A Biologically Inspired Computational Model for Image Saliency Detection

Sheng He[1], Junwei Han[1], Xintao Hu[1], Ming Xu[1], Lei Guo[1], Tianming Liu[2]
[1]School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi 710072 China.
[2]Department of Computer Science, Bioimaging Research Center, the University of Georgia,Athens, GA 30602, USA.
{heshengxgd, junweihan2010, xintao.hu,xuming406}@gmail.com; lguo@nwpu.edu.cn;
tliu@uga.edu;

ABSTRACT

Image saliency detection provides a powerful tool for predicting where human tends to look at in an image, which has been a long attempt for the computer vision community. In this paper, we propose a biologically-inspired model for computing image saliency. At first, a set of basis functions that accords with visual responses to natural stimuli is learned by using eye-fixation patches from an eye-tracking dataset. Three features are then derived based on the learned basis functions including continuity, clutter contrast, and local contrast. Finally, these three features are combined into the saliency map. The proposed approach is easy to implement and can be used in many image and video content analysis applications. Experiments on a large-scale benchmark dataset and comparisons with a number of the state-of-the-art approaches demonstrate its superiority.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Algorithms, Performance, Experimentation

Keywords

Image saliency detection, Biologically-inspired, Sparse coding.

1. INTRODUCTION

Due to the ability of automatically estimating where human tends to focus on in the viewing field, computational modeling of visual saliency is gradually becoming the cornerstone of many multimedia applications such as video summarization [1], image resizing [2], and object segmentation [3]. It has been thus extensively studied in recent years. Existing algorithms can be broadly classified into two categories: biologically inspired and computationally inspired approaches. Inspired by the neuronal architecture of the early primate visual system, Itti [4] proposed the first biologically plausible framework, which implemented the saliency map by "center-surround" operating on three contrast-based features across multi-scales. Motivated by the evidence that the receptive fields of simple-cells in the primary visual cortex (V1) are similar to sparse codes learned from natural image patches, another school of methods [5-7] built the saliency map based on sparse codes mining. The mining schemes include Information Maximization [5], Incremental Coding Length [6], and Site Entropy Rate [7].

More algorithms are purely computationally inspired. The work in [8-10] detected the saliency in the frequency domain. Yu et al. [11] proposed a complementary saliency map by combining sketch-like and envelop-like features. Ren et al. [12] considered super-pixels as the basic unit and adopted the clustering technique to improve the performance of saliency detection. More recently, Cheng et al. [15] developed an elegant histogram descriptor indicating global contrast under the assumption that large objects are more attentive than tiny ones. Other approaches pose saliency modeling as a classification problem. Supervised learning algorithms, such as Support Vector Machines [13] and Conditional Random Fields [14], were used to optimize features guided by a ground truth dataset labeled by human or obtained from human eye-tracking data.

In some sense, the purely computationally inspired approaches are ad hoc. Their performance largely relies on the image database they work on since many attention factors, features, and the learning procedure are oriented towards the prior assumptions built on the image database. The human visual attention mechanism is much more complicated than what we can understand so far. This limits the performance of biologically inspired approaches to some extent. However, this category of methods is to mimic the capability of human and attempts to discover the intrinsic cues of visual attention. As long as the related areas like biology, psychology, and neurology make the breakthrough in visual attention mechanisms, it is certainly to benefit the biologically inspired methods significantly. Therefore, biologically inspired methods are potentially more promising.

The saliency maps computed by many current methods still suffer from such drawbacks as low resolution, ill-defined salient object boundary, non-uniform entire salient object, which also has been pointed out in [10] and [11]. These problems may arise from the fact that many existing approaches only take either local contrast or global contrast into consideration. They do not balance these two aspects of cues well. The emphasis on local contrast tends to produce higher saliency values near salient object boundaries. On the contrary, the use of global contrast leads to ill-defined boundaries.

In this paper, we propose a biologically inspired bottom-up computational model to detect image saliency based on sparse coding representations that have been acknowledged to be similar to the receptive fields of simple cells in V1. The proposed model is motivated by the following three commonly agreed biological

Conference'10, Month 1–2, 2010, City, State, Country.

evidences: 1) According to Gestalt principles, continuous objects are more likely to attract attentions. The continuity is one of essential factors for saliency detection. 2) The clutter contrast of a visual scene can influence our visual search and perception of the stimuli [16]. 3) A neuron's activities are mostly driven by its neighbors [4].

Three features are derived based on the sparse coding theory to reflect the above evidences and then combined to yield the saliency map. The first two features emphasize the global consideration and the last one is to detect the local contrast. The major contributions of the proposed work can be summarized as follows: 1) The sparse codes are learned from eye-fixation patches obtained from eye tracking data rather than arbitrary patches, which is demonstrated to be able to achieve better results. 2) All features are extracted based on sparse coding coefficients, which is simple, efficient, and analogous to neuron's activities. 3) Clutter contrast features are modeled using sparse coding and formally applied to detect image saliency. 4) Empirical comparisons of the proposed work with a number of state-of-the-art approaches are performed on a publicly available database.

The remainder of this paper is organized as follows. Section 2 describes sparse coding representations. Section 3 presents the development of the saliency map. Experimental results are given in Section 4. Finally, conclusions are drawn in Section 5.

2. SPACE CODING RESPRESENTATIONS

It has been mostly agreed that the sparse coding principle is an appropriate simulation to the intrinsic activation property of visual neurons in V1 [17]. An image patch $\mathbf{I}$ can be represented as a linear superposition of a set of patch bases:

$$\mathbf{I} = \sum_{k=1}^{n} \alpha_k \mathbf{B}_k \tag{1}$$

where $\mathbf{B}_k$ and α_k denote the k^{th} basis function and the corresponding coefficient, respectively. Let $\mathbf{F}_k = \mathbf{B}_k^{-1}$ be the k^{th} filter function and α_k can be calculated by:

$$\alpha_k = \sum_{(x,y)} \mathbf{F}_k(x,y)\mathbf{I}(x,y) \tag{2}$$

The sparse coding principle believes the receptive fields of simple cells in V1 can be characterized by basis functions each accounting for a neuron's response. As studied in [17], finding a complete set of basis functions which spans the image space is a critical issue. Traditionally, basis functions are learned using Independent Component Analysis (ICA) from a set of general-purpose image patches randomly selected from a large-scale database [5-7]. In principle, this strategy can obtain a complete set of bases approximately if we have boundless training data. However, in practice, with a view to limited computation resource, a task-oriented scheme may achieve better performance for the specific task. In this paper, we propose to learn basis functions using eye-fixation patches obtained from a publicly available eye tracking database [13], instead of using patches randomly selected from a general-purpose database. The learned bases account for image structures gaining attentions more likely.

88918 eye-fixation patches which are more attractive from all 1000 images in the eye tracking database proposed in [13] are collected. ICA algorithm is utilized to learn the basis functions with the size of 7×7×3 =147. Fig. 1 illustrates basis functions learned from fixation patches and learned from image patches randomly selected from the same database, respectively.

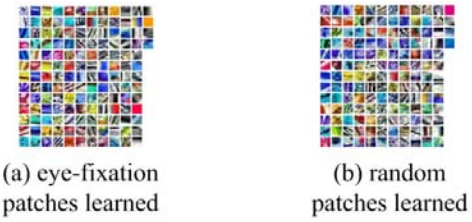

Figure 1: The basis functions learned from (a) eye-fixation patches and (b) random-selection patches.

3. THE SALIENCY MODEL

3.1 Continuity-based Features

According to Gestalt principles, continuity is an important property for attentive objects. This paper proposes to model it based on sparse codes. The framework is shown in Fig. 2.

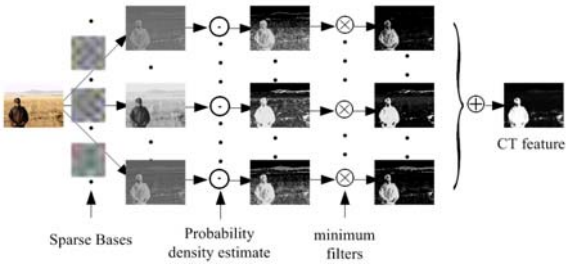

Figure 2: The proposed framework of generating continuity-based features.

At first, the input image is filtered by the set of learned basis functions (according to Eq. (2)) to obtain the corresponding coefficient maps. Afterwards, in the k^{th} coefficient map, we construct the coefficient histogram. The counts in the histogram bins are converted to a discrete probability distribution $H_k(x,y)$, where (x,y) is a location in the map. $H_k(x,y)$ is modified by a Gaussian kernel as:

$$P_k(x,y) = \exp\left(-\frac{H_k(x,y)}{2\sigma^2}\right) \tag{3}$$

Since the coefficient histogram technique only models the coefficient coherence without taking spatial connectivity into account, a post-process is utilized to remove the isolated patch noise. Due to its simplicity, the minimum filter is applied within the neighborhood of 5×5. Finally, the continuity-based feature map is generated as the summation of all coefficient maps:

$$CT(x,y) = \sum_k P_k(x,y) \tag{4}$$

3.2 Clutter Contrast Features

As mentioned in [16], visual clutter contrast is one of inherent factors to decide the image saliency. In [16], a feature congestion measure is proposed to estimate the clutter of any scene. To the best of our knowledge, very little work of saliency detection has formally adopted clutter contrast as a feature. In this paper, we define clutter as the complexity of the structure of the patches in images, which is measured based on the sparse coding representations.

The framework of generating clutter-based features is displayed in Fig. 3. Given an input image and a set of basis functions, each location (x, y) can be represented by a set of coefficients $\boldsymbol{\alpha}(x, y)$ according to the Eq. (2). Then, the clutter contrast features are calculated as follows:

$$CC(x,y) = \exp(-\frac{(\lambda A(x,y) + (1-\lambda)E(x,y))}{2\sigma^2}) \quad (5)$$

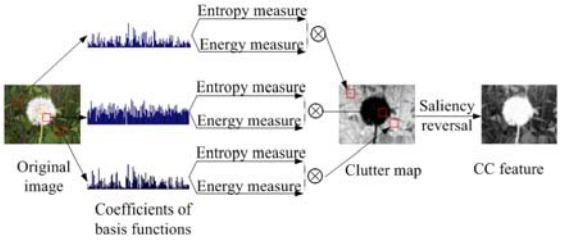

Figure 3: The framework of generating clutter contrast features.

Here, $A(x, y)$ and $E(x, y)$ are entropy and energy over $\boldsymbol{\alpha}_{(x,y)}$, respectively, which are formulated as:

$$A(x,y) = \sum_i p_i \log p_i, \quad p_i = \frac{|\alpha_i|}{\sum_j |\alpha_j|}, \ \alpha_i, \alpha_j \in \boldsymbol{\alpha}(x,y) \quad (6)$$

$$E(x,y) = \sum_i \alpha_i^2 \quad (7)$$

In our implementation $\lambda = 0.4$.

As pointed out in [20], there are two kinds of contrast patterns generally: smooth background with cluttered salient objects and smooth salient objects with cluttered background. The clutter contrast calculation by Eq. (5) can handle the former contrast pattern only but is inappropriate for the latter pattern for the purpose of highlighting salient objects. A saliency reversal procedure [14, 20] is utilized to refine the results, which is based on variance comparison between background and salient object.

3.3 Local Contrast Features

Local contrast is a major cue to detect saliency. It is normally obtained by center-surround differencing (CSD). This paper models local contrast by means of CSD over sparse coding coefficients, which is defined as follows:

$$LC(x,y) = \sum_{l=1}^{n} \Theta_l(x,y) \quad (8)$$

$$\Theta_l(x,y) = \sum_{(x',y') \in \Omega} \left\{ \exp(\frac{\left\| \boldsymbol{\alpha}(x,y) - \boldsymbol{\alpha}(x',y') \right\|_2}{2\sigma^2}) - 1 \right\} \quad (9)$$

Here Θ is the CSD operator and Ω is the neighborhood of (x, y), and l is the scale. In our implementation, the size of Ω was set to 5×5 and the scale number was set to 5. The framework of obtaining local contrast is shown in Figure 4.

3.4 Feature Combination

In visual attention modeling, the feature combination is a tricky problem as biological theories behind it are still unclear. Although many algorithms [13, 14] adopted learning-driven scheme, they are ad hoc and require an extensive training stage. In the proposed work, inspired by the conclusion in [21] that V1 cells tuned to different features interact through lateral connections and activity in cells responding to stimuli is suppressed through mutual inhibition, we develop a novel combination strategy to reveal the mutual interaction across features.

$$S(x,y) = \frac{1}{N}(CT^2(x,y) + CC^2(x,y) + LC^2(x,y) + CT(x,y) \cdot CC(x,y) + CT(x,y) \cdot LC(x,y) + CC(x,y) \cdot LC(x,y)) \quad (10)$$

Here N is a normalizator.

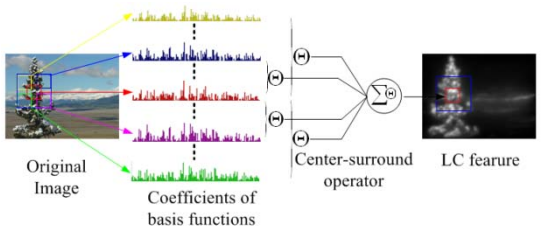

Figure 4: Framework of obtaining local contrast features.

4. EXPERIMENTAL RESULTS

Two evaluation studies were conducted to test the proposed work. The first evaluation is to compare the performance of two basis functions learning strategy: eye-fixation patches based and random patches based methods. The second test is to quantitatively evaluate the proposed saliency detection algorithm by comparing with a number of state-of-the-art algorithms. In our experiments, 1000 images in the benchmark dataset presented in [10] with their ground truth which is manually labeled by subjects, were utilized as testing data.

4.1 Evaluations of Eye-fixation Patches based Learning

Fig. 5 shows the ROC curves of saliency detection computing on all testing data by using basis functions learned from eye-fixation patches and learned from random patches, respectively. As can be seen from the comparison, eye-fixation patches learning strategy can improve the performance of sparse coding representations for the task of visual saliency detection.

4.2 Evaluations of the Proposed Algorithm

Fig. 6 displays some exemplar results that include corresponding feature maps, overall saliency maps, and ground truth. To objectively evaluate the proposed work, we compare it with a number of state-of-the-art algorithms including the Information Maximization (IM) [5], Coding Length Increments (CLI) [6], Special Residual (SR) [8], Frequency-Tuned (FT) [10], Feature Congestion Measure (FCM) [16], Self-ReSemble (SRS) [18], SUN [19]. Fig. 7 shows the comparison results and Fig. 8 shows the ROC curves of saliency detection calculating on all testing data. The comparison results can demonstrate the superiority of the proposed work.

5. CONCLUSIONS

In this paper, a computational model of visual saliency motivated by biological evidences was reported. Three attention-oriented features were derived based on sparse coding representations and then combined to generate the saliency map. Comprehensive

experimental results on a benchmark dataset and comparisons with state-of-the-art approaches have demonstrated the effectiveness of the proposed work.

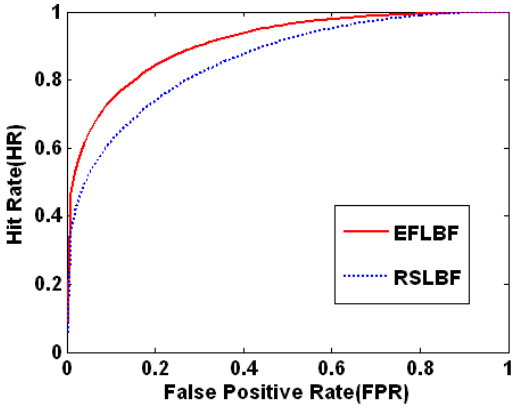

Figure 5: The ROC curve of saliency detection by using Eye-Fixation Learned Basis Functions (EFLBF) and Random-Selection Learned Basis Functions (RSLBF).

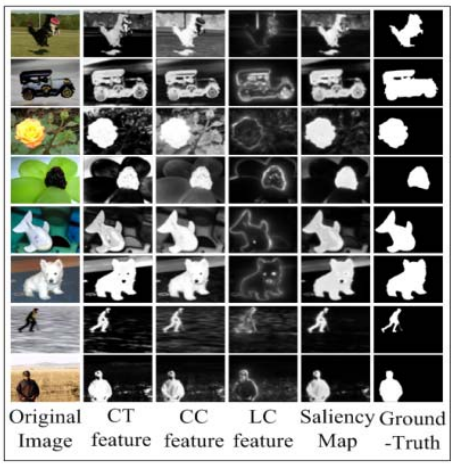

Figure 6: Examples of saliency maps and corresponding feature maps obtained by the proposed work.

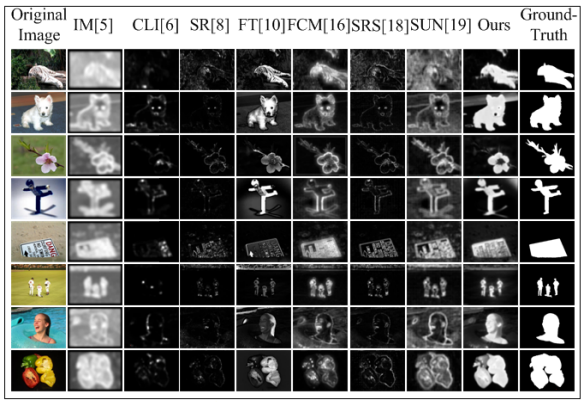

Figure 7: Saliency maps by different algorithms.

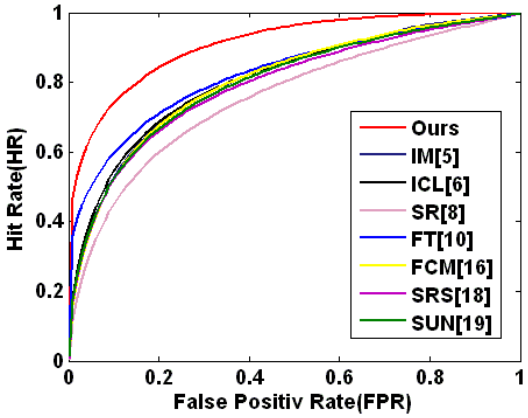

Figure 8: ROC curves of saliency detection by different algorithms.

6. REFERENCES

[1] Ma, Y., Lu, L., Zhang, H. and Li, M. A user attention model for video summarization. *ACM Multimedia* 2002.

[2] Avidan,S. and Shamir, A.. Seam carving for content-aware image resizing. *ACM Transactions on Graphics.* 26(3), 2007.

[3] Han, J., King, N., Li, M. and Zhang, H. Unsupervised extraction of visual attention objects in color images. *IEEE Trans. Circuits Syst. Video Techn.* 16(1): 141-145, 2006.

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

[5] Bruce, N. and Tsotsos, J. Saliency based on information maximization. In *Advances in neural information processing systems.* 18: 155, 2006.

[6] Hou, X. and Zhang, L. Dynamic visual attention: Searching for coding length increments. In *Advances in neural information processing systems.* 21: 681–688, 2008.

[7] Wang, W., Wang, Y., Huang, Q. and Gao, W. Measuring Visual Saliency by Site Entropy Rate. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[8] Hou, X. and Zhang, L. Saliency detection: A spectral residual approach. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] Guo, C., Ma, Q., and Zhang, L. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S. Frequency-tuned salient region detection. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[11] Yu, H., Li, J., Tian, Y. and Huang, T. Automatic Interesting Object Extraction From Images Using Complementary Saliency Maps. *ACM Multimedia* 2010.

[12] Ren, Z., Hu, Y., Chia, L. and Rajan, D. Improved saliency detection based on superpixel clustering and saliency propagation. *ACM Multimedia* 2010.

[13] Judd, T., Ehinger, K., Durand, F. And Torralba, A. Learning to predict where humans look. In *Proceeding of International Conference on Computer Vision.* 2009.

[14] Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X. and Shum, H. Y. Learning to detect a salient object. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[15] Cheng, M., Zhang, G., Niloy, J., Huang, X. and Hu, S. Global Contrast based Salient Region Detection. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[16] Ruth, R., Li, Y. and Lisa, N. Measuring visual clutter. *Journal of Vision.* 7(2):17, 2007.

[17] Olshausen, B. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature.* 381(6583): 607-609, 1996.

[18] Seo, H. and Milanfar, P. Nonparametric bottom-up saliency detection by self-resemblance. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[19] Zhang, L., Tong, M., Marks, T., Shan, H. and Cottrell, G. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision* 8(7):32, 2008.

[20] Wang, Z. and Li, B. A two-stage approach to saliency detection in images. In *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[21] Koene, A. and Zhaoping, L. Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom–up saliency map in V1. *Journal of Vision.* 7(7):17, 2007.