A Computational Model of Selecting Visual Attention Based on Bottom-up and Top-down Feature Combination

Wenyong Chen, Furao Shen, and Jinxi Zhao

***Abstract*— Selecting attention is an important cognitive psychology concept originally which has received much attention from scholars in the field of computer science. Nowadays, selecting attention has much application in computer vision. Most current computational models of attention focus on bottom-up features and ignore scene information. In this paper, a model of selecting visual attention guidance based on both bottom-up and top-down features was proposed. We used two datasets to evaluate the performance of the model, and also compare ours with Itti's model, which is particularly famous for visual attention. Experiments indicate that our model is applicable to the simulation of visual attention, and it achieves better performance in attention transferring than the models existed.**

I. Introduction

ATTENTION is often compared to a spotlight because of a selective gating mechanism [1], which means that the visual system can reduce the amount of input visual information to a small but important amount of data for processing [2]. And the process of selecting and gating visual data is not only based on saliency in the image itself (bottom-up), but also prior knowledge about the scene (top-down) [3].

The attention mechanism can efficiently deal with the balance between computing resources, time cost and performing different visual tasks in a complex, cluttered and dynamic environment [4]. So it has attracted an increasing interest in the field of salient region detection which is useful for image processing such as image indexing, matching, and retrieval and so on.

Commonly attention is believed to act before objects are recognized [5]. Attention mechanisms can suggest strategies for finding shortcuts for object detection and recognition [6]. The question is how we can attend to objects before we recognize them. Several computational models of visual attention have been suggested.

L. Itti [7] introduced a model for bottom-up selective attention based on serially scanning a saliency map, which is computed from local feature contrasts, for salient locations in the order of decreasing saliency. Marat et al [8] presented a bottom-up method for spatio-temporal saliency prediction in video stimuli. The model extracted two signals from video stream, and then derived two static and dynamic saliency maps from these signals. Finally, a spatio-temporal map was fused by using saliency maps. More models can be seen in [9]. The repetitious details need not be given here.

Wenyong Chen(wychen0204@163.com), Furao Shen(frshen@nju.edu.cn) and Jinxi Zhao(jxzhao@nju.edu.cn) are with the National Key Laboratory for Novel Software Technology, and Department of Computer Science and Technology at Nanjing University, Nanjing, 210046, P.R.China.

Previous studies have indicated that human observers use not only bottom-up information, but also top-down information relevant to the scene and prior knowledge, when pay attention to a specific region in a complex and cluttered scene [6][10][16].

Guided by these considerations, this paper aims to develop a new model of selective attention. We calculate a new attention value by using some top-down features, and reproduce a new attentional scan path of this "spotlight", which more corresponds to the human eyes.

This paper has been organized as follows. Section II describes the procedure of modeling attention. Section III shows our experimental results obtained, and also demonstrates the evaluation and analysis of the performance of our model. Section IV is the conclusion and discussion.

II. Model

This paper proposes a model of selecting attention combining the features based on both bottom-up and top-down. The architecture of model is shown as $Figure$ 1.

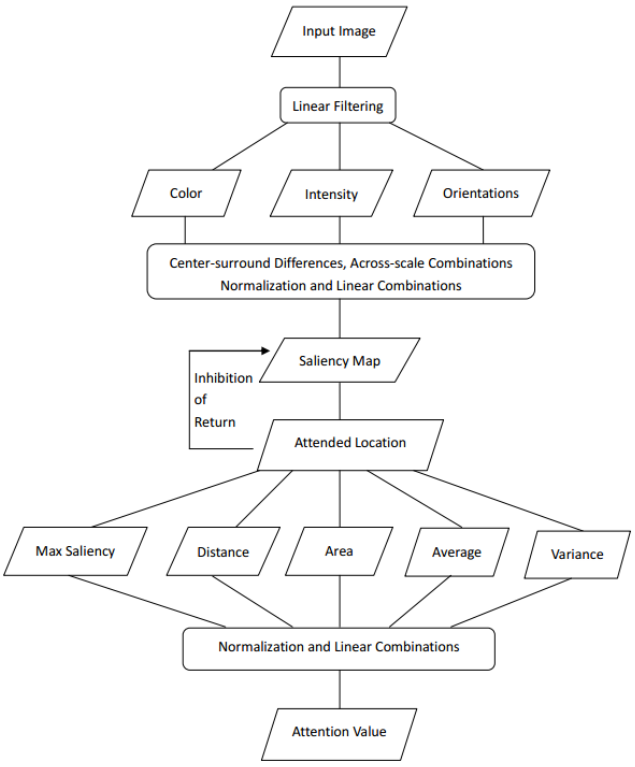


Fig. 1. General architecture of the model.

The model contains two parts roughly:

(1) The searching of attended region: figure out the saliency of the image, and then find the focus of attention. Finally mark the attentional region according the saliency and focus of attention.

(2) Top-down feature combination: according to the attended region got by step(1), and combined with the area of the region, location of the region, as well as the mean value and variance of saliency in the region, we calculate a new attention value and then draw up the path of attention transferring.

A. The Searching of Attended Region

The input image I is sub-sampled by convolution with a linearly separable Gaussian filter. By repeating sub-sampling and decimation process, a pyramid with multi-level is obtained [7].

Let r, g and b are red, green and blue channels of the input image respectively, the intensity map is computed as

$$M_I(\sigma) = \frac{r+g+b}{3} \tag{1}$$

where σ represents the level of the pyramid, the below is same.

For color features, red-green(RG) and blue-yellow(BY) opponencies are used to represent [11].

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

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

Orientation features are obtained from image I using oriented gabor pyramids $O(\sigma,\theta)$, where $\theta \in \{0^0, 45^0, 90^0, 135^0\}$ is the preferred orientation.

The saliency is obtained by computing center-surround difference, in which we use difference of Gaussian(DOG) with the following equation.

$$\begin{aligned} DOG(x,y) = &\frac{1}{2\pi {\sigma_c}^2} exp(-\frac{x^2+y^2}{2{\sigma_c}^2}) \\ &-\frac{1}{2\pi {\sigma_s}^2} exp(-\frac{x^2+y^2}{2{\sigma_s}^2}) \end{aligned} \tag{4}$$

The DOG is simply the difference between two Gaussian distributions with different σ values, where the center c of the model corresponds with the excitatory center Gaussian, and the surround s with the inhibitive surround Gaussian.

To get the saliency map, a normalization algorithm is needed. According to [12], firstly normalize all the feature maps to the same range, such as $[0,1]$, and then find its global maximum M and the average m of all the other local maxima for each map. Next is multiplying the map by $(M-\bar{m})$ globally. After a simply linear addition, a saliency map is obtained. The location with the highest saliency value in the saliency map is the focus of attention.

The following is a mechanism for extracting an image region around the focus of attention. First we find the feature map that contributes most to saliency map at winning location(focus of attention). In the winning feature map, activation spreads from the winning location over the shape of the object at this location. The winning feature map is denoted by F_w, and the winning location is set as (x_w, y_w).

In image processing terms, a threshold method is used [7]. The operation can be expressed by

$$B(x,y) = \begin{cases} 1 & if F_w(x,y) \geq 0.1 \times F_w(x_w,y_w) \\ 0 & otherwise \end{cases} \tag{5}$$

And after labeling the resulting binary map B around (x_w, y_w), an attended region can be obtained.

The locations in the saliency map compete for the highest saliency value by means of winner-take-all network. After the winning location is attended to, a mechanism named inhibition of return is adopt. Continuing competition produces the second most salient location.

B. Top-down Feature Combination

After obtaining underlying information about color, intensity and orientations, we also need some top-down prior knowledge from experience of human being to guide the modeling process. According to the method of section 2.1, we can mark some significant attended regions, and then calculate the following five features: the highest saliency value of each region, the area of the region, the position where each region locates in the image, the mean saliency value and variance of the region.

Imagine an image with the size of $m \times n$, when calculating the feature value of a region R, we can mark the pixel point in this region as 1, namely $label(x,y) = 1$, the point outside of the region R labeled as 0. The saliency value of pixel in $point(x,y)$ can be expressed as $Sal(x,y)$.

For the highest saliency value, its meaning is to find the maximum value of the region.

$$MaxSal_R = \max_{label(x,y)=1} Sal(x,y) \tag{6}$$

Because the large object will always be paid more attention by human being, we choose the region area as one of significant factor for visual attention [13]. For the area, we employ a ratio to express, namely the percent that the region is made up of in the whole image. The number of the points in the region can be recorded as N, which refers to the same meaning in the following equations.

$$N = \sum_{i=1}^{m}\sum_{j=1}^{n} label(i,j) \tag{7}$$

$$Area_R = \frac{N}{m \times n} \tag{8}$$

When a scene is observed, the object in middle of the sight is easier paid attention to. So the location an object lies in the image is also an important factor for visual attention. For

the regional location, we can express it by using a distance, which measures the length from the attended region to the center of the image. Here an average distance from all points of the region to the center point is calculated.

$$Dist_R = \frac{\sum_{label(x,y)=1} \sqrt{(x - centerX)^2 + (y - centerY)^2}}{N} \tag{9}$$

where $(centerX, centerY)$ is the center point of the whole image.

$$\begin{aligned} centerX &= \frac{m}{2} \\ centerY &= \frac{n}{2} \end{aligned} \tag{10}$$

According to [14], considering only the contribution to the strongest point as described in Section 2.1 cannot indicate the contribution to the whole region. Moreover, the combination of the feature maps could also lead to an erroneous strongest point resulting in an erroneous selection of the feature map as the winning map, and then result in erroneous selection of attended region.

So expect for the highest saliency value in the region, we need to consider introducing two new features: the mean and variance of saliency values. These two parameters can be used to ensure the point with the highest saliency is not a noise point. They are calculated as following equations:

$$Avg_R = \frac{\sum_{label(x,y)=1} Sal(x,y)}{N} \tag{11}$$

$$Var_R = \frac{\sum_{label(x,y)=1} (Sal(x,y) - Avg_R)^2}{N} \tag{12}$$

After the calculation of each feature value, the following is a normalization problem, namely how to adjust all the features to the same range for the purpose of consolidating five features into a scalar. Because these features represent different meanings, they cannot be linear added directly. For example, to obtain the region with the largest attention value, for the highest saliency, the area and the mean saliency of the region, the bigger value of these three features is the better, while the smaller of the distance and the variance value is the better.

Based on the considerations above, the following normalization method is proposed. Firstly we calculate the maximum and minimum of each feature, and then study by two cases:

$$feature_\sigma R = \frac{feature_\sigma R - \min(feature_\sigma)}{\max(feature_\sigma) - \min(feature_\sigma)} \tag{13}$$

$$feature_\delta R = \frac{\max(feature_\delta) - feature_\delta R}{\max(feature_\delta) - \min(feature_\delta)} \tag{14}$$

where $\sigma \in \{MaxSal, Area, Avg\}, \quad \delta \in \{Dist, Var\}$

After the operation of normalization, the following work is quite easy. For each region, after linear addition of the five normalized feature value, we can get the final attention value.

$$AttentionValue = \sum_{i \in \{\sigma,\delta\}} feature_i R \tag{15}$$

Then according to the attention value, the order of the attended region will be rearranged. Finally we can get a new sequence of attention transferring.

III. Experiment

A. Experimental Results

The model is extensively tested with different images to ensure proper functioning. The images are from two datasets in experiments. The first one contains the images used in [11], and the second one is the Berkeley segmentation dataset. Here we just select one image from each dataset to observe the result.

Firstly, we use an image from [11] shown as in $Figure.\ 2(a)$. $Figure.\ 2(b)$ show the result obtained by Itti's model. Here yellow lines drawn in the image are the attended regions, and the red lines represent the trail of attention transferring. We can see that the first attended object is the balloon with the shape of elephant in the upper-right corner of the image. And then the focus of attention turns to the yellow balloon in the bottom of the image. In this image, we just choose five earlier attended regions to do the comparative test. The following trajectory is as $Figrue.\ 2(b)$ shown.

According to these five attended regions, we use our model to make experiments. The attention values obtained using the feature combination proposed in this paper are listed in $Table$ I.

TABLE I
THE ATTENTION VALUE OF FIGURE. 2(A) OBTAINED BY OUR MODEL

Itti's Region	1st	2nd	3rd	4th	5th
Attention Value	2.1818	4.1203	1.7634	1.7775	1.8673

As can be seen from the table, the maximum attention value was achieved by the second attended region. In other words, the order of attention was changed. The previous sequence $1-2-3-4-5$ became $2-1-5-4-3$. We can draw the trail in accordance with this order as shown in $Figure.\ 2(c)$.

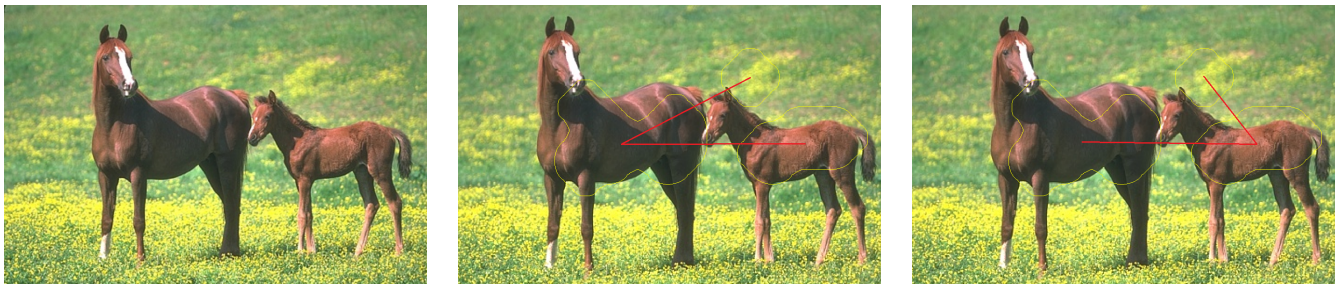
Then an image from the second dataset was used to test. The attention values and trail of attention transferring are shown as $Table$ II and $Figure.$ 3, which is drawn by the same way as the $Figure.$ 2. Here we just select three attended regions to test in this image.

TABLE II
THE ATTENTION VALUE OF FIGURE. 3(A) OBTAINED BY OUR MODEL

Itti's Region	1st	2nd	3rd
Attention Value	2.4895	4.1356	1.6667

(a) Original image (b) The sequence of Itti's model (c) The sequence of our model

Fig. 2. Experimental Results of our model using the image from dataset of [11]



(a) Original image (b) The sequence of Itti's model (c) The sequence of our model

Fig. 3. Experimental Results of our model using the image from Berkeley segmentation dataset

From $Figure.$ 3, we can see that the smaller horse is first attended by using Itti's model, while the bigger horse is attended before the smaller one by exploring our model.

B. Evaluation and Analysis

According to [15], many scholars are making strong attempts at looking for an evaluation method to answer the question "Which is the best model of visual attention". Nevertheless, the pity is that there is not a unified scheme yet. So a common view is that the effectiveness of a model is determined by how well it provides explanations for its function.

To evaluation the performance of our model proposed in Section II, namely rationality of attention transferring trail, we recruited 20 people to join the test. In the absence of any hint, they were expected to provide the order of attending the regions in given image. Through collecting the data received from these twenty people, we can get the results, which are the sequences of 2-5-4-1-3 and 2-1-3 corresponding to two images in the test. Then by comparing with the results obtained by Itti's model and ours, we can find the results of our model are closer to observation of human eyes. For example, people invited are apt to pay first attention to the yellow balloon in bottom of the image in $Figure.$ $2(a)$, and the bigger horse in $Figure.$ $3(a)$.

According to the results, we can infer that a location only with high saliency value will be a noise point in selecting visual attention. To avoid this situation, we need to consider more top-down features which are relevant to the whole scene.

IV. Conclusion

There has been an increasing interest in modeling selecting visual attention. In this paper, we propose a model based on feature combination, not only using bottom-up feature, but also top-down. Both Itti's dataset and Berkeley segmentation dataset are used in the study. The experimental results clearly suggest that our model is effective and outperforms Itti's model by testing two datasets. The achievement is reflected not only in selecting the first focus of attention, but also the transferring of attention.

To ensure the objectivity of evaluation of results, we will use more images and invite more people to test the model. In addition, the model doesn't consider the issue of saccade, which will be incorporated into the model and the experiment in the following work. Moreover, our future work should attempt to give a more efficient evaluation method to assess the performance of our method.

Acknowledgments

This work was supported in part by the 973 Program 2010CB327903, the Fund of the National Natural Science Foundation of Jiangsu NSF grant BK2011567.

The authors would like to thank reviewers for their helpful comments on the paper.

References

[1] D. Sagi, and B. Julesz, *Enhanced detection in the aperture of focal attention*, Nature, 321, 693-695, 1986.

[2] R. Desimone and J. Duncan,*Neural mechanisms of selective visual attention*, Annual Review of Neuroscience, 18:193-222. 2000

[3] L. Itti and C. Koch, *Computational modeling of visual attention*, Nature Reviews Neuroscience, 2(3):194-203, 2001.

[4] H. Pashler, *The psychology of attention*, Cambridge, MA: MIT Press, 1998.

[5] R. Egly, J. Driver, and R. D. Rafal, *Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects*, Journal of Experimental Psychology General, 123(2), 161-177,1994

[6] J. Tunnermann, C. Born, and B. Mertsching, *Top-Down Visual Attention with Complex Templates*, 2012

[7] L. Itti, C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254-1259, 1998

[8] S. Marat, T. Ho-Phuoc, and L. Granjon, *Modeling Spatio-temporal Saliency to Predict Gaze Direction for Short Videos*, IJCV, 2009.

[9] A. Borji, and L. Itti, *State-of-the-art in Visual Attention Modeling*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012

[10] A. Oliva, A. Torralba, and M. S. Castelhano, *Top-down control of visual attention in object detection*, Proc.IEEE Int.Conf. Image Process, 1:253-256, 2003

[11] D. Walther, and C. Koch, *Modeling attention to salient proto-objects*, Neural Networks, 19, 1395-1407, 2006.

[12] L. Itti, and C. Koch, *Feature combination strategies for saliency-based visual attention systems*, Journal of Electronic Imaging, 10(1), 161-169, 2001

[13] J. Zhang, L. Zhuo, and J. Gao, *A study of top-down visual attention model based on similarity distance*, Image and Signal Processing, CISP 09:1-5, 2009

[14] Y. Hu, and X. Xie, *Salient Region Detection suing Weighted Feature Maps based on the Human Visual Attention Model*, Lecture Notes in Computer Science, Volume 3332: 993-1000, 2005

[15] J. K. Tsotsos, A. Rothenstein, Scholarpedia, 6(1):6201, (2011), $http://www.scholarpedia.org/article/Computational\_models\_of\_visual\_attention\#Evaluating\_a\_model$

[16] V. Navalpakkam, C. Koch, A. Rangel, and P. Perona, *Optimal Rewared Harvesting in Complex Perceptual Environments*, PNAS, vol. 107, No. 11, pp. 5232-5237, 2010