BAYESIAN LEARNING BASED VISUAL SALIENCY DETECTION

Jinxia Zhang, Jundi Ding, Chuancai Liu, Jingyu Yang

School of Computer Science and Technology
Nanjing University of Science and Technology, NJUST
Nanjing, China
jxzhang210094@gmail.com, dingjundi2010@mail.njust.edu.cn,
liu.ccnj@yahoo.com.cn, yangjy@mail.njust.edu.cn

Keywords: saliency, Bayesian learning, rarity, distinctiveness, central bias

Abstract

This paper is to present a Bayesian learning based framework for visual saliency detection in natural scenes. Especially, for any point in the scene, this framework has considered whether it is salient or not; but previous methods by Bayesian learning seem not to do so. This framework includes two steps. First, the framework indicates that visual saliency is constituted with three main saliency modules. In a free-viewing manner, these main saliency modules are rarity, distinctiveness and central bias. Second, they are non-linearly combined for the final saliency map by a regularized neural network. The experimental results on two fixation datasets indicate that our framework outperforms other representative methods.

1 Introduction

Human visual system can rapidly detect important items from a tremendous amount of visual information in our surrounding world. Just inspired by this biological discovery, Itti and Koch are the first to build up a visual saliency model [5]. This model computes the central-surround differences in different feature channels and generates a finally single topographical saliency map in a winner-take-all manner. Since then, the term *visual saliency* is widely known by many researchers and they have developed a great number of different saliency maps helpful for many tasks in computer vision, *e.g.,* object detection.

Overall, in a low-level perspective, the core of visual saliency is a bottom-up, stimulus-driven mechanism. This indicates that the location, item or object in the scene sufficiently different from its surroundings is easy to attract our visual attention. Note, for instance, a red flower in the green grassland will be visually salient and can immediately pop out. But, sometimes the bottom-up saliency may be strongly modulated by the top-down, task-driven factors. As an example, if we need to look for one person in a crowded environment who is wearing a black jacket, the top-down desire will surely make the black objects more salient than all others.

Most existing models focus on bottom-up saliency, where the subjects are free viewing a scene. Harel et al employ the graph method to compute the visual saliency [3]. Bruce and Tsotsos compute a visual saliency based on information maximization [1]. Hou et al present a method for visual saliency detection by extracting the spectral residual of an image in spectral domain [4]. Zhang et al use natural statistics to develop a Bayesian framework for visual saliency [9]. Recently, there have been some overall saliency models, where the subjects are asked to look for a specific object from scenes. Torralba et al propose a probabilistic model to search an object by global features [8].

We propose a different framework which classifies the points in the scene into two classes: the salient class and the non-salient class using Bayesian learning. This framework includes two steps. First, the framework indicates that visual saliency is constituted with three main saliency modules. In the free-viewing manner, these main modules are rarity, distinctiveness and central bias. And then a regularized neural network is used to non-linearly combine these modules. Our framework is most related to the methods in [9] and [8]. However, the main difference is as follows. The probability of interest in [9] is how probably a target is present at each point in the scene. And the probability in the work [8] is whether or not a target is present in the scene conditioning on global features. Whereas, the probability we are concerned with is whether or not a point belongs to the salient class. We have implemented the bottom-up saliency of our framework. And experimental results on two fixation datasets indicate that our framework outperforms other representative methods using the metric ROC.

The rest of this paper has been organized as follows: Section 2 presents the visual saliency detection framework based on Bayesian learning. Section 3 gives the implementation details. Section 4 demonstrates the promising results comparing with other representative saliency methods over two fixation data sets. And Section 5 concludes the paper.

2 The Visual Saliency Detection Framework

The goal of our framework is to solve a classification problem of two classes, which would classify each point in the static image to be the salient class or non-salient class. To achieve this goal, the framework will estimate the probability of each point belonging to the salient class given the visual features and the location in the image. And we call this probability the saliency degree of the point. In this section, we will introduce the framework of visual saliency detection.

Let x denote a point in the image. The point can be a pixel, a region or an object. Let the binary random variable S_x denote whether the point x belongs to a salient class or a non-salient

class. Let the random variable F denote the features of a point and the random variable L denote the location of a point. So the saliency degree SD_x of the point x can be formulated as the probability $p\left(S_x = 1 \mid F = f_x, L = l_x\right)$, where f_x represents the feature values of the point x and l_x represents the coordinate of the point x in the image. Then we can calculate the saliency degree of a point using Bayesian theorem:

$$SD_x = p\left(S_x = 1 \mid F = f_x, L = l_x\right) \quad (1)$$

$$= \frac{p\left(F = f_x, L = l_x \mid S_x = 1\right) p\left(S_x = 1\right)}{p\left(F = f_x, L = l_x\right)} \quad (2)$$

For simplicity, we assume that the features and location are independent and conditionally independent give $S_x = 1$. Then the formula above can be simplified as follows:

$$SD_x = \frac{p(F = f_x \mid S_x = 1) p(L = l_x \mid S_x = 1) p(S_x = 1)}{p(F = f_x) p(L = l_x)} \quad (3)$$

$$= \frac{p\left(F = f_x \mid S_x = 1\right)}{p(F = f_x)} \frac{p(L = l_x \mid S_x = 1) p(S_x = 1)}{p(L = l_x)} \quad (4)$$

$$= \frac{1}{p(F = f_x)} p(F = f_x \mid S_x = 1) p(S_x = 1 \mid L = l_x) \quad (5)$$

To compare the saliency degree among the points in an image, it is also effective to estimate the log probability because logarithm is a monotonically increasing function. Here we choose to use the log probability because it is much easier for computation. Then the saliency degree of a point can be redefined as follows:

$$SD_x = -\log p(F = f_x) + \log p(F = f_x \mid S_x = 1) + \log p(S_x = 1 \mid L = l_x) \quad (6)$$

The first term $-\log p(F = f_x)$ shows that the salient point has the property of "rarity". If the feature of a point has a small probability, the point would provide large amounts of information and therefore has large saliency degree. This term depends only on the features observed at the point and therefore is a pure bottom-up factor.

The second term $\log p(F = f_x \mid S_x = 1)$ is a module which can include both the stimulus-driven factor and the task-driven factor. For the bottom-up saliency, where the subjects are free viewing a scene, this term only considers the stimulus-driven factor. It represents the property of "distinctiveness" of the point. The visual receptive field has the architecture that visual neurons are most sensitive in the center, while stimuli presented in the surround would inhibit the neuronal response. Here we call the property which makes the point stand out from the surroundings "distinctiveness". Moreover, for overall saliency, where the subjects are asked to look for a specific object from scenes, this term would also include the task-driven factor besides the stimulus-driven factor and favor the feature values of the target object. And we call this property "consistency". That is, if the task is to find your friend wearing a black jacket, then this term will be much larger for a black point.

The third term $\log p(S_x = 1 \mid L = l_x)$ provides the location prior and is independent of the visual features. According to the work of Tatler et al [7], Judd et al [6] and Zhao et al [10], the central bias plays an important role to detect visual saliency. So we propose that when people viewing the scene freely, that is for bottom-up saliency, the location prior is the central position. And for overall saliency the location prior is where the target is likely to appear.

In the rest of this paper, we will concentrate on bottom-up saliency for static images where the observers are free-viewing the scene and no specific task is given. Then, the second term only shows the property of "distinctness" and the third term provides the central position as the location prior. So in a free-viewing manner, these three saliency modules are "rarity", "distinctness" and "central bias". We use a regularized neural network to non-linearly combine these three modules. Figure 1 shows the process of our visual saliency detection framework.

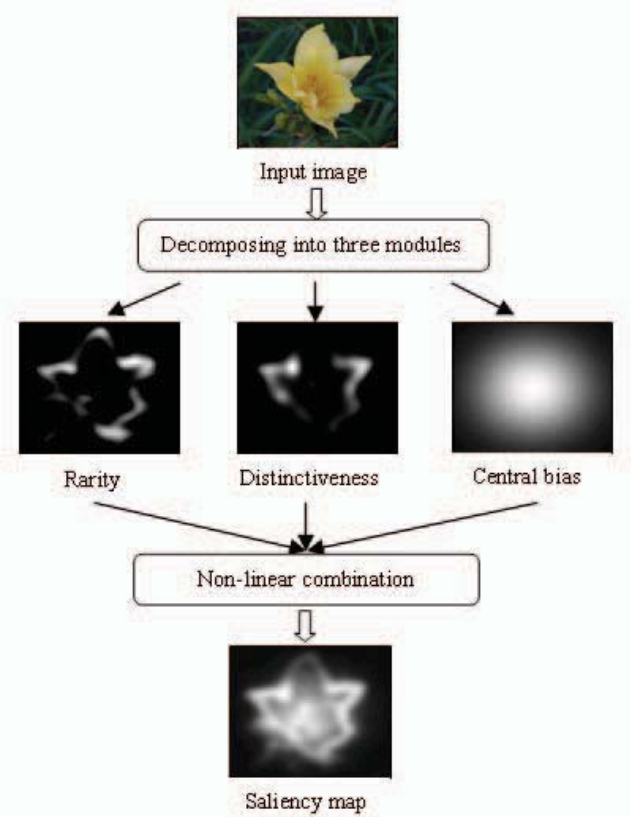

Figure 1: Process of our visual saliency detection framework.

3 Implementation Details

In this section, we will exposit the implementation details of three modules in the visual saliency detection framework and the non-linearly weighted combination method.

3.1 Three modules

"Rarity": The module "rarity" is implemented by computing the self-information of each point in the image across different scales in three feature channels color, intensity and orientation. We use the similar method to work [5] to build the Gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, $Y(\sigma)$ and $I(\sigma)$. And the orientation information is obtained using the steerable pyramids $O(\sigma, \theta)$, where $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ represents the preferred orientation. In all the nine pyramids, we choose four scales for further computation and let $\sigma \in [0, 1, 2, 3]$ represent the scale of the input image.

In every sub-image, we calculate the self-information of the feature value for each point. The specific process is as follows:

first we use a histogram of 100 bins to compute the frequency in every bin for the sub-image. After that, we use the interpolation to compute the probability of the feature value for each point in the sub-image. Then the self-information for a point is calculated in the sub-image using the formula: $h(x) = -\log p(x)$. Finally, for every pyramid the images of different scales are averagely combined into a feature map.

"Distinctiveness": There are many methods to simulate the property of "distinctiveness". Itti et al [5] uses the center-surround operator. Goferman et al [2] uses the dissimilarity between different patches to simulate this property. In our implementation, we model the property "distinctness" using graph-based visual saliency [3].

"Central bias" : When humans take pictures, they naturally frame an object of interest in the central position of the image. And when they look at an image freely, the fixation point often starts from the central position of the image. This is often assumed to arise because of the central bias. As a result, we propose that the central position is the location prior when there is no specified task.

We use the radial basis function to model the central bias. Then the central bias function is a real-valued function whose value depends only on the distance between a point and the center point. The value of the function is small when the point is far from the central point. There are different types of radial basis functions including Gaussian kernel function, inverse quadratic function and so on. In this paper, we use the Gaussian kernel function. So the central bias function is:

$$C(x, y) = \exp\left(-\frac{(x - x_c)^2 + (y - y_c)^2}{2\delta^2}\right) \quad (7)$$

In this function, (x, y) denotes the coordination of a random point in the image and (x_c, y_c) denotes the coordination of the center point. δ is the scale parameter. Different δ corresponds to different central bias modulation. We can learn the best scale parameter δ from the fixation data set.

3.2 The Weighted Combination Method

The neurons in brain are highly non-linear and complex devices. To mimic this property of neurons, we further non-linearly combine different modules based on regularized neural network.

In this paper, we use the network of three layers: one input layer, one hidden layer and one output layer. The number of nodes in the input layer corresponds to the number of features of a point. In the fixation data set there are hundreds of images and in every image there are thousands of points. Therefore the number of training sample points can be very large. But the number of features in every point is small. To achieve a good result, the number of nodes in hidden layer must be large. So we specify that the number of nodes for the hidden layer is seventy in our implementation. We use the back-propagation algorithm to learn the optimal model. To avoid the problem of over fitting, we add a regularized term in the neural network. And we use a validation set to choose the parameter of the regularized term.

4 Experimental Comparisons

We implement the bottom-up saliency of our framework and compare it with three other visual saliency detection methods over two fixation data sets: MIT and Toronto data sets [1, 6].

4.1 MIT data set

We have evaluated our method over the MIT data set which records fixations from 15 viewers who freely viewed 1003 natural indoor and outdoor images. To achieve more precise result, we have 3 trials and obtain the averaged performance. For the non-linear combination method regularized neural network, we randomly choose 100 training images, 100 validation images and 200 testing images. The validation images are used to choose the optimal parameter for the regularized term.

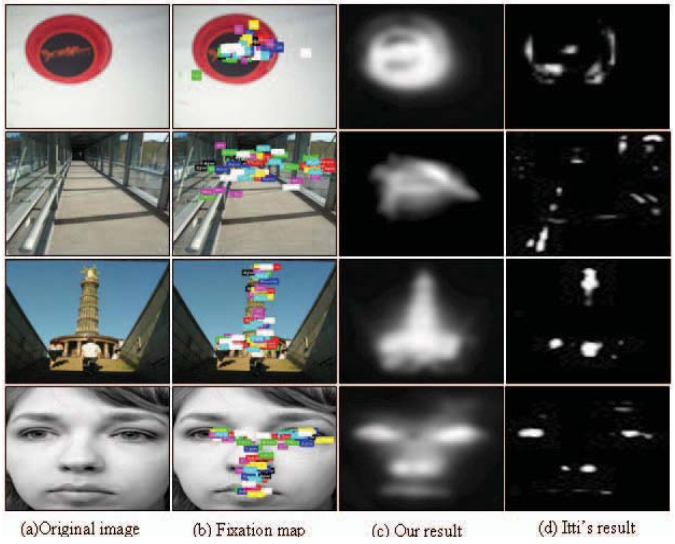

Figure 2. Some examples for visual comparison.

Some examples of our method are shown in Figure 2. For comparison, the saliency maps generated by Itti et al and the fixation maps are also included. Each row contains, from left to right: the original image; the fixation map which records fixations from 15 viewers; the saliency map generated by our method and the saliency map generated by Itti's saliency model. From Figure 2, we can make the following observations: Itti's saliency map is based on the central-surround operator and the salient points often locate near the edges. But the salient points of our method are not only over the edges but also locate inside an object, which are more consistent with the fixation points.

We have chosen three state-of-the-art methods for quantitative comparisons. These saliency methods are IT, SUN and GB which represent the methods of Itti et al, Zhang et al and Harel et al respectively. Because we detect visual saliency based on Bayesian learning, we call our method BL for short. And the quantitative comparison results are showed in Figure 3.

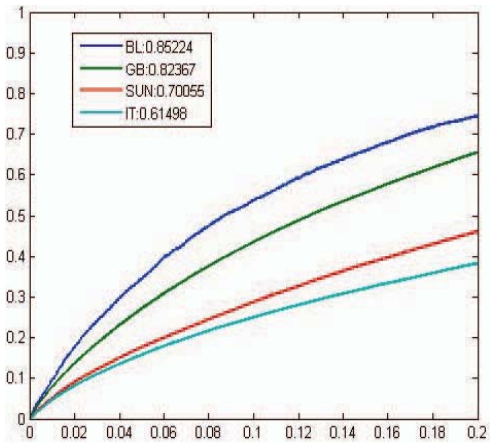

Figure 3. Quantitative comparisons with other visual saliency detection methods over MIT Data Set.

4.2 Toronto data set

We also compare the performances of different methods over Toronto data set to confirm the verdict. This data set contains data from 11 viewers who freely viewed 120 colour images of outdoor and indoor scenes. For the non-linear combination method regularized neural network, we randomly choose 30 training images, 30 validation images and 30 testing images. We also have 3 trials for every model and obtain the averaged performance.

The performances of our method comparing with three other state-of-the-art methods over Toronto data set are included in Figure 4.

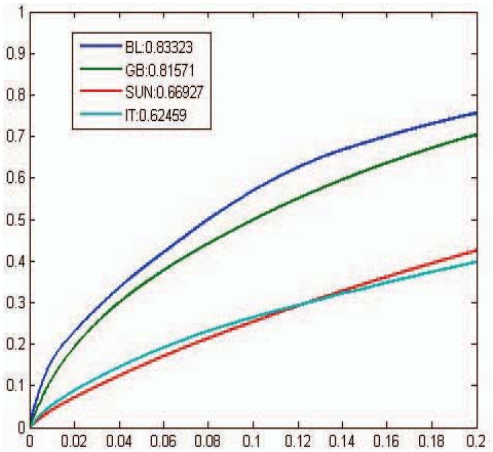

Figure 4. Quantitative comparisons with other visual saliency detection methods over Toronto Data Set.

5 Conclusion and Future Work

We propose a visual saliency detection framework based on Bayesian learning to compute the saliency degree at each point in the scene. From the framework, we get three important modules for visual saliency detection. And we use regularized neural network to non-linearly combine different modules. The experimental comparisons demonstrate the promising result of the bottom-up saliency in our framework. In the future, we plan to implement the overall saliency and generalize our method to detect the visual saliency in dynamic videos.

Acknowledgements

The research was supported by the National Natural Science Fund of China (Grant Nos. 90820306, 60632050, 9082004, 61103058), the basic key technology project of Ministry of Industry and Information Technology of China (Grant No. E0310/1112/JC01) and the National 863 Project (Grant No. 2006AA04Z238).

References

[1] N. D. B. Bruce and J. K. Tsotsos. “Saliency, attention and visual search: an information theoretic approach”, *Journal of Vision*, vol. 9, pp. 1-24, (2009).

[2] S. Goferman, L. Zelnik-Manor, and A. Tal. “Context-aware saliency detection”, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010, p. 8.

[3] J. Harel, C. Koch, and P. Perona. “Graph-based visual saliency,” in *Proc. Neural Information Processing Systems*, 2006, p. 8.

[4] X. Hou and L. Zhang, “Saliency detection: a spectral residual approach,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2007, no. 800, p. 8.

[5] 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*, vol. 22, no. 11, pp. 1254-1260, (1998).

[6] T. Judd, K. Ehinger, F. Durand, and A. Torralba. “Learning to predict where humans look”, in *Proc. IEEE International Conference on Computer Vision*, 2009, p. 8.

[7] B. W. Tatler. “The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions,” *Journal of Vision*, vol. 7, pp. 1-17, (2007).

[8] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. “Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search”, *Psychological Review*, vol. 113, no. 4, pp. 766-86, (2006).

[9] L. Zhang, M. H. Tong, T. K. Marks, and G. W. Cottrell, “SUN: a Bayesian framework for saliency using natural statistics” , *Journal of Vison*, vol. 0, pp. 1-20, (2008).

[10] Q. Zhao and C. Koch. “Learning a saliency map using fixated locations in natural scenes”, *Journal of Vision*, vol. 11, pp. 1-15, (2011).