Artif Intell Rev (2010) 34:109–119
DOI 10.1007/s10462-010-9162-1

A biologically inspired object-based visual attention model

Longsheng Wei · Nong Sang · Yuehuan Wang

Published online: 28 May 2010

Abstract A biologically inspired object-based visual attention model is proposed in this paper. This model includes a training phase and an attention phase. In the training phase, all training targets are fused into a target class and all training backgrounds are fused into a background class. Weight vector is computed as the ratio of the mean target class saliency and the mean background class saliency for each feature. In the attention phase, for an attended scene, all feature maps are combined into a top-down salience map with the weight vector by a hierarchy method. Then, top-down and bottom-up salience map are fused into a global salience map which guides the visual attention. At last, the size of each salient region is obtained by maximizing entropy. The merit of our model is that it can attend a class target object which can appear in the corresponding background class. Experimental results indicate that: when the attended target object doesn't always appear in the background corresponding to that in the training images, our proposed model is excellent to Navalpakkam's model and the top-down approach of VOCUS.

Keywords Visual attention · Object-based · Salience map · Attentional selection

1 Introduction

The visual system requires attention and guidance of that attention because the eyes provide the central nervous system with more information than it can process. Attention has been

L. Wei (✉) · N. Sang · Y. Wang
Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China
e-mail: weilongsheng@163.com

N. Sang
e-mail: nsang@hust.edu.cn

Y. Wang
e-mail: yuehwang@mail.hust.edu.cn

classified into two types based on whether its deployment over a scene is primarily guided by scene features or volition: one is often called bottom-up and is mainly driven by low-level processes depending on the intrinsic features of the visual stimuli; the other refers to knowledge-based top-down processes (Henderson 2003). The top-down attention is more complex to model because it needs to represent object in long-term memory (LTM) (Hollingworth et al. 2001; Hollingworth and Henderson 2002; Hollingworth 2004) and uses the memory to detect likely object in attended scenes (Rensink 2000, 2002; Watanabe 2003).

Most bottom-up visual attention models (Le Meur et al. 2006; Shi and Yang 2007) are inspired by the concept of feature integration theory (Treisman and Gelade 1980). The most popular is the one proposed by Itti et al. (1998) and it has become a standard model of bottom-up visual attention, in which salience according to primitive features such as intensity, orientation and color are computed independently. Lots of top-down visual attention models (Itti 2000; Navalpakkam and Itti 2002; Yu et al. 2009) combine conspicuity maps by weights to form a salience map. Each weight indicates how important the map is for target object. Therefore, how to compute the weight is very important to top-down visual attention. Well-known models include Navalpakkam's top-down model Navalpakkam and Itti (2006) and Visual Object detection with a CompUtational attention System (VOCUS) Frintrop (2006). Navalpakkam et al. used statistical knowledge to obtain the weight of feature map by maximizing signal-to-noise ratio of target object and its background. Top-down visual attention of VOCUS proposed a method to compute weight vector: For each training image, a weight vector was computed as the ratio of the mean target saliency and the mean background saliency; then final weight vector was geometric mean of all the weight vectors. That means that target always appears in the corresponding background for every times, but in fact, a training target can appear in several different training backgrounds and a training background can contain different training targets.

In order to address the above problem, a biologically inspired object-based visual attention model is proposed in this paper. This model includes a training phase and an attention phase. In the training phase, all training targets are fused into a target class and all training backgrounds are fused into a background class. Weight vector is computed as the ratio of the mean target class saliency and the mean background class saliency for each feature. In the attention phase, for an attended scene, all feature maps are combined into a top-down salience map with the weight vector by a hierarchy method. Then, top-down and bottom-up salient map are fused into a global salience map which guides the visual attention. At last, the size of each salient region is obtained by maximizing entropy. The merit of our model is that it can attend a class target object which can appear in the corresponding background class. It is agreement with human visual perception. This model presents a new method for biologically inspired visual attention. Our proposed model is shown in Fig. 1.

The remainder of this paper is organized as follows. Section 2 presents the object representation including feature extract and the training of object representation. While salience map is described in Sect. 3, this part introduces how to acquire top-down salience map and bottom-up salience map. Attentional selection is described in Sect. 4, this part introduces how to acquire global saliency map and how to calculate the size of salient region. Section 5 shows experimental results, and Sect. 6 concludes this paper.

2 Object representation

Basic visual features such as intensity, color and orientation are extracted in this part. All training targets are fused into a target class and all training backgrounds are fused into a

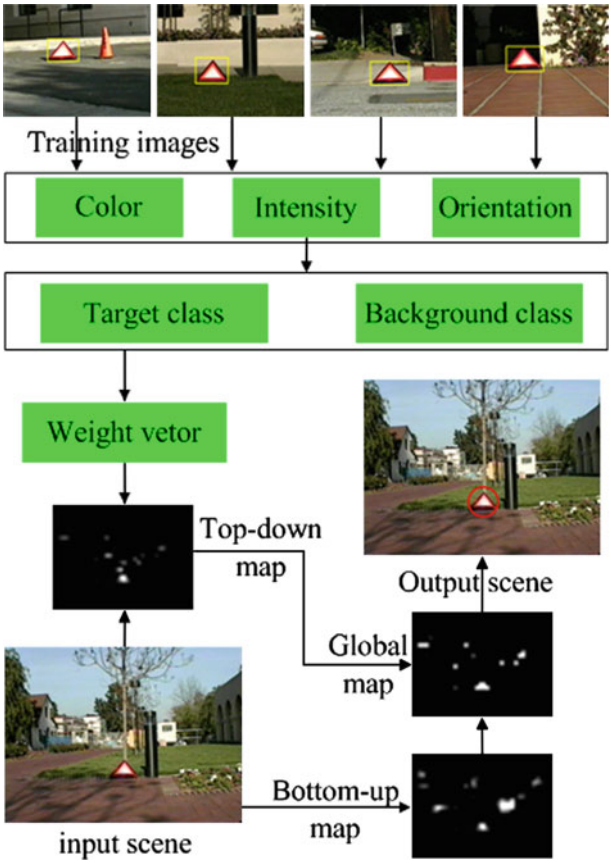

Fig. 1 Our model: Given a task such as "find a *triangle* in the scene". A weight vector is obtained from training images. For an input scene, all feature maps are combined into a top-down salience map with the weight vector by a hierarchy method. Then, top-down and bottom-up salient map are fused into a global salience map which guides the visual attention

background class. Weight vector is computed as the ratio of the mean target class saliency and the mean background class saliency for each feature.

2.1 Feature extraction

For every training image, ten low-level visual features including two intensity features, four color features and four orientation features are extracted in this passage. We divide intensity feature into intensity on (light-on-dark) and intensity off (dark-on-light). The reason is that the ganglion cells in the visual receptive fields of the human visual system are divided into two types: on-center cells respond excitatory to light at the center and inhibitory to light at the surround, whereas off-center cells respond inhibitory to light at the center and excitatory to light at the surround (Palmer 1999). In this paper, we convert the color input image into gray-scale image to obtain an intensity image and let center part be intensity on, surround part be intensity off. Let r, g and b are three color channels of input image, four broadly tuned color channels are created: $R = r - (g + b)/2$ for red, $G = g - (r + b)/2$ for green,

$B = b - (r + g)/2$ for blue and $Y = (r + g)/2 - |r - g|/2 - b$ for yellow (negative values are set to zero). Therefore, color features are divided into red, green, blue and yellow four parties. Four orientation features ($0°, 45°, 90°, 135°$) are computed by Gabor filters detecting bar-like features according to a specified orientation. Gabor filters, which are the product of a symmetric Gaussian with an oriented sinusoid, simulate the receptive field structure of orientation-selective neurons in primary visual cortex (Palmer 1999).

2.2 Training of object representation

To achieve a more stable, general target object representation, it is necessary to train the target properties from n training images. For every training image, the target object is provided as a rectangle which is usually determined manually by the user but might also be the output of a classifier that specifies the target. Every training image is decomposed into ten feature maps. VOCUS top-down proposed a method to compute weight vector: For each training image, a weight vector was computed as the ratio of the mean target saliency and the mean background saliency; then final weight vector was geometric mean of all the weight vectors. That means that target always appears in the corresponding background for every times, but in fact, a training target can appear in several different training backgrounds and a training background can contain different training targets. Therefore, we fuse all training targets as a target class and fuse all training backgrounds as a background class. The weights indicate how important a conspicuity map is for the attended target. Weight vector is computed as the ratio of the mean target class saliency and the mean background class saliency for each feature: the weight w_f of feature map Map_f is computed by

$$w_f = \frac{1}{n}\sum_i m_{i,f,target} \Big/ \frac{1}{n}\sum_i m_{i,f,background} \tag{1}$$

$$= \sum_i m_{i,f,target} \Big/ \sum_i m_{i,f,background} \tag{2}$$

$$f \in \{1, \ldots, 10\}, i \in \{1, \ldots, n\}. \tag{3}$$

For the ith training image and the fth feature, $m_{i,f,target}$ denotes the mean intensity value of the pixels at the target object in map Map_f, showing how strong this map contributes to the saliency of target object, and $m_{i,f,background}$ is the mean of the rest of the image in map Map_f, showing how strong the feature is present in the surroundings. The weights are computed for ten feature maps; together they form the weight vector $w = (w_1, \ldots, w_{10})$.

3 Salience map

For a given attention scene, we obtain a top-down visual salience map according to the target's weight vector represented in Sect. 2. In the same time, a bottom-up salience map is acquired by the contrast of the attention scene itself.

3.1 Top-down salience map

To attend a specific target object in input scene, we search for a target with help of the previously learned weights in the visual LTM to bias the combination of different feature

maps to form a top-down salience map. The weights are used to excite or inhibit the feature maps according to the search task. The weighted maps contribute to a top-down salience map highlighting those regions that are salient with respect to the target and inhibiting others.

According to Sect. 2, we know that when the weights are more than 1, the target is more salient than background in these feature maps, so we need to excite these feature maps; when the weights are smaller than 1, the background is more salient than target in these feature maps, so we need to inhibit these feature maps. Weights with value 1 are ignored since they indicate that the mean saliency of the target region is exactly the same as the mean saliency of the surrounding; such a feature is completely useless for attending the target. However, in practice this usually does not occur unless a feature is not present at all, e.g., color is not present in a gray-scale image and the color weights are set to 1. In order to achieve the effect of excitation, inhibition and neglect, we modify the feature weight as following.

$$w(f) = \begin{cases} w_f & w_f > 1 \\ 0 & w_f = 1 \\ -1/w_f & w_f < 1 \end{cases} \tag{4}$$

For a given attended scene, it is decomposed into ten feature maps and each feature map includes several spatial scales. We create a channel hierarchy H as follows (Navalpakkam and Itti 2005). $H(0)$ (leaves): the set of all features at different spatial scales; $H(1)$: the set of subchannels formed by combining features of different spatial scales and the same feature type; $H(2)$: the set of channels formed by combining subchannels of same modality; ...; $H(n)$: the salience map (where n is the height of H). In order to promote the target in all the feature channels in the channel hierarchy, each parent channel promotes itself proportionally to the maximum feature weight of its children channels.

$$\forall p \in \bigcup_{k=0}^{n} H(k), w(p) \propto \max_{c \in children(p)} (|w(c)|) \tag{5}$$

For instance, if the target has a strong horizontal edge at some scale, then the weight of the subchannel increases and so does the weight of the orientation 0° channel. Hence, those channels that are irrelevant for this target are weighted down and contribute little to the salience map (e.g., for detecting a horizontal object, color is irrelevant and hence the color channel's weights are decreased). At each level of the channel hierarchy, weighted maps of the children channels Map_c are summed into a unique map at the parent channel Map_p, resulting in the salience map at the root of the hierarchy.

$$\forall p \in \bigcup_{k=0}^{n} H(k), Map_p(x, y) = f\left(\max_{c \in children(p)} w(c) \times Map_c(x, y)\right) \tag{6}$$

where f refers to the spatial competition. For details regarding its implementation, please see Sect. 2.4 in Itti and Koch (2001). Top-down salience map S_{td} is obtained by discarding negative values in the nth hierarchy weighted map.

3.2 Bottom-up salience map

In human behavior, top-down and bottom-up attention are always intertwined and may not be considered separately, although one part may outweigh the other in certain situations. Even in

a pure exploration mode, each person has own preferences resulting in individual scan-paths for the same scene. Even if searching highly concentrated for a target, the bottom-up pop-out effect is not suppressible, an effect called attentional capture. The bottom-up salience map part in our proposed approach is implemented based on the model proposed by by Itti et al. (1998).

In order to be consistent with top-down salience map, we still use two intensity features, four color features and four orientation features. Center and surround scales are obtained using Gaussian pyramids with nine scales (from scale 0, the original image, to scale 8, the image reduced by a factor 256). Center-surround differences are then computed as pointwise differences across pyramid scales, for combinations of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$); thus, six feature maps are computed for each of the 10 feature types, yielding a total of 60 feature maps. Each feature map is endowed with internal dynamics that operate a strong spatial within-feature and within-scale competition for activity, followed by within-feature, across-scale competition. The feature maps are fused step by step, thereby strengthening important aspects and ignoring others. Resultingly, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings. All feature maps are then summed into the unique scalar salient location that guides attention.

4 Attentional selection

In this part, top-down salience map and bottom-up salience map are fused into a global salience map. The most salient region is determined and the focus of attention is directed there. Thus, we obtain salient regions, and then, we obtain salient sizes by maximizing entropy.

4.1 Global salience map

Top-down salience map and bottom-up salience map are described previously. The global salience map is the weighted sum of the top-down and the bottom-up map. Both maps compete for saliency: the top-down map emphasizing the features of the learned target; the bottom-up map showing regions that are salient because of scene-specific conspicuities. To make the maps comparable, S_{td} is normalized in advance to the same range as S_{bu}. When fusing the maps, it is possible to determine the degree to which each map contributes to the sum. This is done by weighting the maps with a top-down factor $t \in \{0, \ldots, 1\}$.

$$S_{global} = t \times S_{td} + (1 - t) \times S_{bu} \tag{7}$$

The global salience map is used to guide visual attention.

4.2 The size of salient region

The entropy maximum is considered to analyze the sizes of salient regions (Kadir and Brady 2001). The most appropriate scale x_s for each salient region centered at location x in the global salience map is obtained by (8) which aims to consider spatial dynamics at this location:

$$x_s = arg \max_s \{H_D(s, x) \times W_D(s, x)\} \tag{8}$$

where D is the set of all descriptor values which consist of the intensity values corresponding the histogram distribution in a local region with size s around an attended location x in global salience map, $H_D(s, x)$ is the entropy defined by (9) and $W_D(s, x)$ is the inter-scale measure defined by (10).

$$H_D(s, x) = -\sum_{d \in D} p_{d,s,x} \log_2 p_{d,s,x} \tag{9}$$

$$W_D(s, x) = \frac{s^2}{2s - 1} \sum_{d \in D} |p_{d,s,x} - p_{d,s-1,x}| \tag{10}$$

where $p_{d,s,x}$ is the probability mass function, which is obtained by normalizing the histogram generated using all the pixel values in a local region with a scale s at position x in the global salience map, and the descriptor value d is an element in a set of all descriptor values D, which is the same set of all the pixel values in a local region.

5 Experimental results

We apply three schemes of visual attention mechanism such as Navalpakkam's approach (Navalpakkam and Itti 2006), VOCUS top-down approach (Frintrop 2006) and our proposed approach on same scene images and same training images. We have done 100 group experiments and each group experiment includes 20 training images containing different views of the target, appearing at different locations in the scene. There are 50 nature scenes including 20 single object scenes and 30 multi-object scenes. At the same time, there are 50 artificial scenes including 30 single object scenes and 20 multi-object scenes.

We deal with the task which indicates the task-specific object. It includes two categories. One specifies the task-relevant features. For instance, if the task is searching for a horizontal red aligned object, the color red (the third features) and orientation 0° (the seventh features) are the task-relevant features, so let $w_3 = 1, w_7 = 1$, and other weights equal zero. The experimental result is shown in Fig. 2. The other does not specify task-relevant features. In this case, our approach uses salience descriptors in the object representation to deduce the task-relevant features. Taking a nature scene as example, Fig. 3 is some example of training images. A triangle is included in the entire training images and it is chosen as region of interest (ROI) marked by yellow rectangle. The result of weights for ten feature maps obtained from the training images is given in Table 1.

We begin to attend target object in new test scenes where the target object can appear in different backgrounds, different locations, views and sizes. According to the weights in

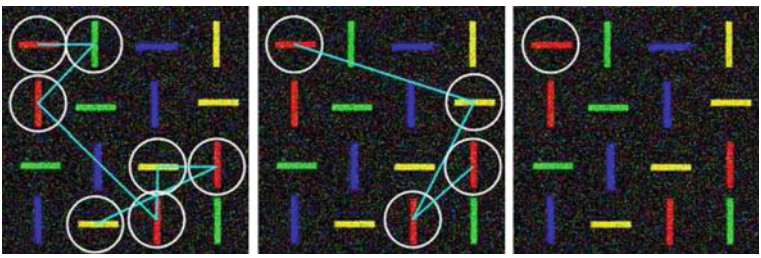

Fig. 2 *Left*: Navalpakkam's approach, the object is found in the seventh times; *Middle*: VOCUS top-down approach, the object is found in the fourth times; *Right*: our approach, the object is found in the first times

Fig. 3 Example of training images

Table 1 The weights for ten feature maps learned from the training images

Feature	Weights	Feature	Weights
Intensity on	0.3728	Color yellow	3.4520
Intensity off	5.3486	Orientations 0°	8.4733
Color red	6.7840	Orientations 45°	3.2912
Color green	0.4216	Orientations 90°	0.0584
Color blue	0.0068	Orientations 135°	7.0141

Fig. 4 *Left*: Navalpakkam's approach, the object is found in the fifth times; *Middle*: VOCUS top-down approach, the object is found in the second times; *Right*: our approach, the object is found in the first times

Table 1, we compute the excitation map, the inhibition map and obtain the global salience map. In experiment, we take $t = 0.9$ which expresses that the top-down map is very important and the bottom-up map just plays a supporting role in global salience map. The attention result of our approach marked by red circle is given in right of Fig. 4. By contrast, VOCUS top-down approach's attention result is also given in middle of Fig. 4 and Navalpakkam's approach attention result is also given in left of Fig. 4.

More examples are expressed in Fig. 5. We attend a white cup in the first column, a mobile telephone in the second column and a green cup in last column. The first row is Navalpakkam's approach attention results, the second row is VOCUS top-down approach attention results and the last row is our approach attention results. All object scenes experimental results of three approaches are expressed in Fig. 6.

According to above experimental results, artificial scenes experimental results are better than nature scenes experimental results. The reason is that the backgrounds of artificial scenes are simple, but the backgrounds of nature scenes are easy to be influenced by others, such as noise, illumination and clutter. Single object scenes experimental results are better than multi-object scenes experimental results. That is because the saliency of single object is easy to pop out, but multi-object is easy to be inhibited by itself. Our proposed approach is the better than other two approaches because it uses whole information of object and background in training images.

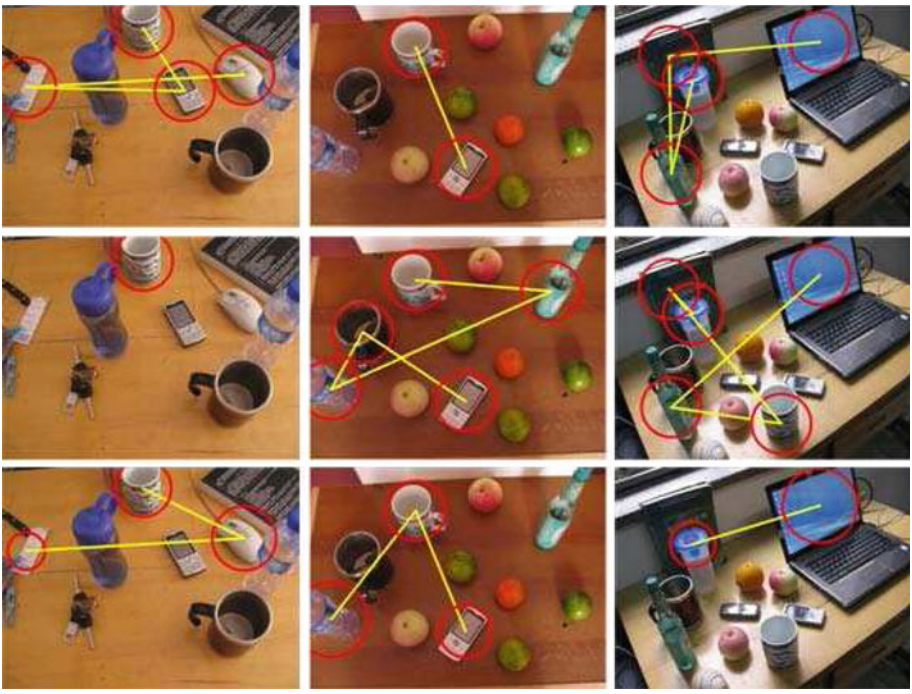

Fig. 5 The first column, the white cup is found in the fourth times, the first times and the third times by Navalpakkam's approach, VOCUS top-down approach and our approach, respectively; The second column, the mobile telephone is found in the second times, the fifth times and the third times by those three approaches, respectively; The last column, the green cup is found in the fourth times, the fifth times and the second times by those three approaches, respectively

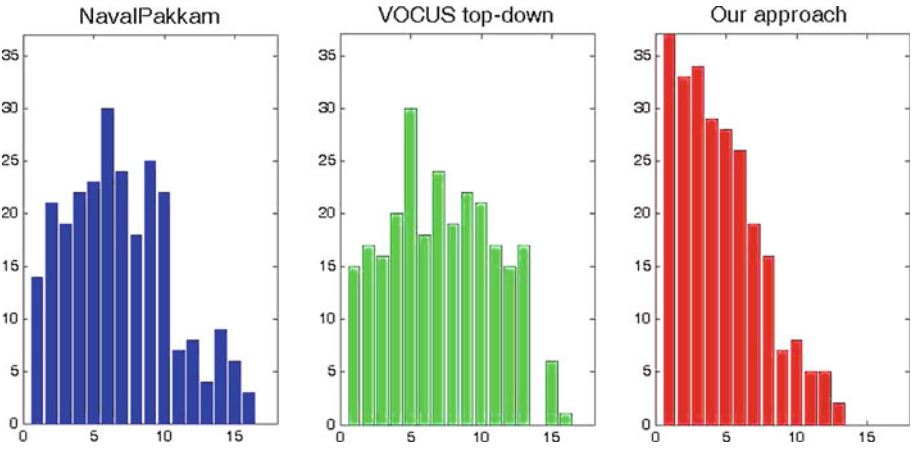

Fig. 6 All object scenes experimental results of three approaches (x-axis expresses the how many times the target object is found; y-axis expresses the total number of emerged in that times)

6 Conclusion

In this paper, a biologically inspired object-based visual attention model is proposed. The main process is described as following. Firstly, all training targets are fused into a target class and all training backgrounds are fused into a background class. Weight vector is computed as the ratio of the mean target class saliency and the mean background class saliency for each feature. Secondly, for an attended scene, all feature maps are combined into a top-down salience map with the weight vector by a hierarchy method. Then, top-down and bottom-up salient map are fused into a global salience map which guides the visual attention. Thirdly, the size of each salient region is obtained by maximizing entropy. The merit of our model is that it can attend a class target object which can appear in the corresponding background class. Experimental results indicate that: when the attended target object does not always appear in the background corresponding to that in the training images, our proposed model is excellent to Navalpakkam's model and the top-down approach of VOCUS. This model presents a new method for object-based visual attention.

In our future works, we will extend our top-down visual attention mechanism to work in dynamic environment by adding some new primitives such as optical flows.

Acknowledgments This work was supported by the National Natural Science Foundation of China under contracts 60736010 and the Chinese National 863 Grant No. 2009AA12Z109.

References

Frintrop S (2006) VOCUS: a visual attention system for object detection and goal-directed search. Lecture Notes in Artificial Intelligence (LNAI), 3899

Henderson JM (2003) Human gaze control during real-world scene perception. Trends Cogn Sci 7:498–504

Hollingworth A (2004) Constructing visual representations of natural scenes: the roles of short- and long-term visual memory. J Exp Psychol Hum Percept Perform 30:519–537

Hollingworth A, Henderson JM (2002) Accurate visual memory for previously attended objects in natural scenes. J Exp Psychol Hum Percept Perform 28:113–136

Hollingworth A, Williams CC, Henderson JM (2001) To see and remember: visually specific information is retained in memory from previously attended objects in natural scenes. Psychon Bull Rev 8:761–768

Itti L (2000) Models of bottom-up and top-down visual attention. Ph.D. thesis, California Institute of Technology. Pasadena, CA, January

Itti L, Koch C (2001) Feature combination strategies for saliency-based visual attention systems. J Electron Imaging 10(1):161–169

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259

Kadir T, Brady M (2001) Saliency, scale and image description. Int J Comput Vis 45(2):83–105

Le Meur O, Le Callet P, Barba D (2006) A coherent computational approach to model bottom-up visual attention. IEEE Trans Pattern Anal Mach Intell 28:802–817

Navalpakkam V, Itti L (2002) A goal oriented attention guidance model. Lect Notes Comput Sci 2525:453–461

Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. Vis Res 45:205–231

Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimal objeet deteetion speed. IEEE computer society conference on computer vision and pattern recognition, pp 2049–2056

Palmer SE (1999) Vision science, photons to phenomenology. MIT Press, Cambridge, MA

Rensink RA (2000) The dynamic representation of scenes. Vis Cogn 7:17–42

Rensink RA (2002) Change detection. Ann Rev Psychol 53:245–277

Shi H, Yang Y (2007) A computational model of visual attention based on saliency maps. Appl Math Comput 188:1671–1677

Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12:97–136
Watanabe K (2003) Differential effect of distractor timing on localizing versus identifying visual changes. Cognition 88(2):243–257
Yu Y, Mann GKI, Gosine RG (2009) Modeling of top-down object-based attention using probabilistic neural network. IEEE Canadian conference on electrical and computer engineering, pp 533–536