

# Adaptive Semantic Bayesian Framework for Image Attention

Wei Zhang, Q. M. Jonathan Wu, and Guanghui Wang

Department of Electrical and Computer Engineering, University of Windsor  
weizhang1216@hotmail.com

## Abstract

Image attention is the basic technique for many computer vision applications. In this paper, we propose an adaptive Bayesian framework to detect the image attention in color image. Firstly, three simple semantics and subtractive clustering are used to construct Attention Gaussians Mixture Model (AGMM) and Background Gaussians Mixture Model (BGMM). Secondly, the Bayesian framework is utilized to classify each pixel into attention objects and background objects. Thirdly, EM algorithm is used to update the parameters of AGMM, BGMM, and Bayesian framework according to the detection results. Finally, the above classification and update procedures are repeated until the detection results become steady. Experimental results on typical images exhibit the robustness of the proposed method.

## 1. Introduction

Image attention (visual attention) detects the salient region in an image and is the basic technique used in many computer vision applications, such as video compression, advertisement design, visual tracking, and image display. As an interdisciplinary subject of physiology, visual psychology, and visual neural system, the challenge of image attention detection rests with the subjective aspects of the problem.

The salient region is commonly detected based on its center-surround difference [1] in variety of feature spaces (e.g. color, intensity, and texture). In principle, image attention is a problem of image classification. According to the architecture employed, existed approaches on image attention can be classified into two categories as bottom-up and top-down.

The bottom-up methods commonly evaluate each pixel's saliency by its center-surround difference [1][4][7]. The top-down methods are always task-derived, in which semantics are popularly added to accomplish specific aims [2][6][9][10]. Recent researches demonstrate that the combination of the two kinds of methods produce much better results [3][5][8].

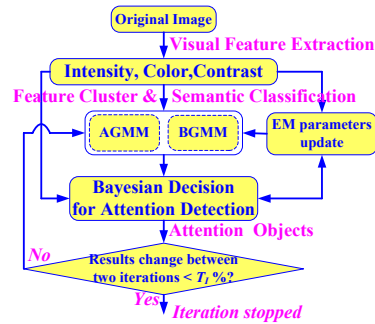


Fig.1. Basic workflow of the proposed method.

In essence, bottom-up approaches demonstrate the property of human eyes neural systems [1], while top-down approaches exhibit the prior knowledge of human recognition system. However, the semantics used in the top-down models are commonly object-based, and the implementation of semantics highly depends on image segmentation, which is not only complicated but also computational expensive [2].

In this paper, we combine the bottom-up and top-down methods, and propose an adaptive Bayesian framework to detect the image attention, which is a four-fold method. Firstly, three simple semantics and subtractive clustering are used to construct Attention Gaussians Mixture Model (AGMM) and Background Gaussians Mixture Model (BGMM), and to obtain the prior knowledge of the Bayesian framework. Secondly, the Bayesian framework is utilized to classify each pixel into attention objects and background objects. Thirdly, EM algorithm is used to update the parameters of AGMM, BGMM, and Bayesian framework according to the detection results. Finally, the above procedures are repeated until the detection results become steady. In this way, the image semantic information is extracted without accuracy image segmentation. Figure 1 illustrates the basic workflow of the proposed method.

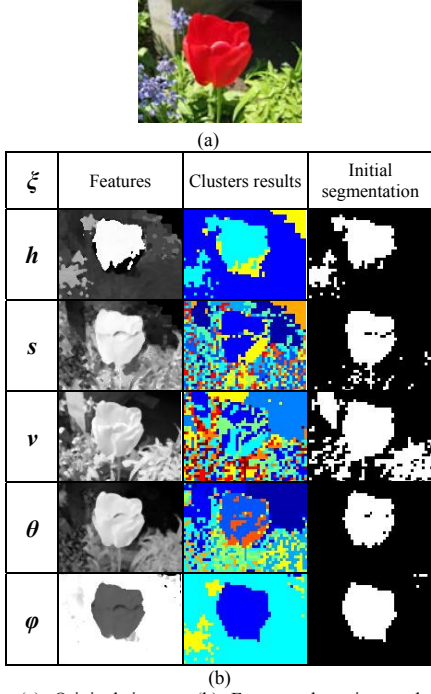


Fig.2. (a) Original image, (b) Feature clustering and initial segmentation.

## 2. Feature Extraction and Clustering

Let the original image be  $I$ , we utilize five channels of features in our framework: intensity, color, and contrast, which are described as follows.

-*Intensity*: the Value component of the HSV color space, let it be  $v$ .

-*Color*: the Hue and Saturation components of the HSV color space, let them be  $h$  and  $s$ .

-*Contrast*: the C and H components of CIE-LCH color space, let them be  $\theta$  and  $\varphi$ .

For each of the above features, we employ subtractive clustering to make an initial segmentation. Subtractive clustering does not require the number of clusters to be specified and is fit to determine the number of clusters for a given set of data. Let the clustering results of feature  $\zeta$  be  $C_1^\zeta, C_2^\zeta, \dots, C_{n_\zeta}^\zeta$ . Here  $n_\zeta$  is the number of the clusters of feature  $\zeta$ , and  $C_i^\zeta$  is binary image with the pixels being 1 for the  $i$ -th cluster of feature  $\zeta$ .

An example of above feature and clustering results are given in Fig. 2, where different clusters are denoted by different colors.

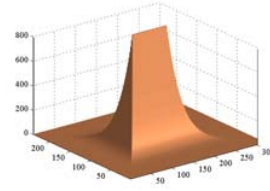


Fig.3. Attention weight map.

## 3. Bayesian Framework for Image Attention

### 3.1. Semantic-based initialization

For feature  $\zeta$ , we model its clusters  $C_i^\zeta, i=1, 2, \dots, n_\zeta$ , as an AGMM and a BGMM by using three semantics, which are listed as follows.

*Semantic 1*: The pixel close to the center of the image has higher possibility to be attended. Therefore, an attention weight map  $\Omega$  is defined as follows.

$$\begin{aligned} \Delta(x, y) &= \min(W - x, x, L - y, y); \\ \Omega(x, y) &= [\Delta(x, y) / M(\Delta)]^r; \end{aligned} \quad (1)$$

where  $M(\cdot)$  represents the mean value operation,  $W$  and  $L$  are the width and length of the image, respectively. We illustrate above attention weight map in Fig.3.

*Semantic 2*: The possibility of one cluster to be attended decreases with the area of the cluster, while small area of cluster cannot be attended. Thus we define area attention value  $\Phi_i^\zeta$  of cluster  $C_i^\zeta$  as follows.

$$\Phi_i^\zeta = \begin{cases} (\Psi_i^\zeta)^{-1} & \text{if } \Psi_i^\zeta > T_A; \\ 0 & \text{otherwise} \end{cases}; \quad \Psi_i^\zeta = \sum_{(x,y)} C_i^\zeta(x, y) \quad (2)$$

where  $\Psi_i^\zeta$  is the area of cluster  $C_i^\zeta$ .

*Semantic 3*: The absolute difference between the average value of the cluster and the average value of the feature is another useful quantity, and is defined as follows.

$$\Gamma_i^\zeta = \left| \frac{1}{\Psi_i^\zeta} \cdot \sum_{C_i^\zeta(x,y)=1} \zeta(x, y) - M(\zeta) \right| \quad (3)$$

With the above semantics, we can obtain a combined saliency value  $\Theta_i^\zeta$  for cluster  $C_i^\zeta$  as follows.

$$\Theta_i^\zeta = \Phi_i^\zeta \cdot \Gamma_i^\zeta \cdot \sum_{C_i^\zeta(x,y)=1} \Omega(x, y) \quad (4)$$

$$\zeta = (v, h, s, \theta, \varphi), i = 1, 2, \dots, n_\zeta;$$

By utilizing a threshold  $T_\Theta$  on  $\Theta_i^\zeta, i=1, 2, \dots, n_\zeta$ , we can classify  $C_i^\zeta$  into attention clusters and background clusters, which are models as AGMM and BGMM. Let

them be  $\sum_{i: C_i^\xi \in AGMM} \eta_{\xi, A, i} [\xi(x, y), \mu_{\xi, A, i}, \sigma_{\xi, A, i}^2]$

and  $\sum_{i: C_i^\xi \in BGMM} \eta_{\xi, B, i} [\xi(x, y), \mu_{\xi, B, i}, \sigma_{\xi, B, i}^2]$ , respectively:

$$\eta_{\xi, \lambda, i} [\xi(x, y), \mu_{\xi, \lambda, i}, \sigma_{\xi, \lambda, i}^2] = \frac{1}{(2\pi)^{1/2} \sigma_{\xi, \lambda, i}} e^{-\frac{1}{2\sigma_{\xi, \lambda, i}^2} [\xi(x, y) - \mu_{\xi, \lambda, i}]^2} \quad (5)$$

$$\xi = (v, h, s, \theta, \varphi); \quad \lambda = A, B$$

where  $\mu_{\xi, \lambda, i}$  and  $\sigma_{\xi, \lambda, i}$  are computed according to the corresponding clusters as follows.

$$\mu_{\xi, \lambda, i} = \frac{1}{\Psi_i^\xi} \cdot \sum_{C_i^\xi(x, y)=1} \xi(x, y);$$

$$\sigma_{\xi, \lambda, i}^2 = \frac{1}{\Psi_i^\xi} \cdot \sum_{C_i^\xi(x, y)=1} [\xi(x, y) - \mu_{\xi, \lambda, i}]^2; \quad (6)$$

$$i = 1 \dots n_\xi; \quad \xi = (v, h, s, \theta, \varphi); \quad \lambda = \begin{cases} A, & \text{if } C_i^\xi \in AGMM \\ B, & \text{otherwise} \end{cases};$$

Note that the AGMM and BGMM are modeled for every feature.

### 3.2. Attention detection and model update

We incorporate the above AGMM and BGMM into a Bayesian framework to detect image attention. Let the pixels be partitioned into two classes:  $\alpha$  (attention object) and  $\beta$  (background object). For pixel  $I(x, y)$ , the conditional probability that it belongs to the attention object is given by:

$$P(\alpha | I(x, y)) = \frac{P[I(x, y) | \alpha]P(\alpha)}{P[I(x, y) | \alpha]P(\alpha) + P[I(x, y) | \beta]P(\beta)} \quad (7)$$

The corresponding pixel is classified as the attention object if the following decision rule is satisfied:

$$P[I(x, y) | \alpha]P(\alpha)\Omega(x, y) > P[I(x, y) | \beta]P(\beta); \quad (8)$$

Different with the traditional Bayesian decision rule, the attention weight map is included in Eq.(8).  $P(\alpha)$  and  $P(\beta)$  are initialized according to the clustering results of each feature:

$$P(\alpha) = \frac{1}{5WL} \cdot \sum_{\xi=[v, h, s, \theta, \varphi]} \left[ \sum_{i: C_i^\xi \in AGMM} \Psi_i^\xi \right]; P(\beta) = 1 - P(\alpha); \quad (9)$$

Assuming that the features,  $v, h, s, \theta, \varphi$ , are identical and independently distributed (i.i.d), we have:

$$P[I(x, y) | \gamma] = \prod_{\xi=[v, h, s, \theta, \varphi]} P[\xi(x, y) | \gamma], \gamma = \alpha, \beta; \quad (10)$$

In every channel of feature, one object is always attended because of its highest fitness with the AGMM of the feature channel. Therefore, we fit  $\xi(x, y)$  with the AGMM and BGMM, and the conditional probability of  $\xi(x, y)$  is calculated as the maximum resultant value:

$$P[\xi(x, y) | \alpha] = \text{Max}_{i: C_i^\xi \in AGMM} \eta_{\xi, A, i} [\xi(x, y), \mu_{\xi, A, i}, \sigma_{\xi, A, i}^2],$$

$$P[\xi(x, y) | \beta] = \text{Max}_{i: C_i^\xi \in BGMM} \eta_{\xi, B, i} [\xi(x, y), \mu_{\xi, B, i}, \sigma_{\xi, B, i}^2], \quad (11)$$

$$\xi = [v, h, s, \theta, \varphi];$$

EM algorithm [11] is commonly used for parameter estimation and is included in our framework to update AGMM and BGMM. If pixel  $I(x, y)$  is classified to be attention object, the AGMM is updated according to Eq. (12) and the parameters of BGMM keep unchanged.

$$\mu_{\xi, A, j} = (1 - \rho_{\xi, A, j})\mu_{\xi, A, j} + \rho_{\xi, A, j}\xi(x, y)$$

$$\sigma_{\xi, A, j}^2 = (1 - \rho_{\xi, A, j})\sigma_{\xi, A, j}^2 + \rho_{\xi, A, j} [\xi(x, y) - \mu_{\xi, A, j}]^2 \quad (12)$$

$$\rho_{\xi, A, j} = \delta \cdot \eta[\xi(x, y), \mu_{\xi, A, j}, \sigma_{\xi, A, j}^2]$$

$$i = \text{argmax}_{i: C_i^\xi \in AGMM} \{ \eta_{\xi, A, i} [\xi(x, y), \mu_{\xi, A, i}, \sigma_{\xi, A, i}^2] \}$$

$$\xi = (v, h, s, \theta, \varphi)$$

where  $\delta$  controls the speed of updating.

After the attention objects are detected in the entire image, and let the binary image of the detected attention image be  $R$ , the Bayesian framework is then updated as follows.

$$P(\alpha) = \frac{1}{WL} \sum_{(x, y)} R(x, y); P(\beta) = 1 - P(\alpha); \quad (13)$$

The above detection and updating are repeated until the change of the detection results between two iterations is less than  $T_I\%$ . In experiments, the iteration number is always 2~5 when  $T_I$  is set to 1.

### 3.3. Post-processing

After the attention objects are detected, we adopt a post-processing to further improve the detection results as follows. Firstly, a morphological close operation is implemented on the  $R$ ; secondly, a flood-fill operation is applied because background region should not be in the interior of attention objects in most situations; thirdly, the small area of attention objects is removed.  $R$  is updated correspondingly after the above procedures.

## 4. Experimental Results

The proposed method has been tested on typical images to evaluate its effectiveness. The MSRA Salient Object Database<sup>1</sup> is used in our experiments. The database contains 5000 high quality images. Some of the detection results are illustrated in Fig. 4. It can be seen that the proposed method can detect the attention objects correctly.

The performance of the proposed method is also evaluated quantitatively to get a systematic evaluation. The ground-truth of the image attention is manually labeled by nine different volunteers [7], and let it be  $G$ .

<sup>1</sup> Available from [http://research.microsoft.com/~jiansun/SalientObject/salient\\_object.htm](http://research.microsoft.com/~jiansun/SalientObject/salient_object.htm)

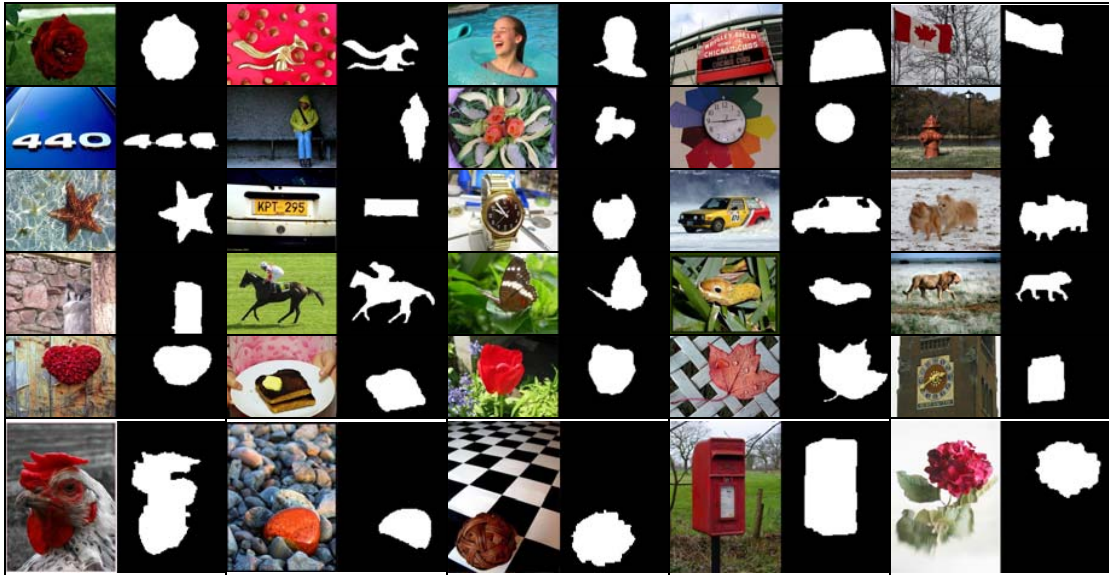


Fig. 4. The experimental results of the proposed method.

$G$  is binary image with the pixels being 1 for the labeled attention region. In our experiments, we select the recall-precision rate for quantitative evaluation, which are defined as follows.

$$Precision = \frac{\sum_{G(x,y)=1} R(x,y)}{\sum_{(x,y)} R(x,y)}; Recall = \frac{\sum_{G(x,y)=1} R(x,y)}{\sum_{(x,y)} G(x,y)}; \quad (14)$$

Furthermore, we define the *Overall* detection rate as the sum of *Recall* and *Precision*. The results of the proposed method are given in Table 1 together with the results of [1][7][10] for comparison. We can see in Table 1 that the proposed method gets the highest *Precision* rate, a lower *Recall* rate than [10], and the highest *Overall* detection rate.

Table 1. Quantitative evaluation of the proposed method.

	[1]	[7]	[10]	Proposed
<i>Precision</i>	0.72	0.843	0.632	0.85
<i>Recall</i>	0.783	0.778	0.905	0.794
<i>Overall</i>	1.503	1.621	1.537	1.644

## 5. Conclusions

In this paper we have proposed an adaptive Bayesian framework to detect image attention. The top-down and bottom-up models are combined. By utilizing a simple image cluster, the semantics are extracted and an initial segmentation is achieved, which is used as the prior-knowledge of the Bayesian framework. An attention Gaussian mixture model and background Gaussian mixture model are maintained, and the pixel is classified by using the Bayesian classifier. Experimental results and quantitative

evaluation exhibit the robustness of the proposed method.

## References

- [1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, vol. 20, no. 11, pp. 1254-1259, 1998.
- [2] Z. Yu and H-S. Wong, "A Rule Based Technique for Extraction of Visual Attention Regions Based on Real-Time Clustering," *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 766-784, 2007.
- [3] A. Levin and Y. Weiss, "Learning to combine bottom-up and top-down segmentation," *ECCV*, pp. 581-594, 2006.
- [4] O. L. Meur, O. L. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *PAMI*, vol. 28, pp. 802-817, 2006.
- [5] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," *CVPR*, pp. 2049-2056, 2006.
- [6] J. Han, K. N. Ngan, M. J. Li, and Hong-Jiang Zhang, "Unsupervised extraction of visual attention objects in color images," *CSVT*, vol. 16, pp. 141-145, 2006.
- [7] T. Liu, J. Sun, N-N. Zheng, X. Tang, and H-Y. Shum, "Learning to Detect A Salient Object," *CVPR*, pp. 1-8, 2007.
- [8] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," *CVPR*, pp. 1-8, 2007.
- [9] A. Oliva, A. Torralba, M. S. Castelhana, and J. M. Henderson, "Top-down control of visual attention in object detection," *ICIP*, vol. 1, pp.253-256, 2003.
- [10] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," *ACMM*, pp. 374-381, 2003.
- [11] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc., Ser. B*, vol. 39, pp.1-38, 1977.