LETTER

A Novel Bayes' Theorem-Based Saliency Detection Model

Xin HE[†], **Huiyun JING**[†], **Qi HAN**[†], *Nonmembers*, *and* **Xiamu NIU**[†a)], *Member*

SUMMARY We propose a novel saliency detection model based on Bayes' theorem. The model integrates the two parts of Bayes' equation to measure saliency, each part of which was considered separately in the previous models. The proposed model measures saliency by computing local kernel density estimation of features in the center-surround region and global kernel density estimation of features at each pixel across the whole image. Under the proposed model, a saliency detection method is presented that extracts DCT (Discrete Cosine Transform) magnitude of local region around each pixel as the feature. Experiments show that the proposed model not only performs competitively on psychological patterns and better than the current state-of-the-art models on human visual fixation data, but also is robust against signal uncertainty.

key words: *visual attention, saliency map, Bayes' theorem, kernel density estimation*

1. Introduction

The Human vision system rapidly detects important or interesting parts of images or videos to reduce the computational complexity. It is well known that visual saliency plays an important role that makes important or interesting parts stand/pop out from their surrounding and drives our perceptual attention. Thus, saliency detection is an indispensable component in many theories of visual attention. Applications of saliency detection have been reported in many fields such as object detection [1], image cropping [2], image browsing [3], and image/video compression [4].

Saliency has two major categories: bottom-up saliency and top-down saliency. The bottom-up saliency approach refers to mechanisms which are generally fast, stimulus-driven and independent of the knowledge in the scene, whereas the top-down saliency approach refers to mechanisms which are slow, goal-oriented and require the prior knowledge.

There have been many studies focusing on the bottom-up saliency detection over the last two decades. Based on the feature integration theory [5], Itti et al. [6] presented a saliency-based computational model for scene analysis. In Itti et al.'s work, visual input is first decomposed into several multi-scale feature maps. Followed by a center-surround operation, early visual features are extracted in parallel through linear filtering for the three types: intensity, color and orientation. All feature maps are then combined to produce a single saliency map. Bruce and Tsotsos [7] proposed a bottom-up model of overt attention based on the principle of maximizing information sampled from a scene. In their work, saliency is determined by quantifying the self-information of each local image patch after ICA (Independent Component Analysis) decomposition. Zhang et al. [8] proposed a similar scheme based on a Bayesian framework, which defines bottom-up saliency as the self-information of visual features and overall saliency (incorporating top-down information with bottom-up saliency) as the pointwise mutual information between the visual features and the desired target. Gao et al. [9] proposed a discriminant saliency model based on the center-surround mechanism, which measures saliency as the discriminant power of a set of features with respect to the classification problem that opposes stimuli at center and surround. Recently, Seo and Milanfar [10] presented nonparametric saliency detection approach based on LSK (Local Steering Kernel) features and utilized self-resemblance mechanism to compute saliency map.

Without full consideration of Bayes' theorem, previous works [8], [10] utilized different parts of Bayes' equation to measure saliency. In this letter, we integrate the two parts of Bayes' equation for better performance. We model saliency as a function of local kernel density estimation of features in the center-surround region and global kernel density estimation of features at each pixel across the whole image. This model not only considers local statistics of features, but also global statistics of features.

2. Proposed Saliency Detection Model

Motivated by the approach in [8], [9] and [10], we measure saliency of each pixel by Bayes' theorem. Firstly, representing saliency of each pixel i under the feature F_i and the location L_i as a binary random variable, we define binary random variables $\{y_i{}_{i=1}^{M}\}$ as follows

$$y_i = \begin{cases} 1, & \text{if pixel } i \text{ is salient,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $i = 1, \ldots, M$ and M is the total number of pixels in the image.

Thus, the saliency of a pixel i is defined as a posterior probability $Pr(y_i = 1|F, L)$ as follows

$$S_i = Pr(y_i = 1|F_i, L_i) \tag{2}$$

where $F_i = [f_i^1, f_i^2, \ldots, f_i^K]$ contains a set of features $\{f_{ik=1}^{K}\}$

Manuscript received June 10, 2011.
Manuscript revised July 23, 2011.

[†]The authors are with the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.
a) E-mail: xm.niu@hit.edu.cn

DOI: 10.1587/transinf.E94.D.2545

extracted from the local neighborhood of the corresponding pixel, K is the number of features in that neighborhood and L_i represents the pixel coordinates.

Equation (2) can be rewritten using Bayes' rule:

$$S_i = Pr(y_i = 1|F_i, L_i) = \frac{p(F_i, L_i|y_i = 1)Pr(y_i = 1)}{p(F_i, L_i)} \quad (3)$$

Inspired by Zhang et al. [8], we also assume that the feature and location are independent and conditionally independent given $y_i = 1$, which implies that location doesn't depend on the distribution of the features.

$$p(F_i, L_i) = p(F_i)p(L_i) \quad (4)$$

$$p(F_i, L_i|y_i = 1) = p(F_i|y_i = 1)p(L_i|y_i = 1) \quad (5)$$

Then (3) is simplified as follows

$$\begin{aligned} S_i &= \frac{p(F_i, L_i|y_i = 1)Pr(y_i = 1)}{p(F_i, L_i)} \\ &= \frac{p(F_i|y_i = 1)}{p(F_i)} \frac{p(L_i|y_i = 1)Pr(y_i = 1)}{p(L_i)} \\ &= \frac{1}{p(F_i)} p(F_i|y_i = 1)Pr(y_i = 1|L_i) \end{aligned} \quad (6)$$

We assume that under location prior, $Pr(y_i = 1|L)$ is equal to be salient and is omitted for simplicity. Equation (6) can be rewritten as follows

$$S_i = \frac{1}{p(F_i)} p(F_i|y_i = 1) \quad (7)$$

$p(F_i)$ depends on the visual features and implies that the feature of less probability seems to have higher saliency. In Seo et al. [10], $p(F_i)$ is considered uniform over features. In Bruce et al. [7] and Zhang et al. [8], $p(F_i)$ is used to detect saliency, where F_i is the feature vector and the features are calculated as the responses to filters learned from natural images. Different from Bruce et al. [7] and Zhang et al. [8], we directly calculate $p(F_i)$ using normalization kernel density estimation for F_i. Then we obtain Eq. (8).

$$\frac{1}{p(F_i)} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{M} \kappa(F_i - F_j)}{\sum_{j=1}^{M} \kappa(F_i - F_j)} \quad (8)$$

where κ is the kernel density function and M is the total pixels number of the image.

In Zhang et al. [8], $p(F_i|y_i = 1)$ of Eq. (7) is considered with knowledge of the target and is not used when calculating saliency. However, Seo et al. [10] adopt local "self-resemblance" measure to calculate $p(F|y_i = 1)$ using nonparametric kernel density estimation. Similar to Seo et al. [10], we make a hypothesis that $y_i = 1$ of the center pixel i in the center-surround region. It means that F_i is the only sampled feature in the center-surround features' space. Under this hypothesis, we estimate all $F = [F_1, F_2, \ldots, F_N]$ including Fi using kernel density estimation in the center-surround region where F is a feature set containing features from the center and surrounding region and N is the number of pixels in the center-surround region. Then we normalize $p(F_i|y_i = 1)$ under the hypothesis of $y_i = 1$.

$$\begin{aligned} p(F_i|y_i = 1) &= \frac{\kappa(F_i - F_i)}{\sum_{j=1}^{N} \kappa(F_i - F_j)} \\ &= \frac{1}{\sum_{j=1}^{N} \kappa(F_i - F_j)} \end{aligned} \quad (9)$$

Now we rewrite Eq. (7) using Eqs. (8) and (9) and obtain the saliency formula of each pixel

$$S_i = \frac{\sum_{i=1}^{M} \sum_{j=1}^{M} \kappa(F_i - F_j)}{\sum_{j=1}^{M} \kappa(F_i - F_j)} \frac{1}{\sum_{j=1}^{N} \kappa(F_i - F_j)} \quad (10)$$

Equation (10) could be represented as follows

$$S_i = \frac{K_{local}(F_i)}{K_{global}(F_i)} \quad (11)$$

where $K_{local}(F_i)$ represents normalization kernel density estimation in the local center-surround region and $K_{global}(F_i)$ represents normalization kernel density estimation across the whole image. Thus, we define pixel saliency using local and global kernel density estimation of features of the corresponding pixel.

3. Implementation

We only extract DCT (Discrete Cosine Transform) magnitude of local region around each pixel as the feature to calculate $K_{local}(F)$ and $K_{global}(F)$, which represents the energy information of local region in the image. DCT has been widely applied in many fields related to human vision system, such as image/video compression, content-based image retrieval and indexing, and image/video watermarking. The implementation could be detailed by the following steps:

1. For the input image of M pixels, 2D-DCT operates on 3×3 block around pixel $\mathbf{i}$. The absolute values of DCT coefficients are considered as the feature of each pixel, $F_i = [f_i^1, f_i^2, \ldots, f_i^9]$. i is the i^{th} pixel. For the sake of robustness, we ignore f_i^7, f_i^8 and f_i^9, and use $F_i = [f_i^1 f_i^2, \ldots, f_i^6]$ to represent the feature.

2. Calculating $K_{global}(F_i)$ is usually time-consuming. In order to speed up the operation, we perform a kernel density estimation using a Gaussian kernel with the rule-of-thumb bandwidth†.

3. We choose 7×7 window around pixel $\mathbf{i}$ as the center-surround region of pixel $\mathbf{i}$. Each window has 49 features which is represented as $F_1, F_2, \ldots, F_{49}$. An adaptive kernel $G(\cdot)$ [10] and "Matrix Cosine Similarity" [10], [13] are used to calculate $p(F_i|y_i = 1)$ effectively. Then we use the following equation to calculate $K_{local}(F_i)$

$$K_{local}(F_i) = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(F_i,F_j)}{\sigma^2}\right)} \quad (12)$$

†We use the Kernel Density Estimation Toolbox for Matlab provided by Alexander Ihler (available at http://www.ics.uci.edu/~ihler/code/kde.html).

where $\rho(F_i, F_j)$ is equal to $trace\left(\frac{F_i^T F_j}{\|F_i\|\|F_j\|}\right)$ and σ is set to 0.007 as default value.

4. By step 2 and step 3, we obtain $K_{global}(F_i)$ and $K_{local}(F_i)$. Then saliency of each pixel is calculated using Eq. (11).

4. Experimental Results

4.1 Psychological Patterns

We tested our model on psychological patterns, which are widely used in attention experiments not only to explore the mechanism of visual search but also to test the effectiveness of saliency map. We used 8 patterns to test our method, including density, curvature, color (hue), intersection, length, intensity, number, orientation and terminators pattern. The results are shown in Fig. 1. The experimental results show that our model performs competitively on psychological patterns.

4.2 Visual Fixation Data

We evaluated our model on human visual fixation data from natural images. The dataset we used was collected by Bruce and Tsotsos [7] as the benchmark dataset for comparing human eye predictions between methods. The dataset contains eye fixation data from 20 subjects for a total of 120 natural images. The Kullback-Leibler divergence (KLD) and the area under receiver operating characteristic (AUC) were computed as performance metrics. A high value of two metrics means better performance. In Zhang et al. [8], Zhang et al. noted that the original KL divergence and ROC area measurement are corrupted by an edge effect which yielding artificially high results. For eliminating border effects, we adopt the same procedure described by Zhang et al. [8] to measure KL divergence and ROC area. We compared our method against state-of-the-art methods including Itti et al. [6], Bruce et al. [7], Gao et al. [9], Zhang et al. [8], and Seo et al. [10]. The mean and the standard error are reported in Table 1. Our model performs better than the current state-of-the-art models in KLD and AUC metrics.

Limited by space, we only present some examples of visual results of our model compared with Seo et al. [10] and Bruce et al. [7] in Fig. 2. Visually, our model also exceeds the other two models in term of accuracy.

Table 1 Experimental results.

Model	KLD(SE)	AUC(SE)
Itti et al. [6]	0.1130(0.0011)	0.6146(0.0008)
Bruce and Tsotsos [7]	0.2029(0.0017)	0.6727(0.0008)
Gao et al. [9]	0.1535(0.0016)	0.6395(0.0007)
Zhang et al. [8]	0.2097(0.0016)	0.6570(0.0008)
Seo and Milanfar [10]	0.3432(0.0029)	0.6769(0.0008)
Our Model	0.4386(0.0034)	0.6970(0.0008)

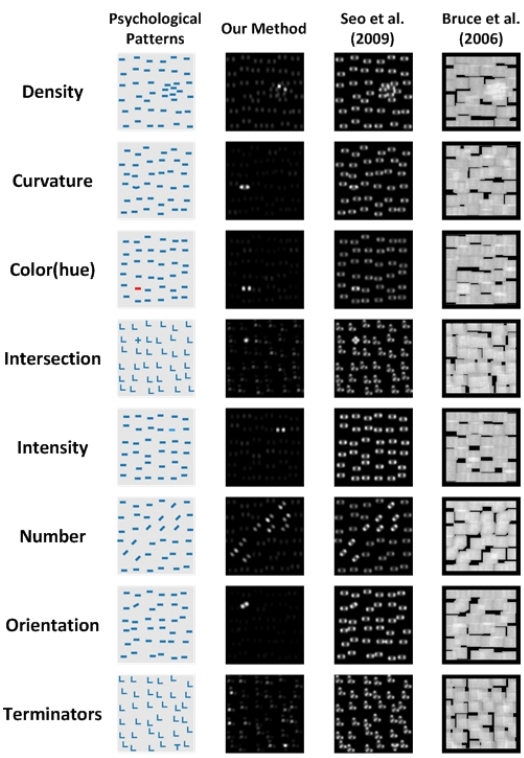

Fig. 1 Saliency map on psychological patterns.

Fig. 2 Examples of saliency map on human visual fixation data.

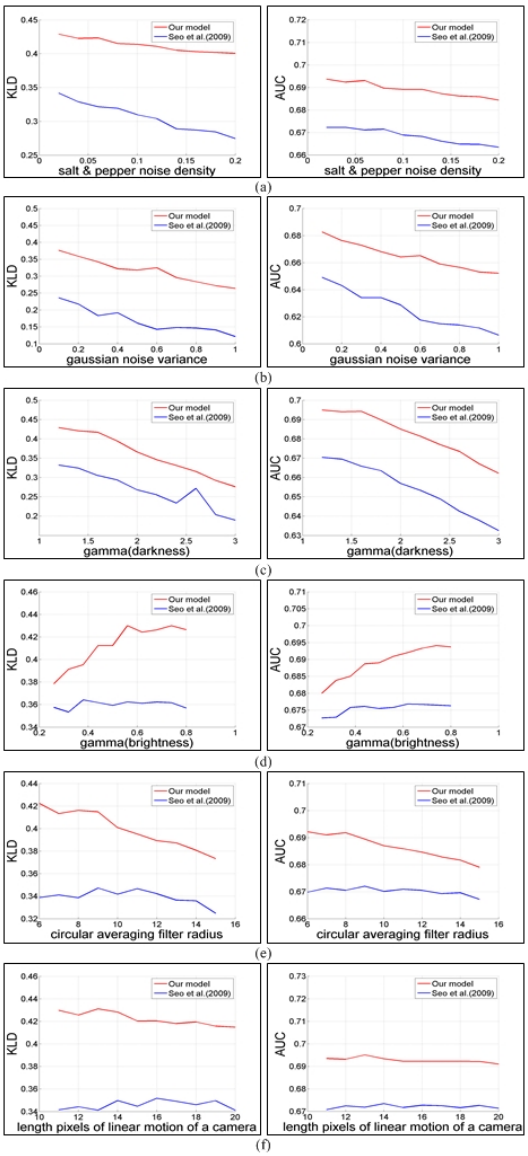

Fig. 3 Robust experimental results.

4.3 Robust Experiments

We also evaluated our model on distorted images. The dataset and experimental protocol are the same as in the experiment on visual fixation data, but each image of dataset is distorted using six different types of distortions at ten different levels of distortion. We compared our model against Seo et al. [10]. The curves of mean KLD and AUC are shown in Fig. 3. The experimental results show that our model achieves high robustness on distorted images.

5. Conclusions and Future Work

This letter presents a novel saliency detection model based on Bayes' theorem. Our model integrates the two parts of Bayes' equation and defines saliency as a function of local kernel density estimation of features in the center-surround region and global kernel density estimation of features at each pixel across the whole image. Moreover, a saliency detection method based on the DCT magnitude is proposed. Experiments demonstrate that the proposed model achieves good performance and robustness.

In future work, we will try to incorporate more features (e.g. local color features) to our model and make use of spare features to measure saliency. Also, we will extend our model to spatial-temporal domain so as to detect video saliency.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Project Number: 60832010) and the Fundamental Research Funds for the Central Universities (Grant No. HIT. NSRIF. 2010046).

References

[1] C. Papageorgiou and T. Poggio, "A trainable system for object detection," Int. J. Comput. Vis., vol.38, no.1, pp.15–33, 2000.
[2] B. Suh, H.B. Ling, B.B. Bederson, and D.W. Jacobs, "Automatic thumbnail cropping and its effectiveness," Proc. 16th Annual ACM Symposium on User Interface Software and Technology, pp.95–104, Vancouver, Canada, Oct. 2003.
[3] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "AutoCollage," ACM Trans. Graphics, vol.25, pp.847–852, July 2006.
[4] L. Itti, "Automatic foveation for video compressing using a neurobiological model of visual attention," IEEE Trans. Image Process., vol.13, no.10, pp.1304–1318, Oct. 2004.
[5] A. Treisman and G. Gelade, "A feature-integration theory of attention," Cognitive Psychology, vol.12, no.1, pp.97–138, Jan. 1980.
[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.20, no.11, pp.1254–1259, Nov. 1998.
[7] N.D.B. Bruce and J.K. Tsotsos, "Saliency based on information maximization," Advances in Neural Information Processing Systems, vol.18, pp.155–162, June 2006.
[8] L.Y. Zhang, M.H. Tong, T.K. Marks, H.H. Shan, and G.W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," J. Vision, vol.8, no.7, pp.32–51, Dec. 2008.
[9] D.S. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," J. Vision, vol.8, no.7, pp.13–31, Dec. 2008.
[10] H.J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," J. Vision, vol.9, no.12, pp.15–41, Nov. 2009.
[11] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," IEEE Trans. Image Process., vol.16, no.2, pp.349–366, Feb. 2007.
[12] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," IEEE Trans. Image Process., vol.18, no.9, pp.1958–1975, Sept. 2009.
[13] H.J. Seo and P. Milanfar, "Training-free, generic object detection using locally adaptive regression kernels," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1688–1704, Sept. 2010.