

# Measuring Visual Saliency by Site Entropy Rate

Wei Wang<sup>1,3,4</sup>, Yizhou Wang<sup>1,2</sup>, Qingming Huang<sup>4</sup>, Wen Gao<sup>1,2</sup>

<sup>1</sup>National Engineering Lab for Video Technology

<sup>2</sup>Key Lab. of Machine Perception (MoE), Peking University, Beijing 100871, China

<sup>3</sup>Key Lab. of Intelligent Information Processing Chinese Academy of Sciences, Beijing 100080, China

<sup>4</sup>Graduate School, Chinese Academy of Sciences, Beijing 100039, China

wwang@jdl.ac.cn, yizhou.wang@pku.edu.cn, qmhuang@jdl.ac.cn, wgao@pku.edu.cn

## Abstract

*In this paper, we propose a new computational model for visual saliency derived from the information maximization principle. The model is inspired by a few well acknowledged biological facts. To compute the saliency spots of an image, the model first extracts a number of sub-band feature maps using learned sparse codes. It adopts a fully-connected graph representation for each feature map, and runs random walks on the graphs to simulate the signal/information transmission among the interconnected neurons. We propose a new visual saliency measure called Site Entropy Rate (SER) to compute the average information transmitted from a node (neuron) to all the others during the random walk on the graphs/network. This saliency definition also explains the center-surround mechanism from computation aspect. We further extend our model to spatial-temporal domain so as to detect salient spots in videos. To evaluate the proposed model, we do extensive experiments on psychological stimuli, two well known image data sets, as well as a public video dataset. The experiments demonstrate encouraging results that the proposed model achieves the state-of-the-art performance of saliency detection in both still images and videos.*

## 1. Introduction

Visual attention plays an important role in human visual system when perceiving the world. It is able to select the most valuable visual information from a large amount of the sensory data to interpret complex scenes in real time fashion. In the last decades, the psychophysics of visual attention has been extensively studied (e.g. [21, 23]) and many computational models of visual attention have been proposed (e.g. [22, 12]). By exploiting these models, visual attention has been successfully applied to many computer vision applications, e.g. interest region detection [17], ob-

ject recognition [19], and scene classification [20].

In this paper, we propose a biology-inspired bottom-up computational model of attention based on *visual saliency*, which is considered as the impetus for selection of fixation points[12]. Instead of defining visual saliency from the perspective of the commonly used center-surround mechanism, we propose a new saliency measure derived from the principle of *information maximization*. This principle suggests that the human visual system (HVS) tends to focus on the most *informative* points in an image in order to efficiently analyze the scene[23]. Our computational model assigns a point a higher saliency value if it is more “informative”. In addition, our model is constructed based on the following mostly agreed biological evidences. (1) With the understanding of the properties of simple-cells in primary visual cortex(V1), sparse coding is broadly acknowledged as an efficient coding strategy for optimal information transfer and metabolic efficiency [8]. (2) Two different spatial scales of cortical connectivity construct a network by the recurrent local connection and long-range horizontal connection. In the network, neurons communicate with each other via synaptic firings to give rise to an emergent behavior. [15] (3) A neuron’s activities are driven by the total synaptic input from its neighbors [14].

By referring to the above three key evidences, we propose a framework (shown in Fig. 1) in order to simulate the computational function of visual saliency in our brain, rather than the real architecture of the early primate visual system. Fig. 1 shows the procedure of computing the saliency map of an image. (1) The system first filters the input image with a number of sparse coding bases. Each filtering generates a feature map, which is a sub-band image of the corresponding sparse code. As being found that the receptive fields of simple-cells in V1 are similar to sparse codes learned from natural images[16], we learn two sets of the sparse coding basis functions as the early visual features from a large number of natural images in both color and gray (as shown in Fig. 2). (2) To simulate the cortical

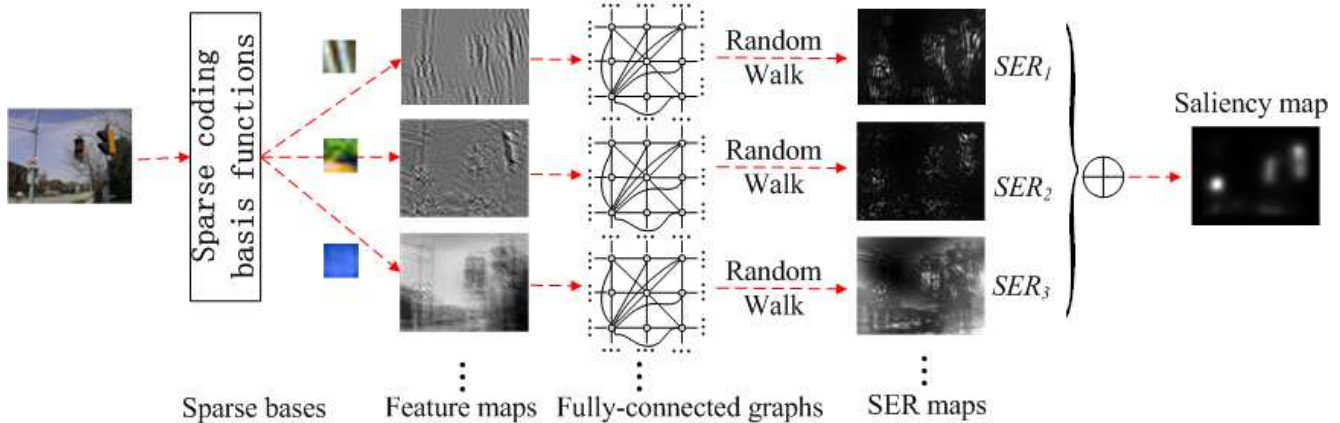


Figure 1. The proposed framework. An input image is filtered by sparse coding basis functions to obtain the corresponding sub-band feature maps. A fully-connected graph is constructed for each feature map, and a random walk is run on each graph to compute a SER map of each channel. Finally, the saliency map is generated by summing over all the SER maps.

neuron connectivity we adopt a fully-connected graph representation for the feature maps. The full connectedness is able to capture the long range relation between two sites in an image. (3) A random walk is adopted on each sub-band graph in order to model the signal transmission among the neurons in the network. As the entropy rate (ER) of the random walk is the average information of one step move, we use it to compute the total transmitted information of all the graph nodes (neurons). By distributing the entropy rate onto each graph node (neuron), we propose a new term, *Site Entropy Rate* (SER), to measure the average information from a node to all the others. In this way, we obtain a SER map for each sub-band graph. It is worth pointing out that the SER describes the accumulative effects of all the interactions between a neuron and the other connected ones, not necessarily the exact total synaptic input from its neighbors. It is also interesting to note that the SER formulation also explains the center-surround mechanism encoded in the computational model. (4) Finally, the saliency map is computed by summing over all the sub-band SER maps in a similar fashion to the Feature-Integration Theory [21]. The higher an integrated SER value at a site, the more salient the site.

We also extend our model to spatial-temporal domain to detect salient spots in videos. Based on the biological evidence that neural response attenuates with prolonged exposure to the same stimuli [16], we assume that it is the novel signal and the change of the signal at a site that make the site salient. Under this assumption, we compute the saliency map at time  $t$  by discounting the effect from the previous frames so as to simulate the temporal change of neural response. More specifically, we update each sub-band feature map by subtracting the temporally weighted

feature responses from the corresponding sub-band feature map of the previous frames. Then we still run random walks on the fully-connected graphs of the updated feature maps to obtain the SER maps and finally the saliency map.

We do extensive experiments on psychological stimuli, two well known image datasets, as well as a public video dataset. The experimental results show that our proposed saliency model achieves the best performance on both still images and videos.

### 1.1. Related work

In this section, we briefly review the existing literature that closely relates to the proposed model.

*The approaches with graph representation* In order to simulate attentional shifts and eye movements, Costa et al. [3] employ random walks on image lattice to compute the visual saliency. The saliency value at a spot is proportional to the frequency of visit to the spot at the equilibrium of the random walk. Harel et al. [9] extend [3]’s method by proposing a better dissimilarity measure to model the transition probability between two nodes. Gopalakrishnan et al. [7] adopt the same saliency measure and formulate the salient region detection as random walks on a fully-connected graph and a  $k$ -regular graph to extract the global and local image properties respectively. In this paper we utilize fully-connected graph structure to simulate the cortical neuron connection, and derive a new saliency measure from the information maximization principle.

*The approaches based on information maximization* This kind of methods consider that *information* is the driving force behind attentive sampling and use the *rarity* of features to measure visual saliency. Bruce et al. [2] adopt the *self-information* of sparse features as a saliency measure.

Hou et al. [11] assume that “salient features can offer entropy gain”. They introduce the *Incremental Coding Length* to allocate different amount of energy to features according to their rarity. Our model defines the *Site Entropy Rate* of the random walk on the graph structure. It measures the average information transmitted from a node to both its local and far neighbors so as to simulate the activity level of the neuron and the saliency degree of its receptive field.

*The approaches for the Center-surround mechanism* This group of methods model the center-surround mechanism of primary visual cortical cells. Itti and Koch [12] propose a biologically-plausible visual saliency model based on the center-surround contrast mechanism. By arguing that [12]’s linear model of the similarity measure on color, intensity, and orientation is inconsistent with the properties of higher level human judgement (which tends to be asymmetric), Gao et al. [6] propose a discriminant center-surround hypothesis using *mutual information*. It poses saliency detection as a classification problem, and obtains an optimal solution from the decision-theoretic perspective. Although our model is derived on a fully-connected graph from the point of view of information maximization, the proposed saliency measure, SER, can be considered as a generalized center-surround model (to be explained in Section 2.3).

*Saliency models for videos* Several approaches have been proposed to measure the saliency on videos. Itti et al.[13] propose a model of surprising/salient event detection. They formulate the surprise/saliency as the Kullback-Leibler divergence between the posterior and prior beliefs of an observer about the scene. In fact, this model extends the spatial center-surround contrast to the spatiotemporal domain. Hou et al.[11] consider the temporal correlation among video frames as a Laplacian distribution and replace the feature activity ratio distribution over space with the cumulative activity ratio distribution over both space and time in their feature-based model. However, our proposed method aims to model the biological behavior of neurons in temporal domain. Specifically, the model simulates attenuation effect when neurons are exposed to the same stimuli over time.

The rest of the paper is organized as follows. In Section 2, we introduce the details of the theory and the model. The experimental results on psychological stimuli, color and gray images and videos are presented in Section 3. Finally, we conclude the paper in Section 4.

## 2. The Model

In this section, we shall first introduce the learning of the sparse coding basis functions, followed by the model representation, the new saliency measure (SER), and finally the computation of saliency maps.

### 2.1. The sparse coding bases

There are evidences showing that when presented to a scene, only a small number of early visual neurons out of a large set will be activated [1]. To simulate this property of simple cells in the primary visual cortex, the sparse coding theory is proposed to extract the intrinsic structure of natural images for efficient coding [16]. The theory assumes that an image  $\mathbf{I}$  is a linear superposition of a number of image bases  $B_k$ , where  $k$  indexes for the location, orientation and scale, etc.

$$\mathbf{I} = \sum_k a_k B_k \quad (1)$$

$a_k$  is the coefficient of basis  $B_k$ ,  $p(a_k) \propto e^{-\alpha|a_k|}$  is the high-order statistics prior to enforce the sparsity. This coefficient can be computed by its corresponding filter function  $G_k$

$$a_k = \sum_{x,y} G_k(x,y)\mathbf{I}(x,y) \quad (2)$$

where  $G_k$  is the inverse/pseudoinverse of  $B_k$ .

In this paper, we adopt Independent Component Analysis (ICA) [10] to learn two sets of sparse coding basis functions from a color image dataset and a gray image dataset (to be introduced in Section 3). And we use the coefficient  $a_k$  as the early visual feature response  $f_k(x,y)$  of  $\mathbf{I}$  being filtered by  $G_k$ . The filter responses of  $G_k$  form the  $k$ -th sub-band feature map  $F_k$ .

### 2.2. The sub-band graph representation

To simulate the recurrent local and long-range connections between neurons, we construct a fully-connected graph  $G_k = \{V_k, E_k\}$  on each feature map  $F_k$ , where  $V_k = \{v_{k1}, \dots, v_{kn}\}$  is the set of nodes at image pixels.  $v_{ki} = (x_i, y_i, f_k(x_i, y_i))$  has two attributes: the location and the feature response.  $E_k = \{e_{kij}, i, j = 1, \dots, n\}$  is the set of weighted edges connecting every pair of nodes, where  $e_{kij} = (i, j, w_{kij})$ . The weight  $w_{kij}$  measures the dissimilarity between node  $i$  and  $j$  from two aspects: the feature dissimilarity denoted by  $\phi_{kij}$  and the spatial distance denoted by  $d_{ij}$ . The weight is simply defined as

$$w_{kij} = \phi_{kij} * d_{ij} \quad (3)$$

where  $\phi_{kij}$  and  $d_{ij}$  are defined as

$$\phi_{kij} = \exp\{|f_k(x_i, y_i) - f_k(x_j, y_j)|/M_k\} \quad (4)$$

$$d_{i,j} = \exp\{-\lambda \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{D}\} \quad (5)$$

where  $M_k$  is the largest feature difference,  $D$  is the larger dimension (horizontal or vertical) of the image.  $\lambda$  is a positive parameter to balance the importance of the two aspects.  $\lambda = 5$  in our experiment.

### 2.3. The site entropy rate

It is difficult to model the large-scale signal transmission on the neuronal network when spontaneous synaptic firings occur simultaneously. Instead, we simulate the information flow from one neuron to another as a random walk process from one site to another on the fully-connected graphs of the sub-band feature maps. We define transition probability of the random walk from site  $i$  to site  $j$  in terms of the normalized edge weights between site  $i$  and  $j$ .

$$P_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} \quad (6)$$

In this section, the sub-band index  $k$  is omitted for simplicity.

As we know, random walk is a stochastic process of a sequence of random variables  $\{X_i\}$ . A Markov chain is a simple case of the random walk process:

$$\begin{aligned} Pr(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) \\ = Pr(X_{n+1} = x_{n+1} | X_n = x_n) \end{aligned} \quad (7)$$

for all  $x_1, x_2, \dots, x_{n+1} \in \chi, n = 1, 2, \dots$

If the finite-state Markov chain is irreducible and aperiodic, the stationary distribution  $\pi$  is unique and  $\pi P = \pi$ . Here  $P$  is the transition matrix. For a random walk process, the element of  $\pi$  at node  $i$  can be simply computed as

$$\pi_i = \frac{W_i}{2W} \quad (8)$$

where  $W_i = \sum_j w_{ij}$  is the total weight of edges emanating from node  $i$ , and  $W = \sum_{i,j:j>i} w_{ij}$  is the sum of the weights of all the edges.

The stationary probability  $\pi_i$  is generally considered as the frequency of visit to the node  $i$  at the equilibrium of the random walk [3]. In the literature, people directly use the  $\pi_i$  as the saliency measure at location  $i$  [3, 9]. However, we think that the total information sent from one neuron to another is decided by two terms: the transmission (visit) frequency and the amount of information at each transmission (visit). Therefore, we need an information theoretic measure for the information transmission so as to account for the two factors during the random walk process.

In information theory, the *entropy rate*  $H(\chi)$  is defined to measure the average entropy of a sequence with  $n$  random variables, which is the average information obtained along with the time as ( $n \rightarrow \infty$ ). It is defined as follows

$$H(\chi) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n) \quad (9)$$

when the limit exists. From the theorem introduced in [4], we can obtain a computable entropy rate of a Markov Chain as follows.

Let  $\{X_i\}$  be a Markov chain with stationary distribution  $\pi = \{\pi_i\}$  and transition matrix  $P = \{P_{ij} : P(X_n = j | X_{n-1} = i)\}$ . Then the entropy rate is equivalent to the conditional entropy  $H(X_n | X_{n-1})$

$$H(\chi) = H(X_n | X_{n-1}) = - \sum_{ij} \pi_i P_{ij} \log P_{ij} \quad (10)$$

By a simple manipulation of the above equation, we have

$$H(\chi) = \sum_i (\pi_i \sum_j -P_{ij} \log P_{ij}). \quad (11)$$

we define a new term, *Site Entropy Rate* (SER)

$$SER_i = \pi_i \sum_j -P_{ij} \log P_{ij} \quad (12)$$

to measure the average information transmitted from node  $i$  to the other connected ones. The SER can be divided into two parts: the stationary distribution term  $\pi_i$  and the entropy term  $\sum_j -P_{ij} \log P_{ij}$ . The  $\pi_i$  tells the frequency at which a random walker visits node  $i$ . It is also the frequency that node  $i$  communicates with the other nodes. The entropy term  $\sum_j -P_{ij} \log P_{ij}$  measures the uncertainty of node  $i$  jumping to the other nodes at one step. It is related to the amount of information transmitted from node  $i$  to the others at one step. SER is the product of the two terms. It measures the average total information transmitted from node  $i$  to the others in one step. We hope to use this measure to simulate the activity level of a neuron. Thus, the higher the SER value, the more active is the neuron, and the more salient is its corresponding receptive field. Thus we adopt SER as a visual saliency measure.

Now, let us analyze the SER definition from another perspective. Eqn.12 suggests that if a site be salient, (1) the site should be frequently visited (large  $\pi_i$ ), and (2) the entropy term needs to be large too. In our model, the  $P_{ij}$  is determined by the edge weight between node  $i$  and node  $j$  (Eqn.6). And the edge weight encodes the feature difference and the spatial distance of the two nodes (Eqn.4&5). It can be considered as a spatially-weighted feature dissimilarity.  $\sum_j -P_{ij} \log P_{ij}$  achieves its maximum value if  $P_{ij}$  is uniform. There are two cases maximize the entropy: (1) the center-surround contrast pattern, and (2) the constant or smooth pattern. However,  $\pi_i$  in these two cases are quite different:  $\pi$  at the center of the center-surround contrast region is much larger than it is in the constant/smooth region. In another word, the visiting frequency enhances the saliency measure at center-surround contrast regions, but decreases the saliency value of constant/smooth regions. Therefore, our model can be considered as an information theoretic model for the generalized center-surround mechanism.

## 2.4. The saliency map

We obtain the SER map for each sub-band feature map. According to the Feature-Integrated Theory [21], the final saliency map is the sum of all the SER maps. The saliency value at pixel  $i$  is

$$S_i = \sum_k SER_{ki} \quad (13)$$

## 3. Experimental Results

To test the performance of the proposed model, we do extensive experiments on psychological stimuli, two public image datasets and a video dataset. On each dataset we compare our model with the state-of-the-art approaches based on commonly-used evaluation criterion.

### 3.1. Learning the sparse coding bases

We learn two sets of sparse coding basis functions as the early visual features of both color and gray images. A set of 192 color sparse basis functions is learned from 120,000  $8 \times 8 \times 3$  image patches randomly extracted from 1500 natural color images using ICA [10]. Using the same method, we learn another set of 64 gray image basis functions from 50,000  $8 \times 8$  gray image patches. The basis functions of color images and gray images are shown in Fig. 2.

As shown in Fig.1, each filtering generates a feature map, which is a sub-band image of the corresponding sparse code.

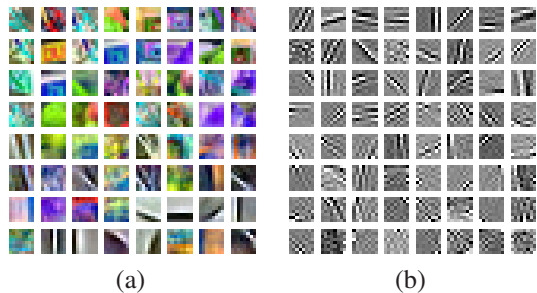


Figure 2. (a) The 64 components of 192 color sparse bases. (b) The 64 gray sparse bases.

### 3.2. Experiments on psychological stimuli

We test our model on several psychological stimuli which are commonly used to represent the pre-attentive visual features. These patterns include “line orientation”, “length”, “size”, “closure”, “curvature”, “density”, “number”, “intersection”, “terminator” and “color”, etc.

In this experiment, we test our model on five of these stimuli. As shown in Fig. 3, we compare our results with the results from Itti et al. [12] by showing the saliency maps and

the saliency heat maps. In the saliency heat maps, the hotter the color of a spot, the more salient it is. The figure clearly shows that our model predicts the saliency spots more accurately than [12] in the “number”, “curvature” and “intersection” stimuli. It is worth noting, in the “intersection” stimulus, the reason for the two extra regions we detect (in red circle) being salient is that the signal spatial density is different from their surroundings.

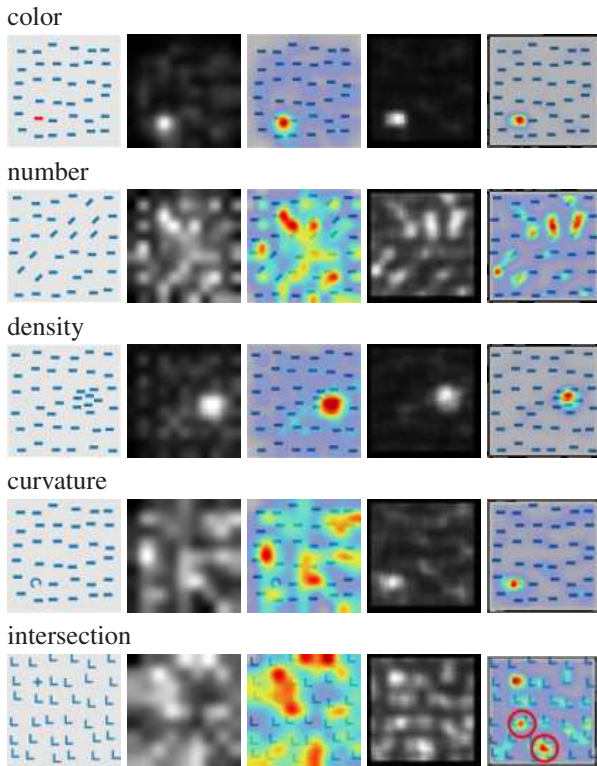


Figure 3. Comparison results of five psychological stimuli between our model and Itti et al.’s [12]. The columns from left to right are: the original stimuli, [12]’s saliency maps and its saliency heat maps, our saliency maps and saliency heat maps. The highlighted two salient regions with the red circles in “intersection” is because that the signal spatial density is different from their surroundings.

### 3.3. Experiments on still images

**The color image dataset** The popular color image dataset collected by Bruce et al. [2] usually serves as the benchmark dataset for comparing visual saliency detection results. This dataset consists of a variety of images about indoor and outdoor scenes. Eye fixations are recorded from 20 subjects on the 120 color images. We evaluate and compare our model with the others using two types of measures introduced in [2]: (1) In qualitative comparison, we show our saliency maps and the fixation density maps generated from the sum of all 2D Gaussians to the human fixations in Fig. 4, and compare the saliency maps to the other four

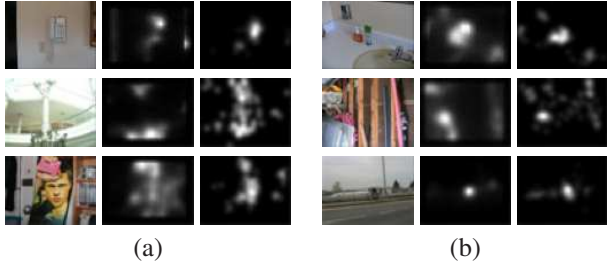


Figure 4. Results for a qualitative comparison between our model and human fixations. The first columns in (a) and (b) show the original images, the second columns are our saliency maps, and the third columns are the human fixation density maps.

state-of-the-art approaches ([12], [2], [6], [11]) in Fig. 5. Although the saliency maps from [11] are very similar to ours, the rank of the salient regions in our saliency map is more consistent with the fixation density map than that of [11]’s. (2) In quantitative performance evaluation, we compare the Receiver Operator Characteristic (ROC) curves and the ROC areas to the methods of [12], [2], [6], [11]. The ROC curve results are shown in Fig. 6 and the ROC areas are compared in Table 1. (The larger the ROC area, the better.) Both the ROC curves and ROC areas are generated by classifying the locations in a saliency map into fixations and non-fixations with varying quantization thresholds. (Note that for the compared four methods, there are small differences between the showed results and their reported results. This discrepancy is due to different sampling densities to obtain quantization thresholds.) It can be seen that our model achieves the best performance on the dataset.

Table 1. The ROC area comparison on the color image dataset

	ROC area
Itti et al. [12]	0.7031
Bruce et al. [2]	0.7522
Gao et al. [6]	0.7644
Hou et al. [11]	0.7808
Our model	0.8049

**The gray image dataset** We use the gray image dataset of natural scenes and the corresponding fixation data collected by Einhäuser et al. [5] to evaluate the performance of our model, and compare the results to the other five approaches. This dataset contains 108 gray images and each image has nine modified versions. Every version is generated by changing luminance-contrast in five randomly selected circular regions. The fixation data are recorded from seven human subjects observing  $108 \times 9$  gray images. (But not all subjects complete the eye-tracking experiments for the nine versions.) We also adopt the *ROC area* to assess our

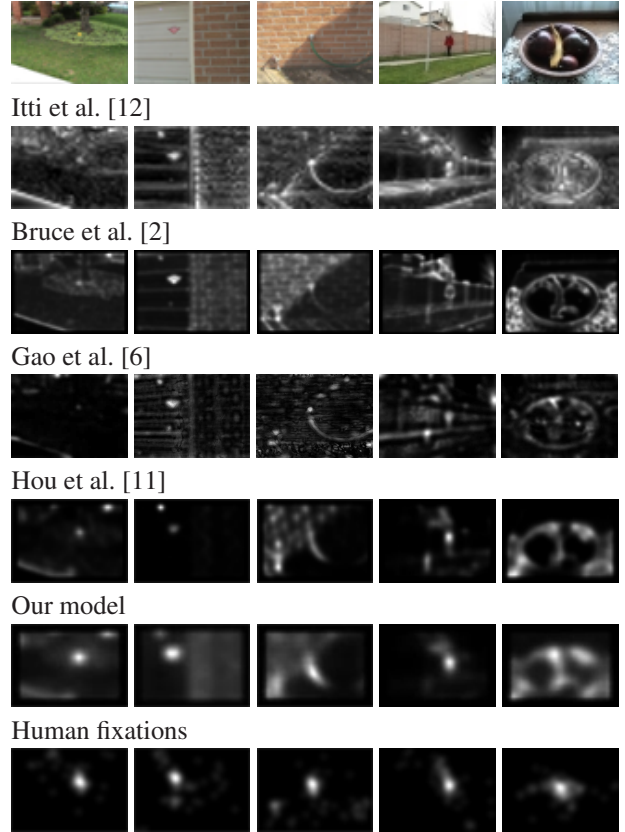


Figure 5. Results for a qualitative comparison between our model and the other four approaches. The rows from the top to the bottom are: the original images, the saliency maps of Itti et al.’s method [12], the saliency maps of Bruce et al.’s method [2], the saliency maps of Gao et al.’s method [6], the saliency maps of Hou et al.’s method [11], the saliency maps of ours, the human fixation density maps.

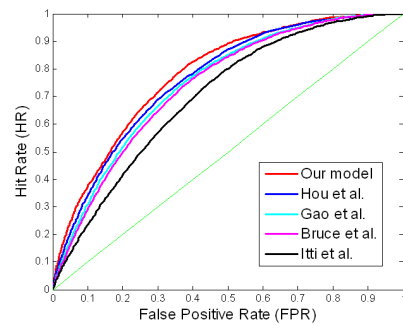


Figure 6. The ROC curves of our model and the other four state-of-the-art approaches on the color image dataset.

model. We select the images with more than three human subjects’ fixation data. The evaluation method takes the fixations as detection targets. The *ROC area* is computed with the saliency map generated by our model. In Table 2 we

illustrate the comparison of our quantitative evaluation results against those of the other five methods ([12], [9], [2], [6], [11]) Fig. 7 shows these saliency maps and human fixation density maps. It can be seen that our model predicts human fixations more accurately than the other five methods.

Table 2. The comparison results on the gray image dataset

	ROC area
Harel et al. [9]	0.5028
Gao et al. [6]	0.5203
Itti et al. [12]	0.5241
Hou et al. [11]	0.6094
Bruce et al. [2]	0.6420
Our model	0.6537

From the prediction results in Table 1 and Table 2, we can see that the ROC areas of color images are much larger than those of gray images. The images from the color image dataset generally contain only a few semantic objects. Thus, the human fixations are relatively consistent. But in the gray image dataset, there contain few semantic objects but raw signals. The image entropy is much higher than the color ones, the human fixations are very diversified. Moreover, the number of human subjects of this dataset is too small to accurately estimate the true distribution of human fixation. This causes the worse performance of our model on the gray image dataset compared with the color one. But even in this case, our model still outperforms the other models.

### 3.4. Experiments on videos

In sensory neuroscience, there is evidence showing that only the unexpected signal at one stage is transmitted to the next [18]. Moreover, the electrophysiological studies also demonstrate that neural response greatly attenuates with repeated or prolonged exposure to an initially novel stimulus [16]. By referring to these facts, we assume that it is the novel signal and the change of the signal at a site that make the site salient. In another word, if the signal at a site remains constant, even it is prominent in space, its saliency value will decrease along time. Under this assumption, we compute the saliency map at time  $t$  by discounting the effect from the previous frames so as to simulate the temporal change of neural response. More specifically, we update each sub-band feature map by subtracting the temporally weighted feature responses from the corresponding sub-band feature map of the previous  $k$  frames. Let  $f_j(x, y, t)$  denote the  $j$ -th sub-band feature map of frame  $t$ . The attenuated feature response at  $(x, y, t)$  in sub-band  $j$  is

$$f'_j(x, y, t) = |f_j(x, y, t) - \sum_{\tau=1}^k \exp(-\frac{\tau}{\sigma}) f_j(x, y, t - \tau)|$$

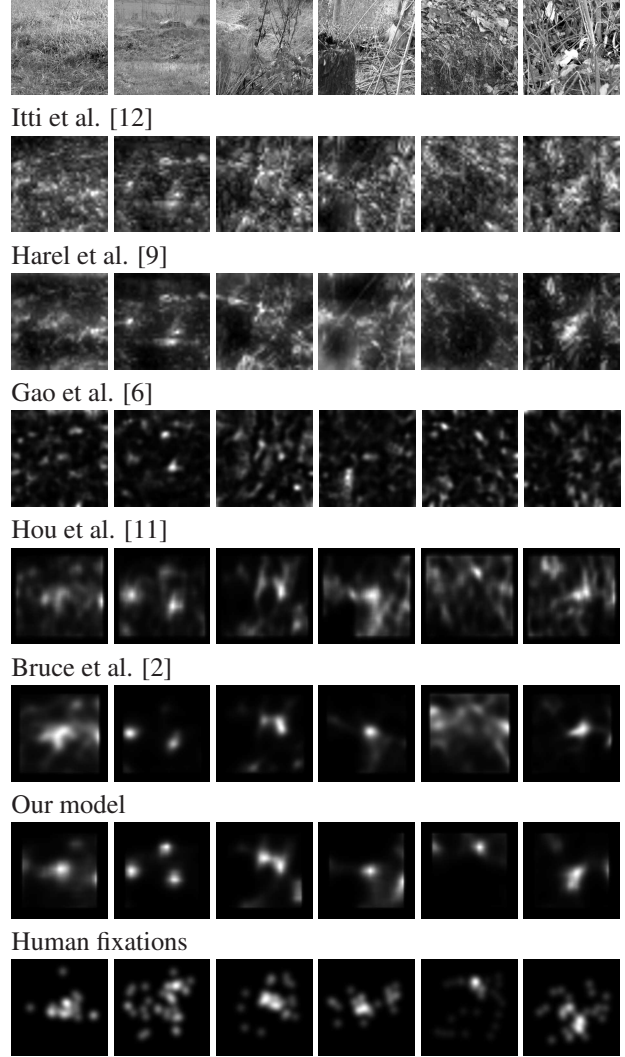


Figure 7. Results for a qualitative comparison between our model and the other five approaches. The rows from the top to the bottom are: the original images, the saliency maps of Itti et al.’s method [12], the saliency maps of Harel et al.’s method [9], the saliency maps of Gao et al.’s method [6], the saliency maps of Hou et al.’s method [11], the saliency maps of Bruce et al.’s method [2], the saliency maps of ours and the human fixation density maps.

where  $\sigma$  controls the attenuating rate, which is set to 1.5 in our implementation, and  $k = 4$  in our experiment.

Then the following steps are similar to those for computing the saliency maps of still images – we run random walks on the fully-connected graphs of the updated feature maps to obtain the SER maps, and finally sum up all the SER maps to get the saliency map of frame  $t$ .

In this experiment, we use a dataset of 50 video clips and their corresponding eye-tracking data from [13]. The 50 video clips contain outdoor scenes, television broadcast and video games. The eye-tracking data are recorded from

eight subjects. We adopt the evaluation method proposed in [13] to assess the performance of our model. To evaluate a saliency detection model, we first compute the saliency map of a given video using the model. Then we collect two sets of locations from the video, the human saccade locations and random saccade locations. Subsequently, two histograms of saliency values at the two set of locations can be obtained respectively. The Kullback-Leibler (KL) divergence between the two histograms is adopted as the model evaluation criterion. The intuitive idea of this evaluation method is that an effective model predicts high saliency values at human saccadic locations. Thus the saliency value histogram from human saccades locations should be very different from the histogram from random locations, i.e. the larger KL divergence, the better the model. In Table 3, we compare the KL distances<sup>1</sup> to the other two methods [13] and [11] on “beverly03”. The ranking of KL distances shows that our model achieves the best performance.

Table 3. Performance comparison on videos

	Itti et al. [13]	Hou et al. [11]	Our model
KL distance	0.3403	0.5432	0.6927

## 4. Conclusion and Future Work

This paper proposes a computational model of visual saliency, in which the saliency is defined as *Site Entropy Rate* (SER) based on the principle of *information maximization*. The experiments demonstrate that the proposed model achieves the state-of-art performance of saliency detection in both still images and videos.

As can be seen, in this paper we simply model the transition probability between two graph nodes by fusing the dissimilarity of sub-band feature responses and the spatial distance with pre-set parameter ( $\lambda$ ). In the future, we will design a data driven algorithm to learn the transition probability from human saccade.

## Acknowledgments

This work was supported by the National Science Foundation of China under grant No. 90920012 and 60872077, the National Basic Research “973” Program of China under grant No. 2009CB320904. We would like to thank Dashan Gao, Xiaodi Hou and Neil Bruce for providing their source codes for result comparison and patiently answering our questions about experiment settings.

<sup>1</sup>We cite the results of Itti et al.’s method [13] and Hou et al’s method [11] reported in [11], and use the evaluation code from [11] to compute our KL distance.

## References

- [1] H. Barlow. Unsupervised learning. *Neural Computation*, 1989.
- [2] N. D. Bruce and J. Tsotsos. Saliency based on information maximization. *NIPS*, 2006.
- [3] L. Costa. Visual saliency and attention as random walks on complex networks. *physics/0603025*, 2007.
- [4] T. M. Cover and J. A. Thomas. Elements of information theory, second edition. *John Wiley & Sons. Inc., Hoboken, New Jersey*, 2006.
- [5] W. Einhäuser, W. Kruse, K. Hoffmann, and P. König. Differences of monkey and human overt attention under natural conditions. *Vision Research*, 2006.
- [6] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. *NIPS*, 2007.
- [7] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. *CVPR*, 2009.
- [8] D. Graham and D. J. Field. Efficient coding of natural images. *New Encyclopedia of Neuroscience*, 2007.
- [9] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *NIPS*, 2006.
- [10] J. Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B*, 1998.
- [11] X. Hou and L. Zhang. Dynamic visual attention: searching for coding length increments. *NIPS*, 2008.
- [12] L. Itti, C. Koch, and E. Niebur. A model of saliency based visual attention for rapid scene analysis. *IEEE TPAMI*, 1998.
- [13] I. Itti and P. Baldi. Bayesian surprise attracts human attention. *NIPS*, 2006.
- [14] W. McCullagh and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 1943.
- [15] V. Mountcastle. An organizing principle for cerebral function: The unit model and the distributed system. *The Mindful Brain Cambridge, MA: MIT Press*, 1978.
- [16] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996.
- [17] C. Privitera and L. Stark. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE TPAMI*, 2000.
- [18] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 1999.
- [19] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? *CVPR*, 2004.
- [20] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE TPAMI*, 2007.
- [21] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 1980.
- [22] J. Tsotsos, S. Culhane, W. Wai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *AI*, 1995.
- [23] A. Yarbus. Eye movements and vision. *NY: Plenum*, 1967.