graph-based model, and spectral analysis for the last two models. Finally, we analyze our model in a specific domain to demonstrate its behavior on detecting salient objects.

A. Heterogeneous Domain

For the first set of simulations, the knowledge about base features of the scene are used to bias the attention process. Figure 2 shows 6 scenes from distinct domains in which objects contrast to the foreground according to the following features: intensity, color, orientation, location and size. The object recognition value was not taken into account, since the

Fig. 3. Mean similarities. (a) Analyzed models regarding to fixation maps (FM) of [20] and [12] for a total of 61 scenes (b) Analyzed models regarding to the fixation maps (FM) for a total of 15 synthetic images (c) Analyzed models regarding to the fixation maps (FM) for a total of 28 scenes with road signs.

scenes deal with distinct domains. In a qualitative analysis of the outcome presented in Figure 2, we can observe that the OSM results are quite consistent with the results of the fixation maps [12]. Qualitatively, we conclude that the OSM model can efficiently predict regions of the scene in which human observers tend to gaze. We also perform a quantitative analysis by using the similarity measure proposed in [12], to analyze the similarity between the object-saliency map and the fixation map.

In Figure 3(a) are presented the similarities mean obtained by the proposed model and the other models analyzed. According to this behavior analysis and the similarity mean, we can observe the effectiveness of the proposed model in comparison to the others.

B. Psychophysic Domain

According to [20], computational models of visual attention have been of great value for the understanding of biological attention. In the experimental counterpoint, several experiments have been conducted with primates to analyze the visual reaction according to the presented stimulus. From these biological experiments, several synthetic images have been constructed and are widely applied in experimental and simulation studies of visual attention. The simulations presented below consider these orthodox images derived from psychophysical literature. More specifically, we use the benchmark data from [20]. The purpose of these simulations is to demonstrate the behavior of the proposed model and the literature models in specific psychophysical conditions. For this set of experiments specifically, the fixation maps were not available. Thus, for comparative purposes, we generate the fixation maps according to the experimental work described in [20].

Initially, the first image depicted in Figure 4 is based on experiments related to the integration feature theory [1], in which objects that have contrasting features to other objects of the scene, *pop-out*, receiving attention. In the second image, primitive features related to color, size and orientation sequentially guide the visual attention, however, among the patterns “5” we can observe a pattern “2”, which can be

selected only with a conjunctive search. In images 3 and 4 we show experiments varying the similarity between the target object and the distractors. In this scenario, according to [23], [24], and [20], when an object has a certain contrast in relation to the others, it should receive the attention. Then the images 5 and 6, we consider the theories presented in [18] and [25], which is based on the asymmetry credited to the presence or absence of salient features for delivering the attention. In these images, the task of finding the object “+” among “-” is easier than finding the “-” among “+”, however the salience should exist in both tasks. The results obtained by the methods proposed in [3], [13], [14] and [15], and the object-saliency map proposed in this work applied on the benchmark proposed in [20] are depicted in Figure 4.

According to the qualitative analysis of the maps depicted in Figure 4, we can observe that the behavior of the proposed model for predicting the most salient regions was consistent with the fixation maps. The models proposed in [14] and [15], owing to the simple color-based contrast mechanisms, presented no significant predictive values. On the other hand, the models [3] and [13] that consider intensities, colors and orientations features to create the saliency map, presented greater accuracy in predicting the salient locations. Quantitatively, the behavior of all models was quite similar. Further, due to the object-based competition mechanism proposed here, objects with similar features were totally inhibited, which significantly increased the selection process. This behavior can be observed in images 1, 2, 3, 4 and 5 presented in Figure 4. Figure 3(b) shows the mean similarity obtained by each model.

C. Homogeneous Domain

Considering the parameter W_k (Eq. (8)), one can modulate the visual attention to the proposed model for attributes of larger membership in the scene. In the following simulations, it has been considered a driver driving by a road, where a large amount of information, both top-down as bottom-up, may emerge dynamically. Here, it is intended to show that HLC network can indeed help solving real world problems, specifically those pertaining to the pattern recognition area.

Fig. 4. Qualitative analysis of OSM in synthetic scenes, compared to the fixation map empirically generated for comparative purposes and also with the saliency maps proposed in [3], [13], [14] and [15] respectively shown from left to right. Benchmark images publicly available by [20].

In this context, HLC network is trained to recognize the following road signs shown in Figure 5. (see [8] for details about this classifier). Then, when a segment is pulsing into LEGION network, it is presented to HLC network and its output indicates the classification membership value of each segment. In this simulation, values have been assigned to the parameter in order to direct the attention to recognized objects that are the road sign. It is worth noting that in this simulation, due to the similarity among images and goals, it was possible to determine the most appropriate choices of the parameter values. Thus, according to Figure 5 (OSM), the objects with the greatest saliency value are the road signs. Figure 3(c) shows the mean similarity obtained by each model.

IV. CONCLUSIONS

In this work, a top-down contextual classifier and bottom-up object-based visual selection model was proposed for location of salient objects in real images and synthetic images. The proposed model was able to select objects regarding to their visual features as well the previous knowledge of the system. Thanks to this model, top-down modulation can overcome

bottom-up saliency by selecting a known object instead the (bottom-up) most salient being even clear in the absence of any bottom-up clue. In the proposed model, the competition for attention, occurs among objects, which differs from other models in which the competition occurs among pixels. A quite interesting feature of the proposed model is the possibility of the modulation in relation to specific features, turning it able to be used in different domains. Several simulations were conducted on synthetic and real scenes and the results showed that the proposed model is a promising mechanism for visual attention systems. As future work, we intend to consider other important features to guide the visual attention such as texture and to apply the proposed model on video, for this purpose we will consider the top-down bias, able to direct visual attention previously to the segmentation process. We also intend to develop automatic parameter settings for specific domains.

REFERENCES

- [1] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97 – 136, 1980.
- [2] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, pp. 194–203, 2001.