Accepted for IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Visual Scene Understanding (ViSU'09), Miami, June, 2009

Nonparametric Bottom-Up Saliency Detection by Self-Resemblance

Hae Jong Seo and Peyman Milanfar
Electrical Engineering Department
University of California, Santa Cruz
1156 High Street, Santa Cruz, CA, 95064
rokaf@soe.ucsc.edu

Abstract

We present a novel bottom-up saliency detection algorithm. Our method computes so-called local regression kernels (i.e., local features) from the given image, which measure the likeness of a pixel to its surroundings. Visual saliency is then computed using the said "self-resemblance" measure. The framework results in a saliency map where each pixel indicates the statistical likelihood of saliency of a feature matrix given its surrounding feature matrices. As a similarity measure, matrix cosine similarity (a generalization of cosine similarity) is employed. State of the art performance is demonstrated on commonly used human eye fixation data [3] and some psychological patterns.

1. Introduction

Visual saliency detection has been of great research interest [3, 6, 8, 10, 13, 26, 31] in recent years. Analysis of visual attention is considered a very important component in the human vision system because of a wide range of applications such as object detection, predicting human eye fixation, video summarization [17], image quality assessment [15, 19] and more. In general, saliency is defined as what drives human perceptual attention. There are two types of computational models for saliency according to what the model is driven by: a bottom-up saliency [3, 6, 10, 13, 31] and a top-down saliency [8, 26]. As opposed to bottom-up saliency algorithms that are fast and driven by low-level features, top-down saliency algorithms are slow and task-driven.

The problem of interest addressed in this paper is bottom-up saliency which can be described as follows: Given an image, we are interested in accurately detecting salient objects from the image without any background knowledge. In order to do this, we propose to use, as features, so-called local steering kernels which capture local data structure exceedingly well. Our approach is motivated by a Bayesian probabilistic framework, which is based on a nonparametric estimate of the likelihood of saliency. As we describe below, this boils down to the local calculation of a "self-resemblance" map, which measures the similarity of a feature matrix at a pixel of interest to its neighboring feature matrices.

1.1. Previous work

Itti et al. [13] introduced a saliency model which was biologically inspired. Specifically, they proposed the use of a set of feature maps from three complementary channels as intensity, color, and orientation. The normalized feature maps from each channel were then linearly combined to generate the overall saliency map. Even though this model has been shown to be successful in predicting human fixations, it is somewhat ad-hoc in that there is no objective function to be optimized and many parameters must be tuned by hand. With the proliferation of eye-tracking data, a number of researchers have recently attempted to address the question of what attracts human visual attention by being more mathematically and statistically precise [3, 6, 7, 8, 12, 31].

Gao et al. [6, 7, 8] proposed a unified framework for top-down and bottom-up saliency as a classification problem with the objective being the minimization of classification error. They first applied this framework to object detection [8] in which a set of features are selected such that a class of interest is best discriminated from all other classes, and saliency is defined as the weighted sum of features that are salient for that class. In [6], they defined bottom-up saliency using the idea that pixel locations are salient if they are distinguished from their surroundings. They used difference of Gaussians (DoG) filters and Gabor filters, measuring the saliency of a point as the Kullback-Leibler (KL) divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region.

Bruce and Tsotsos [3] modeled bottom-up saliency as the maximum information sampled from an image.

More specifically, saliency is computed as Shannon's self-information $-\log p(\mathbf{f})$, where $\mathbf{f}$ is a local visual feature vector (i.e., derived from independent component analysis (ICA) performed on a large sample of small RGB patches in the image.) The probability density function is estimated based on a Gaussian kernel density estimate in a neural circuit.

Oliva and Torralba [20, 26] proposed a Bayesian framework for the task of visual search (i.e., whether a target is present or not.) They modeled bottom-up saliency as $\frac{1}{p(\mathbf{f}|\mathbf{f}_G)}$ where $\mathbf{f}_G$ represents a global feature that summarizes the appearance of the scene and approximated this conditional probability density funtion by fitting to multivariate exponential distribution. Zhang et al. [31] also proposed saliency detection using natural statistics (SUN) based on a similar Bayesian framework to estimate the probability of a target at every location. They also claimed that their saliency measure emerges from the use of Shannon's self-information under certain assumptions. They used ICA features as similarly done in [3], but their method differs from [3] in that natural image statistics were applied to determine the density function of ICA featuers. Itti and Baldi [12] proposed so-called "Bayesian Surprise" and extended it to the video case [11]. They measured KL-divergence between a prior distribution and posterior distribution as a measure of saliency.

Most of the methods [6, 13, 20] based on Gabor or DoG filter responses require many design parameters such as the number of filters, type of filters, choice of the nonlinearities, and a proper normalization scheme. These methods tend to emphasize textured areas as being salient regardless of their context. In order to deal with these problems, [3, 31] adopted non-linear features that model complex cells or neurons in higher levels of the visual system. Kienzle et al. [14] further proposed to learn a visual saliency model directly from human eyetracking data using a support vector machine (SVM).

Different from traditional image statistical models, a spectral residual (SR) approach based on the Fourier transform was recently proposed by Hou and Zhang [10]. Spectral residual does not rely on parameters and detects saliency rapidly. In this approach, the difference between the log spectrum of an image and its smoothed version is the spectral residual of the image. However, Guo and Zhang [9] claimed that what plays an important role for saliency detection is not SR, but the image's phase spectrum.

1.2. Overview of the Proposed Approach

In this paper, our contributions to the saliency detection task are two-fold. First we propose to use local regression kernels as features which capture the underlying local structure of the data exceedingly well, even in the presence of significant distortions. Second we propose to use a

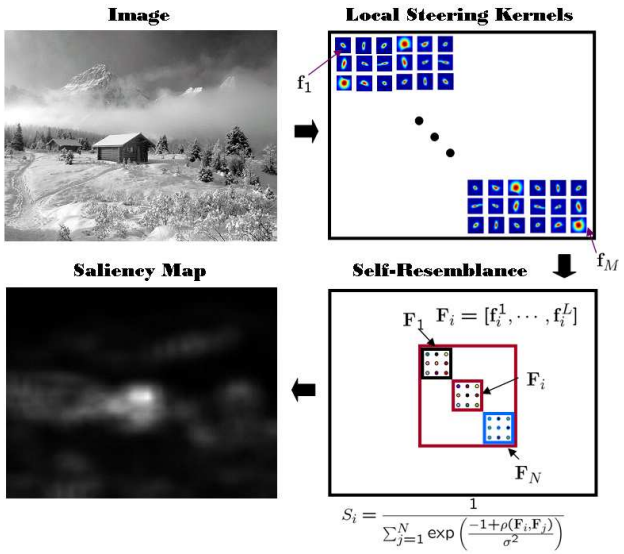

Figure 1. Graphical overview of saliency detection system

nonparametric kernel density estimation for such features, which results in a saliency map consisting of local "self-resemblance" measure, indicating likelihood of saliency. The original motivation behind these contributions is the earlier work on adaptive kernel regression for image reconstruction [24] and nonparametric object detection [21].

As similarly done in Gao et al. [6], we measure saliency at a pixel in terms of how much it stands out from its surroundings. To formalize saliency at each pixel, we let the binary random variable y_i denote whether a pixel position $\mathbf{x}_i = [x_1, x_2]_i^T$ is salient or not as follows:

$$y_i = \begin{cases} 1, & \text{if } \mathbf{x}_i \text{ is salient,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $i = 1, \cdots, M$, and M is the total number of pixels in the entire image. Motivated by the approach in [31, 20], we define saliency at pixel position $\mathbf{x}_i$ as a posterior probability $Pr(y_i = 1|\mathbf{F})$ as follows:

$$S_i = Pr(y_i = 1|\mathbf{F}), \tag{2}$$

where the feature matrix, $\mathbf{F}_i = [\mathbf{f}_i^1, \cdots, \mathbf{f}_i^L]$ at pixel of interest $\mathbf{x}_i$ (what we call a center feature,) contains a set of feature vectors ($\mathbf{f}_i$) in a local neighborhood where L is the number of features in that neighborhood. In turn, the larger collection of features $\mathbf{F} = [\mathbf{F}_1, \cdots, \mathbf{F}_N]$ is a matrix containing features not only from the center, but also a surrounding region (what we call a center+surround region, See Fig. 2.) N is the number of feature matrices in the center+surround region. Using Bayes' theorem, Equation (2) can be written as

$$S_i = Pr(y_i = 1|\mathbf{F}) = \frac{p(\mathbf{F}|y_i = 1)Pr(y_i = 1)}{p(\mathbf{F})}. \tag{3}$$

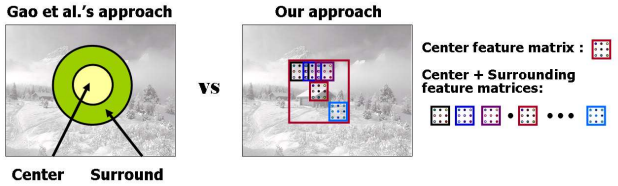

Figure 2. Illustration of difference between Gao el al. [6]'s approach and our approach about a center-surround definition.

By assuming that 1) a-priori $Pr(y_i = 1)$, every pixel is considered to be equally likely to be salient; and 2) $p(\mathbf{F})$ are uniform over features, the saliency we defined boils down to the conditional probability density $p(\mathbf{F}|y_i = 1)$.

Since we do not know the conditional probability density $p(\mathbf{F}|y_i = 1)$, we need to estimate it. It is worth noting that Gao et al. [6] and Zhang et al. [31] have tried to fit a marginal density of local feature vectors $p(\mathbf{f})$ to a generalized Gaussian distribution. However, in this paper, we approximate the conditional density function $p(\mathbf{F}|y_i = 1)$ based on nonparametric kernel density estimation which will be explained in detail in Section 2.2.

Before we begin a more detailed description, it is worthwhile to highlight some aspects of the proposed framework. While the state-of-the art methods [3, 6, 12, 31] are related to our method, their approaches fundamentally differ from ours in the following respects: 1) While they use Gabor filter, DoG filter, or ICA feature as features, we propose to use local steering kernels (LSK) which are highly nonlinear and stable in the presence of uncertainty in the data [24]. In addition, normalized local steering kernels provide a certain invariance to changes as shown in Fig. 3; 2) As opposed to [6, 31] modeling marginal densities of band-pass features as a generalized Gaussian distribution, we estimate the conditional probability density $p(\mathbf{F}|y_i = 1)$ using the idea of nonparametric kernel density estimation; 3) While Itti and Baldi [12] computed, as a measure of saliency, KL-divergence between a prior and a posterior distribution, we explicitly estimate the likelihood function using nonparametric kernel density estimation. From a practical standpoint, it is important to note that our method is appealing because it is nonparametric. Fig. 1 shows an overview of our proposed framework for saliency detection. To summarize the operation of the overall algorithm, we first compute the normalized local steering kernels from the given image I and vectorize them as $\mathbf{f}$'s. Then, we identify $\mathbf{F}_i$ at a pixel of interest $\mathbf{x}_i$ and a set of feature matrices $\mathbf{F}_j$ in a center+surrounding region and compute the self-resemblance measure (See Equations (9) and (10).) The final saliency map is given as a density map as shown in Fig 1.

In the next section, we provide further technical details about the steps outlined above. In Section 3, we demonstrate the performance of the system with experimental results, and we conclude this paper in Section 4.

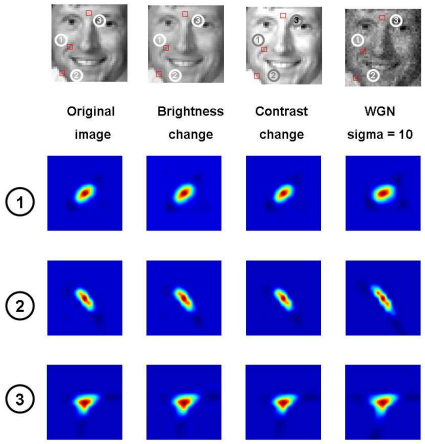

Figure 3. Invariance and robustness of LSK weights $W(\mathbf{x}_l - \mathbf{x}_i)$ in various challenging conditions. Note that WGN means White Gaussian Noise.

2. Technical Details

2.1. Local Steering Kernel as a Feature

The key idea behind local steering kernel is to robustly obtain the local structure of images by analyzing the radiometric (pixel value) differences based on estimated gradients, and use this structure information to determine the shape and size of a canonical kernel. The local steering kernel is modeled as

$$K(\mathbf{x}_l - \mathbf{x}_i) = \frac{\sqrt{\det(\mathbf{C}_l)}}{h^2} \exp\left\{\frac{(\mathbf{x}_l - \mathbf{x}_i)^T \mathbf{C}_l (\mathbf{x}_l - \mathbf{x}_i)}{-2h^2}\right\}, \quad (4)$$

where $l \in \{1, \cdots, P\}$, P is the number of pixels in a local window, h is a global smoothing parameter, and the matrix $\mathbf{C}_l$ is a covariance matrix estimated from a collection of spatial gradient vectors within the local analysis window around a sampling position $\mathbf{x}_l = [x_1, x_2]_l^T$.

In what follows, at a position $\mathbf{x}_i$, we will essentially be using (a normalized version of) the function $K(\mathbf{x}_l - \mathbf{x}_i)$. To be more specific, the local steering kernel function $K(\mathbf{x}_l - \mathbf{x}_i)$ is calculated at every pixel location and normalized as follows

$$W(\mathbf{x}_l - \mathbf{x}_i) = \frac{K(\mathbf{x}_l - \mathbf{x}_i)}{\sum_{l=1}^{P} K(\mathbf{x}_l - \mathbf{x}_i)}, \quad i = 1, \cdots, M. \quad (5)$$

It is worth noting that LSK reliably captures local data structures even in complex texture regions or in the presence of moderate levels of noise. Normalization of this kernel function yields invariance to brightness change and robustness to contrast change as shown in Fig. 3.

From a human perception standpoint [26], it has been shown that local image features are salient when they are distinguishable from the background. Computationally, measuring saliency requires, as we have seen, the estimation of local feature distributions in an image. For this pur-

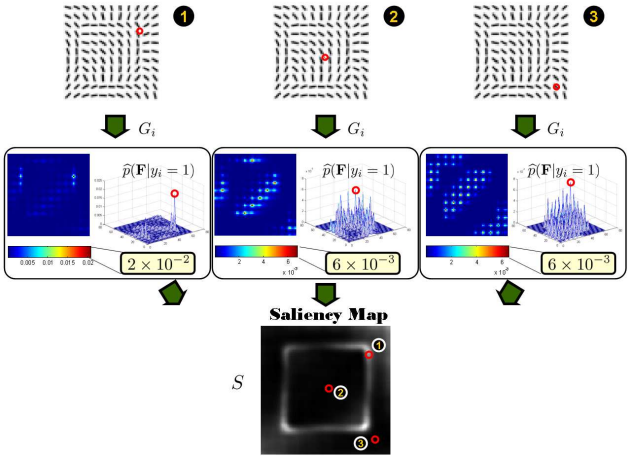

Figure 4. Example of saliency computation in psychological pattern. Note that center+surrounding regions to compute Self-Resemblance is as large as the entire image in this case. i.e., $N = M$

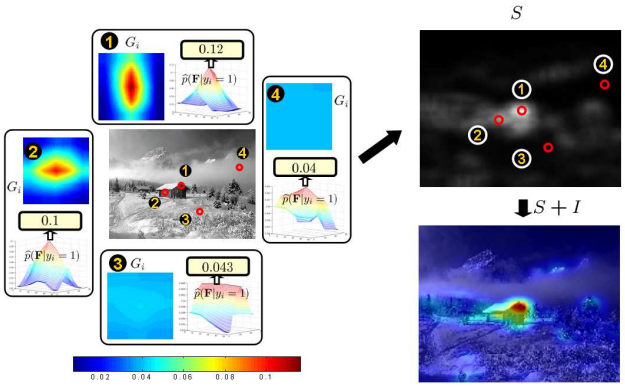

Figure 5. Example of saliency computation in natural gray-scale image. Note that center+surrounding regions to compute self-resemblance is a local neighborhood in this case. i.e., $N << M$. Note that red values in saliency map represent higher saliency, while blue values mean lower saliency.

pose, a generalized Gaussian distribution is often employed as in [6, 26, 31].

However, LSK features follow a power-law distribution (a long-tail distribution) [21]. In other words, the LSK features are scattered out in a high dimensional feature space, and thus there basically exists no dense cluster in the feature space. Instead of using a generalized Gaussian distribution, we employed a locally adaptive kernel density estimation method which we explain in the next section.

2.2. Saliency by Self-Resemblance

As we alluded to in Section 1.2, saliency at a pixel $\mathbf{x}_i$ is measured using the conditional density of the feature matrix at that position: $S_i = p(\mathbf{F}|y_i = 1)$. Hence, the task at hand is to estimate $p(\mathbf{F}|y_i = 1)$ over $i = 1, \cdots, M$. In general, the Parzen density estimator is a simple and generally accurate non-parametric density estimation method [23]. However, in higher dimensions and with an expected long-tail distribution, Parzen density estimator with an isotropic kernel is not the most appropriate method [1, 2, 29]. As explained earlier, the LSK features tend to generically come from long-tailed distributions, and as such, there are generally no tight clusters in the feature space. When we estimate a probability density at a particular feature point, for instance $\mathbf{F}_i = [\mathbf{f}_i^1, \cdots, \mathbf{f}_i^L]$ (where L is the number of vectorized LSKs ($\mathbf{f}$'s) employed in the feature matrix), the isotropic kernel centered on that feature point will spread its density mass equally along all the feature space directions, thus giving too much emphasis to irrelevant regions of space and too little along the manifold. Earlier studies [1, 2, 29] also pointed out this problem. This motivates us to use *a locally data-adaptive version of the kernel density estimator.* We define the conditional probability density $p(\mathbf{F}|y_i = 1)$ at $\mathbf{x}_i$ as a center value of a normalized adaptive kernel (weight function) $G(\cdot)$ computed in the center-surrounding region as follows:

$$\widehat{p}(\mathbf{F}|y_i = 1) = \frac{G_i(\overline{\mathbf{F}}_i - \overline{\mathbf{F}}_i)}{\sum_{j=1}^{N} G_i(\overline{\mathbf{F}}_i - \overline{\mathbf{F}}_j)}, \tag{6}$$

where $G_i(\overline{\mathbf{F}}_i - \overline{\mathbf{F}}_j) = \exp\left(\frac{-||\overline{\mathbf{F}}_i - \overline{\mathbf{F}}_j||_F^2}{2\sigma^2}\right)$, $||\cdot||_F$ is Frobenious norm, $\overline{\mathbf{F}}_i = \left[\frac{\mathbf{f}_i^1}{\|\mathbf{F}_i\|_F}, \cdots, \frac{\mathbf{f}_i^L}{\|\mathbf{F}_i\|_F}\right]$ and $\overline{\mathbf{F}}_j = \left[\frac{\mathbf{f}_j^1}{\|\mathbf{F}_j\|_F}, \cdots, \frac{\mathbf{f}_j^L}{\|\mathbf{F}_j\|_F}\right]$, and σ is a parameter controlling the fall-off of weights.

Inspired by earlier works such as [4, 5, 16, 21] that have shown the effectiveness of correlation-based similarity, the kernel function G_i in Equation (6) can be rewritten using the concept of matrix cosine similarity [21] as follows:

$$G_i(\overline{\mathbf{F}}_i - \overline{\mathbf{F}}_j) = \exp\left(\frac{-1 + \rho(\mathbf{F}_i, \mathbf{F}_j)}{\sigma^2}\right), \quad j = 1, \cdots, N, \tag{7}$$

where $\rho(\mathbf{F}_i, \mathbf{F}_j)$ is the "Matrix Cosine Similarity (MCS)" between two feature matrices $\mathbf{F}_i, \mathbf{F}_j$ and is defined as the "Frobenius inner product" between two normalized matrices $\rho(\mathbf{F}_i, \mathbf{F}_j) =< \overline{\mathbf{F}}_i, \overline{\mathbf{F}}_j >_F = \text{trace}\left(\frac{\mathbf{F}_i^T \mathbf{F}_j}{\|\mathbf{F}_i\|_F \|\mathbf{F}_j\|_F}\right) \in [-1, 1]$. This matrix cosine similarity can be rewritten as a weighted sum of the standard cosine similarities [4, 5, 16] $\rho(\mathbf{f}_i, \mathbf{f}_j)$ between each pair of corresponding feature vectors (i.e., columns) in $\mathbf{F}_i, \mathbf{F}_j$ as follows:

$$\rho_i = \sum_{\ell=1}^{L} \frac{\mathbf{f}_i^{\ell T} \mathbf{f}_j^{\ell}}{\|\mathbf{F}_i\|_F \|\mathbf{F}_j\|_F} = \sum_{\ell=1}^{L} \rho(\mathbf{f}_i^{\ell}, \mathbf{f}_j^{\ell}) \frac{\|\mathbf{f}_i^{\ell}\| \|\mathbf{f}_j^{\ell}\|}{\|\mathbf{F}_i\|_F \|\mathbf{F}_j\|_F}. \tag{8}$$

The weights are represented as the product of $\frac{\|\mathbf{f}_i^{\ell}\|}{\|\mathbf{F}_i\|_F}$ and $\frac{\|\mathbf{f}_j^{\ell}\|}{\|\mathbf{F}_j\|_F}$ which indicate the relative importance of each fea-

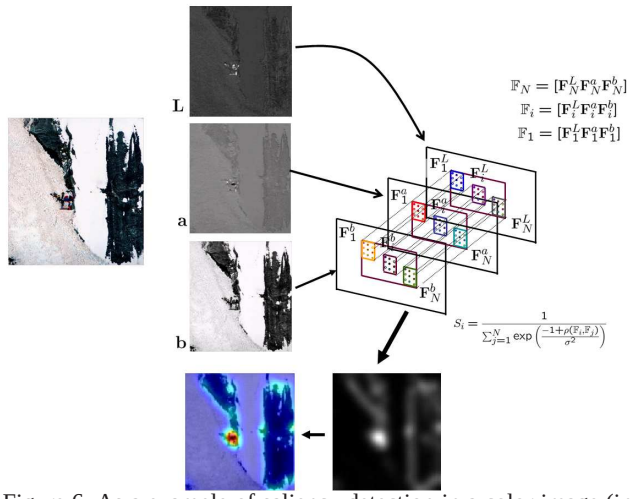

Figure 6. As a example of saliency detection in a color image (in this case, CIE L*a*b*), we show how saliency is computed using matrix cosine similarity.

ture in the feature sets $\mathbf{F}_i, \mathbf{F}_j$. This measure [1] not only generalizes the cosine similarity, but also overcomes the disadvantages of the conventional Euclidean distance which is sensitive to outliers.

Fig. 4 describes what kernel functions G_i look like in various regions of a psychological pattern image[2]. As shown in Fig. 4, each kernel function G_i has a unique peak value at $\mathbf{x}_i$ which represents a likelihood of the pixel $\mathbf{x}_i$ being salient given feature matrices in the center+surrounding region. Therefore, saliency at $\mathbf{x}_i$ ($S_i = \widehat{p}(\mathbf{F}|y_i = 1)$) is the center value of (the normalized version) of the weight function G_i which already contains contributions from all the surrounding feature matrices as follows:

$$S_i = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(\mathbf{F}_i,\mathbf{F}_j)}{\sigma^2}\right)}. \tag{9}$$

As a consequence, $\widehat{p}(\mathbf{F}|y_i = 1)$ reveals how $\mathbf{F}_i$ is salient given all the features $\mathbf{F}_j$'s in a neighborhood. Fig. 5 illustrates how these values computed from a natural image provide a reliable saliency measure.

2.3. Handling color images

Up to now, we only dealt with saliency detection in a grayscale image. If we have a color input image, we need an approach to integrate saliency information from all color channels. To avoid some drawbacks of earlier methods [13, 18], we do not combine saliency maps from each color channel linearly and directly. Instead we utilize

[1] This measure can be efficiently implemented by column-stacking the matrices $\mathbf{F}_i, \mathbf{F}_j$ and simply computing the cosine similarity between two long column vectors.

[2] The image came from the website, http://www.svcl.ucsd.edu/projects/discsalbu/

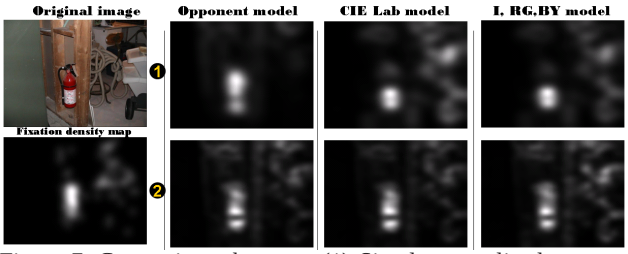

Figure 7. Comparisons between (1) Simple normalized summation and (2) The use of matrix cosine similarity without any fusion in three different color spaces. Simple normalized summation method tends to be dominated by a particular chrominance information. It is clearly shown that using matrix cosine similarity provides consistent results than the simple normalized summation fusion method.

the idea of matrix cosine similarity. More specifically, we first identify feature matrices from each color channel as $\mathbf{F}_i^{c_1}, \mathbf{F}_i^{c_2}, \mathbf{F}_i^{c_3}$, where c_1, c_2, c_3 represent each color channel as shown in Fig. 6. By collecting them as a larger matrix $\mathbb{F}_i = [\mathbf{F}_i^{c_1}, \mathbf{F}_i^{c_2}, \mathbf{F}_i^{c_3}]$, we can apply matrix cosine similarity between $\mathbb{F}_i$ and $\mathbb{F}_j$. Then, the saliency map from color channels can be analogously defined as follows:

$$S_i = \widehat{p}(\mathbb{F}|y_i = 1) = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(\mathbb{F}_i,\mathbb{F}_j)}{\sigma^2}\right)}. \tag{10}$$

In order to verify that this idea allows us to achieve a consistent result and leads us to a better performance than using fusion methods, we have compared three different color spaces [3] : Opponent color channels [28], CIE L*a*b* [21, 22] channels, and I R-G B-Y channels [31]

Fig. 7 compares saliency maps using simple normalized summation of saliency maps from different channels as compared to using matrix cosine similarity. It is clearly shown that using matrix cosine similarity provides consistent results regardless of color spaces and helps to avoid drawback of fusion methods. To summarize, the overall pseudo-code for the algorithm is given in **Algorithm** 1.

3. Experimental Results

3.1. Predicting human visual fixation data

In this section, we show several experimental results on detecting saliency in natural images. We used an image dataset and its fixation data collected by Bruce and Tsotsos [3] as a benchmark for comparison. This dataset contains eye fixation records from 20 subjects for a total of 120 images of size 681×511. Given an image I, we downsample it to an appropriate scale (86×64, 8 times fewer pixels) in order to reduce the time-complexity. We then compute LSK

[3] Opponent color space has proven to be superior to RGB, HSV, normalized RGB, and more in the task of object and scene recognition [28]. Shechman and Irani [22] and Seo and Milanfar [21] showed that CIE L*a*b* performs well in the task of object detection.

Algorithm 1 Saliency Detection Algorithm

I : image, P : size of local steering kernel (LSK) window, h : a global smoothing parameter for LSK, L : number of LSK used in the feature matrix, N : size of a center+surrounding region for computing self-resemblance, σ : a parameter controlling fall-off of weights for computing self-resemblance.

Stage1 : Compute Features

Compute the normalized LSK W_i and vectorize it to $\mathbf{f}_i$, where $i = 1, \cdots, M$.

Stage2 : Compute Self-Resemblance

for $i = 1, \cdots, M$ **do**

  if I is a grayscale image **then**

    Identify feature matrices $\mathbf{F}_i$,$\mathbf{F}_j$ in a local neighborhood.

    $S_i = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(\mathbf{F}_i,\mathbf{F}_j)}{\sigma^2}\right)}$

  else

    Identify feature matrices $\mathbb{F}_i = [\mathbf{F}_i^{C1}, \mathbf{F}_i^{C3}, \mathbf{F}_i^{C3}]$ and $\mathbb{F}_j = [\mathbf{F}_j^{C1}, \mathbf{F}_j^{C3}, \mathbf{F}_j^{C3}]$ in a local neighborhood from three color channels.

    $S_i = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(\mathbb{F}_i,\mathbb{F}_j)}{\sigma^2}\right)}$

  end if

end for

Output : Saliency map $S_i, \quad i = 1, \cdots, M$

of size 3×3 as features and generate feature matrices $\mathbf{F}_i$ in a 7×7 local neighborhood. The smoothing parameter h for computing LSK was set to 0.008 and the fall-off parameter σ for computing self-resemblance was set to 0.07 for all the experiments. We obtained an overall saliency map by using CIE L*a*b* color space throughout all the experiments. Some visual results of our model are compared with state-of-the-art methods in Fig. 8. As opposed to Bruce's method [3] which is quite sensitive to textured regions and SUN [31] which is somewhat better in this respect, the proposed method is much less sensitive to background texture. We also computed the area under receiver operating characteristic (ROC) curve and KL-divergence by following the experimental protocol of [31]. In [31], Zhang et al. pointed out that the dataset collected by Bruce [3] is center-biased and the methods by Itti et al. [13], Bruce et al. [3] and Gao et al. [6] are all corrupted by edge effects which resulted in relatively higher performance than they should have (See Fig. 9.). We compare our model against Itti et al.[4] [13], Bruce and Tsotsos[5] [3], Gao et al. [6], and SUN[6] [31]. For the evaluation of the algorithm, we used the same procedure as in [31]. More specifically, the shuffling of the saliency maps is repeated 100 times. Each time, KL-divergence is computed between the histograms of unshuffled saliency and shuffled saliency on human fixations. When calculating the area under the ROC curve, we also used 100 random permutations. The mean and the standard errors are reported in Table 1. Our model outperforms all the other state-of-the-art methods in terms of both KL-divergence and ROC area.

[4]Downloadable from http://ilab.usc.edu/toolkit/home.shtml

[5]Downloadable from http://web.me.com/john.tsotsos/Visual_Attention/ST_and_Saliency.html

[6]Downloadable from http://www.roboticinsect.net/index.htm

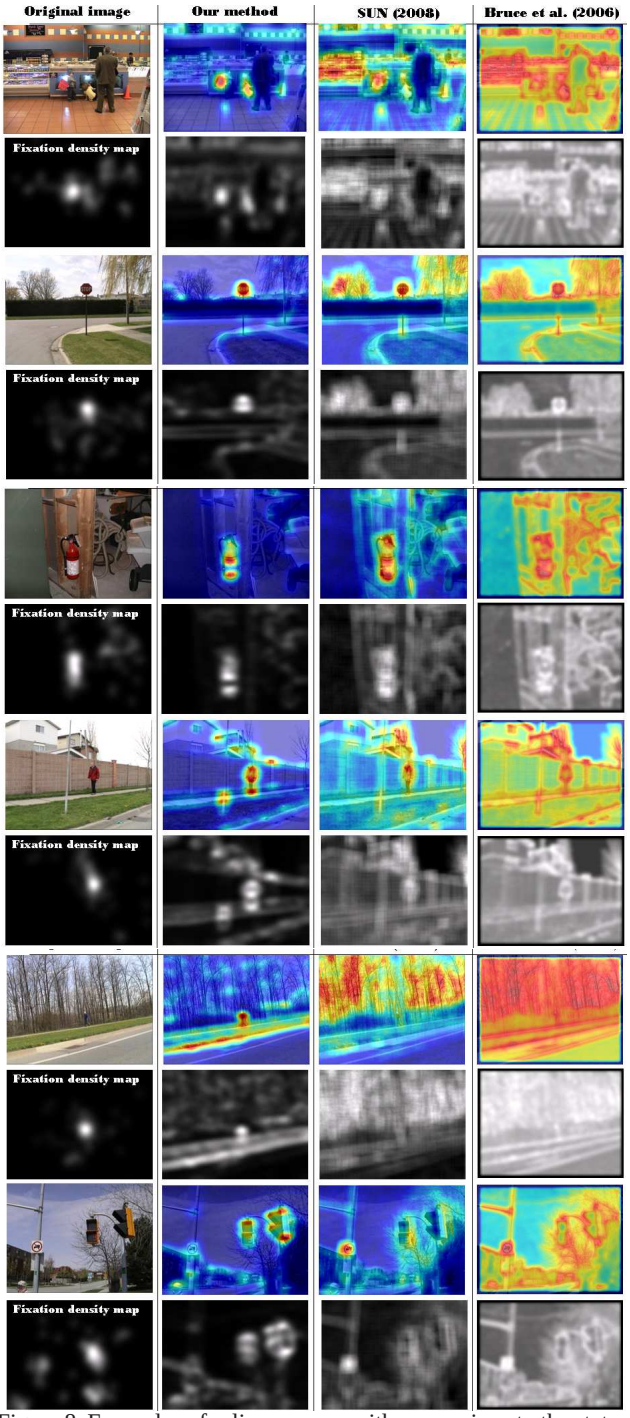

Figure 8. Examples of saliency maps with comparison to the state-of-the-art methods. Visually, our method outperforms other state-of-the-art methods.

3.2. Psychological Pattern

We also tested our method on psychological patterns. Psychological patterns are widely used in attention experiments not only to explore the mechanism of visual search,

Figure 9. Comparison of average saliency maps on human fixaton data by Bruce and Tsotsos [3]. Averages were taken across the saliency maps for a total of 120 color images. Note that Bruce et al.'s method [3] exhibits zero values at the image borders while SUN [31] and our method do not have edge effects

Table 1. Performance in predicting human eye fixations when viewing color images.

Model	KL (SE)	ROC (SE)
Itti *et al.* [13]	0.1130 (0.0011)	0.6146 (0.0008)
Bruce and Tsotsos [3]	0.2029 (0.0017)	0.6727 (0.0008)
Gao *et al.* [6]	0.1535 (0.0016)	0.6395 (0.0007)
Zhang *et al.* [31]	0.2097 (0.0016)	0.6570 (0.0008)
Our method	**0.3432** (0.0029)	**0.6769** (0.0008)

but also to test effectiveness of saliency maps [27, 30]. As shown in Fig. 10, whereas SUN [31] and Bruce's method [3] failed to capture perceptual difference in most cases, Gao's method [6] and Spectral Residual [10] tend to capture perceptual organization rather better. Overall, however, the proposed saliency algorithm outperforms other methods in all cases including closure pattern (Fig. 10 (a)) and texture segregation (Fig. 10 (b)) which seem to be very difficult even for humans to distinguish.

4. Conclusion and Future work

In this paper, we have proposed a bottom-up saliency detection algorithm by employing *local steering kernels*; and by using a nonparametric kernel density estimation based on "Matrix Cosine Similarity" (MCS). The proposed method can automatically detect salient objects in the given image. The proposed method is practically appealing because it is nonparametric and robust to the uncertainty in the data. Challenging sets of real-world human fixation data experiments demonstrated that the proposed saliency detection method achieves a high degree of accuracy and improves upon state-of-the-art methods. The proposed framework is general enough as to be extendable to space-time saliency detection using 3-D LSKs [25]. Due to its robustness to noise and other systemic perturbations, we also expect the present framework to be quite effective in other applications such as image quality assessment and video summarization.

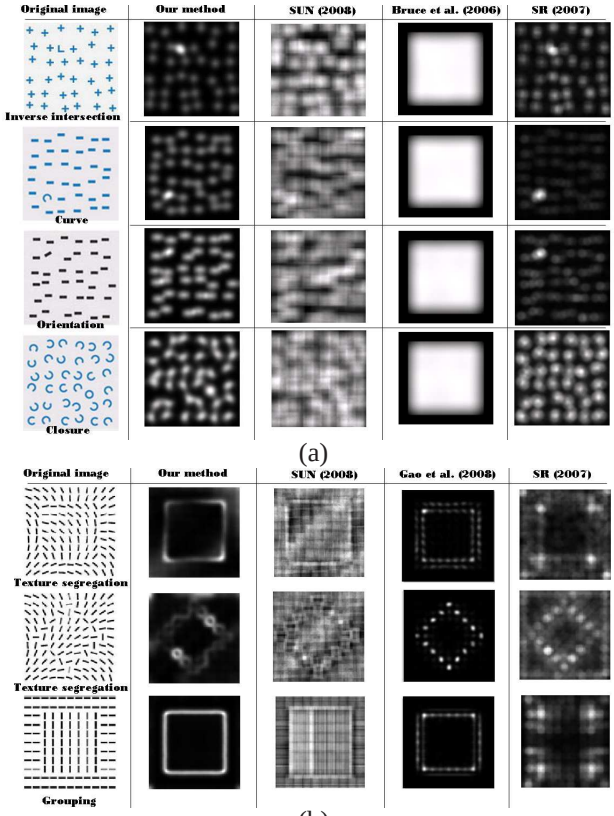

Figure 10. Examples of Saliency map on psychological patterns. (a) images are from [10] (b) images are from [6].

Acknowledgment

The authors would like to thank Neil Bruce and John K. Tsotsos for kindly sharing their human fixation data; and Lingyun Zhang for sharing her Matlab codes and helpful discussion. This work was supported by AFOSR Grant FA 9550-07-01-0365.

References

[1] Y. Bengio, H. Larochelle, and P. Vincent. Non-local manifold parzen windows. *In Advances in Neural Information Processing Systems (NIPS)*, 18:115–122, 2005. 4

[2] T. Brox, B. Rosenhahn, and H.-P. S. D. Cremers. Nonparametric density estimation with adaptive anisotropic kernels for human motion tracking. *2nd. Workshop on Human Motion, Springer-Verlag Berlin Heidelberg (LNCS)*, 4814:152–165, 2007. 4

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. *In Advances in Neural Information Processing Systems*, 18:155–162, 2006. 1, 2, 3, 5, 6, 7

[4] Y. Fu and T. S. Huang. Image classification using correlation tensor analysis. *IEEE Transactions on Image Processing*, 17(2):226–234, 2008. 4

[5] Y. Fu, S. Yan, and T. S. Huang. Correlation metric for generalized feature extraction. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence, 30(12):2229–2235, 2008. 4

[6] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7):13,1–18, 2008. 1, 2, 3, 4, 6, 7

[7] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. *In Advances in Neural Information Processing Systems*, 17:481–488, 2004. 1

[8] D. Gao and N. Vasconcelos. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:282–287, 2005. 1

[9] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 1, 2, 7

[11] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:631–637, 2005. 2

[12] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *In Advances in Neural Information Processing Systems*, 18:1–8, 2006. 1, 2, 3

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transcations on Pattern Analysis and Machine Intelligence (PAMI)*, 20:1254–1259, 1998. 1, 2, 5, 6, 7

[14] W. Kienzle, F. Wichmann, B. Scholkopf, and M. Franz. A nonparametric approach to bottom-up visual saliency. *In Advances in Neural Information Processing Systems*, 19:689–696, 2007. 2

[15] Q. Ma and L. Zhang. Saliency-based image quality assessment criterion. *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues (LNCS)*, 5226:1124–1133, 2008. 1

[16] Y. Ma, S. Lao, E. Takikawa, and M. Kawade. Discriminant analysis in correlation similarity measure space. *International Conference on Machine Learning*, 227:577–584, 2007. 4

[17] S. Marat, M. Guironnet, and D. Pellerin. Video summarization using a visual attentional model. *EUSIPCO, EURASIP*, pages 1784–1788, 2007. 1

[18] O. Meur, P. L. Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47:2483–2498, 2007. 5

[19] A. Niassi, O. LeMeur, P. Lecallet, and D. barba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. *IEEE International Conference on Image Processing (ICIP)*, 2:169–172, 2007. 1

[20] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. *In Proceedings of International Conference on Image Processing*, pages 253–256, 2003. 2

[21] H. J. Seo and P. Milanfar. Training-free, generic object detection using locally adaptive regression kernels. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, October 2008. 2, 4, 5

[22] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. *In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007. 5

[23] B. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability 26, New York: Chapman & Hall, 1986. 4

[24] H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2):349–366, February 2007. 2, 3

[25] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Super-resolution without explicit subpixel motion estimation. *Submitted to IEEE Transactions on Image Processing*, 2008. 7

[26] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113:766–786, 2006. 1, 2, 3, 4

[27] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. 7

[28] K. vande Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 5

[29] P. Vincent and Y. Bengio. Manifold parzen windows. *In Advances in Neural Information Processing Systems (NIPS)*, 15:825–832, 2003. 4

[30] J. Wolfe. Guided search 2.0: A revised model of guided search. *Psychonomic bulletic and Rivew*, 1:202–238, 1994. 7

[31] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32,1–20, 2008. 1, 2, 3, 4, 5, 6, 7