An Object-oriented Visual Saliency Detection Framework Based on Sparse Coding Representations

Junwei Han, Sheng He, Xiaoliang Qian, Dongyang Wang, Lei Guo, and Tianming Liu

***Abstract*—Saliency detection aims at quantitatively predicting attended locations in an image. It may mimic the selection mechanism of the human vision system, which processes a small subset of a massive amount of visual input while the redundant information is ignored. Motivated by the biological evidence that the receptive fields of simple-cells in V1 of the vision system are similar to sparse codes learned from natural images, this paper proposes a novel framework for saliency detection by using image sparse coding representations as features. Unlike many previous approaches dedicated to examining the local or global contrast of each individual location, this paper develops a probabilistic computational algorithm by integrating objectness likelihood with appearance rarity. In the proposed framework, image sparse coding representations are yielded through learning on a large amount of eye-fixation patches from an eye-tracking dataset. The objectness likelihood is measured by three generic cues called compactness, continuity, and center bias. The appearance rarity is inferred by using a Gaussian Mixture Model. The proposed work can serve as a basis for many techniques such as image/video segmentation, retrieval, retargeting, and compression. Extensive evaluations on benchmark databases and comparisons with a number of up-to-date algorithms demonstrate its effectiveness.**

***Index Terms*—Visual attention, Saliency, Sparse coding, Independent Component Analysis, Gaussian Mixture Models**

I. INTRODUCTION

The advancement of computer vision technology is limited largely by the profound challenge of automatically identifying the object of interest in an image. An attempt at simulating the human visual attention mechanism potentially promises to resolve this problem. The intrinsic attribute of visual attention is its selection procedure, which enables our vision system to select a subset of interesting inputs in the visual field for further cognition. This selection relies on the synergy of both bottom-up and top-down factors. Unlike top-down attention which is driven by task, bottom-up attention is driven by saliency. So far, much less is known about top-down attention and its quantitative calculation is still practically impossible since it involves numerous cognitive cues like the observer's background, expectations, and preferences. In contrast, bottom-up attention is simpler and plays a critical role under the scenario of free viewing. Therefore, the study of computation of bottom-up attention is becoming popular. The core component of computational models is called the saliency map invented by Koch et al. [1], which is defined as a 2D topographical map encoding the conspicuity at every location of the image. In recent years, the use of saliency map has benefited a broad range of applications such as image segmentation [2], image/video retargeting [4], video summarization [6], image retrieval [7], image collage [8], video coding [9], and so on.

Manuscript received on May 2, 2012. This work was supported in part by the National Science Foundation of China under Grant 61005018 and 91120005, NPU-FFR-JC20120237, and Program for New Century Excellent Talents in University under grant NCET-10-0079.

Junwei Han, Sheng He, Xiaoliang Qian, Dongyang Wang, and Lei Guo are with School of Automation, Northwestern Polytechnical University, Xi'an, China. (phone: 86-29-88431318; fax: phone: 86-29-88431318;e-mail: junweihan2010@gmail.com).

Tianming Liu is with the Computer Science Department, University of Georgia, Boyd 420, Athens, USA, (e-mail: tliu@cs.uga.edu).

A. *Previous works*

Most approaches to calculating saliency map are based on the observation that locations in the visual field that are distinctive from their contextual background are more likely to attract human attention. The distinctiveness or rarity can be measured by contrast. A milestone work was presented by Itti et al. [10]. It developed a biologically plausible system that invented a "center-surround" operation implemented using a "Difference of Gaussians" (DoG) across multiple scales to model the contrast. The final saliency map was derived by the linear summation of color, intensity, and orientation contrast. Likewise, Ma et al. [11] adopted the "Difference of Windows" (DoW) to calculate color distribution distance between a location and its surrounding location within a window to measure contrast. A work similar to [11] was proposed by Achanta et al. [12], which also leveraged DoW to determine visually salient regions. Recently, Klein et al. [43] detected the saliency in an information-theoretic paradigm, which estimates the distribution difference of visual features between the center and its surround regions by Kullback-Leibler divergence. Other representative works using center-surround mechanisms include those in [13-15] and [32]. The works of [16] and [17] investigated the use of Gaussian Mixture Models (GMMs) for the saliency computational model. The former method [16] adopted GMMs to represent the dominant hue in which the inter-cluster distance between components indicates the saliency. The overall saliency map was automatically selected

as either a color saliency map or an orientation saliency map. The latter approach [17] involved two GMMs that represented attention regions and background, respectively. A pixel was classified into salient regions or background depending on a Bayesian framework. Harel et al. [31] exploited a graph model in which each node represents a lattice and the connection between two nodes is proportional to their dissimilarity. The contrast was inferred by a Markov chain. Goferman et al. [29] combined local contrast, global contrast, visual organizational rules, and high-level cues to form a new type of saliency called context-aware saliency.

Another school of methods [18-20] explored supervised learning methodologies for saliency detection. A number of attention features were firstly extracted. Afterwards, the feature weights were learned based on a ground truth database manually labeled or obtained by eye-tracking experiments. The data in the ground truth indicates the objects of interest or human eye fixations. Finally, the saliency map was generated according to the weighted combination of features. The supervised learning algorithms used in previous works [18-20] include Conditional Random Fields [18], Support Vector Machines (SVMs) [19], and Mixture of SVMs [20].

Based on the assumption that the global contrast is preferable than the local contrast for saliency detection, a newly emerging research stream on modeling image saliency with high computational efficiency in the frequency domain is gaining interest. In [21], the Spectral Residual (SR) defined as the difference between the log Fourier amplitude spectrum of an image and the prior knowledge was used for saliency discovery. Nevertheless, Guo et al. [22] argued that the SR of the amplitude spectrum is indecisive. Alternatively, they explored the saliency using the phase spectrum of the Fourier Transform. In [23], Achanata et al. provided a frequency-tuned (FT) approach to capture global contrast. Alternatively, Hou et al. [34] employed the sign of each Discrete Cosine Transform component, which is equivalent to the phase information of the Fourier transformation. Recently, Li et al. [33] combined global contrast from frequency domain and local contrast from spatial domain for the generation of a saliency map.

Motivated by the biological evidence that the receptive fields of simple cells in the primary visual cortex (V1) are similar to sparse codes learned from natural image patches, researchers [24-28, 44] have attempted to leverage sparse representations to compute visual saliency. The initial step was to learn basis functions by performing Independent Component Analysis (ICA) on a large number of randomly selected image patches. The learned basis functions were applied to filter the image, thus obtaining a set of coefficients as the features. Then, various principles such as Information Maximization (IM) [24], Incremental Coding Length (ICL) [25], Bayesian framework [26], Site Entropy Rate (SER) [27], and Feature Activation Rate [44] were used to detect the distinctiveness in images. The sparse representation based methods are biologically plausible.

B. *Overview and contribution of the proposed approach*

Although visual saliency detection has been studied extensively, many existing approaches still suffer from such drawbacks as low resolution, ill-defined salient boundary, non-uniform entire salient object, and so on, as summarized in [23]. Most of these drawbacks result from the fact that existing algorithms only take the appearance rarity or distinctiveness into consideration and ignore the objectness cues. Essentially, the underlying purpose of saliency detection is to locate meaningful objects that are more likely to attract the user's attention. From the viewpoint of considering objectness, saliency detection is related to the extraction of video object planes (VOPs). Some first works for VOP extraction include [3, 5, 49]. In [49], Doulamis et al. proposed to extract foreground VOPs such as head and shoulder of speakers in video conference applications. Gu et al. [5] developed a semi-automatic system where the precise object segmentation was done by human assistance in I frames followed by automatic object tracking in remaining frames. Kim et al. [3] combined temporal and spatial information to extract VOPs, which adopted temporal information to localize moving objects and spatial information to obtain precise boundaries.

Appearance rarity and objectness are two critical concepts for attention modeling. Inspired by this insight, this paper proposes an object-oriented approach for saliency detection by coupling appearance rarity and objectness into a probabilistic framework using image sparse coding representations. It consists of three major components as shown in Fig. 1. Firstly, images are characterized by a set of sparse codes learned using ICA. Secondly, the rarity probability is modeled by a GMM and the salient objectness likelihood is inferred by measuring GMM components using compactness, continuity, and center bias. These two aspects are integrated to yield the saliency map. Finally, bounding boxes locating salient objects are obtained using an adaptive algorithm.

The novelties that distinguish the proposed work from previous approaches are five-fold. 1) The proposed work integrates appearance rarity with objectness likelihood in a probabilistic paradigm based on sparse coding representations. In contrast to previous work that considers contrast alone, the combination of objectness attributes enables us to extract whole salient objects uniformly. 2) The sparse codes are learned from a large number of eye-fixation patches obtained from an eye tracking dataset rather than random patches, which has been demonstrated to achieve better results. 3) Three generic measurements are developed to characterize the objectness. These measurements are calculated efficiently and effective for saliency detection. 4) It improves on work [30] and proposes an adaptive algorithm to create bounding boxes locating salient objects easily and effectively. 5) Extensive evaluations on publicly available datasets and comparisons with 18 state-of-the-art algorithms are carried out and results demonstrate the effectiveness of the proposed work.

The rest of the paper is organized as follows. Section II describes image sparse coding representations. Section III reports the probabilistic framework involving rarity and objectness measurement. Section IV introduces an algorithm

that can locate salient objects with bounding boxes based on saliency maps. Section V presents experimental results. Finally, conclusions are drawn in Section VI.

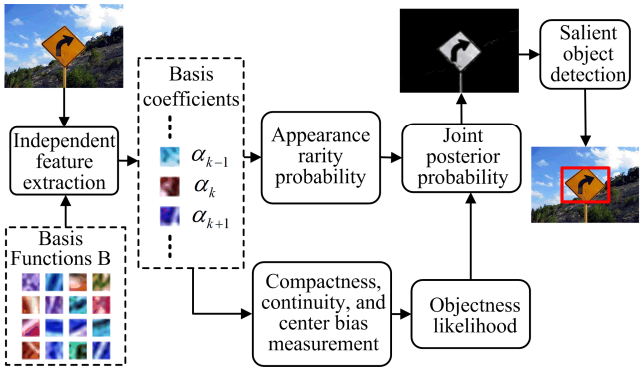


Fig. 1 The architecture of the proposed saliency detection framework based on sparse coding representations.

II. IMAGE SPARSE CODING REPRESENTATIONS

It has been commonly acknowledged that the receptive fields of simple cells in the primate primary visual cortex (V1) are spatially localized, oriented, and band-pass [35]. This intrinsic property can be accounted for by sparse coding representations, which attempt to represent a high-dimensional original signal by using a few representative atoms on a low-dimensional manifold. The investigations in [24] and [25] have found that the sparse coding principle is useful in understanding the cause of saliency mechanisms in the brain. These findings motivate us to detect visual saliency based on sparse coding representations. Furthermore, the study of V4 and MT cortical regions by [36, 37] has demonstrated that attention can be deduced from particular features. These particular features can be inferred from an eye tracking benchmark database. In sparse coding representations, each atom or code is most effective for describing one type structure or a particular feature in the image. The generation of sparse codes is partly sensitive to the training samples, of which, in practice, we have only a limited number of. This naturally motivates us to learn sparse codes using eye fixation patches from eye tracking databases instead of random patches, which is biased in favor of finding image structures or features that are more likely to draw attention.

Given an image patch $\mathbf{I}_i(x,y)$ centered at location $\mathbf{z}_i = (x_i, y_i)$, it can be represented as a linear superposition of a set of sparse coding bases:

$$\mathbf{I}_i = \sum_{j=1}^{n} f_i^j \mathbf{B}_j \quad (1)$$

Here $\mathbf{B}_j$ indicates the j th basis function and f_i^j denotes its associated coefficient, which is referred to as the "feature". Suppose the j th filter function $\mathbf{E}_j$ is the inverse/pseudoinverse of $\mathbf{B}_j$ and f_i^j is derived by:

$$f_i^j = \sum_{(x,y)} \mathbf{E}_j(x,y)\mathbf{I}_i(x,y) \quad (2)$$

Finding a complete set of basis functions which spans the image space is a critical issue. ICA training is a good way to approximately resolve this issue and thus is adopted by many existing algorithms [24-28, 44, 38, 39]. As shown in [39], although this scheme is incapable of achieving entirely independent codes, the yielded codes are independent to third-order statistics. Moreover, the investigations in [26, 35, 38] have demonstrated that the features obtained in this way qualitatively resemble those observed found in the visual cortex. Accordingly, this paper also applies ICA to learn the set of basis functions.

To implement ICA training for sparse codes, many earlier methods [24-28, 44, 38, 39] work on a collection of general-purpose image patches randomly selected from a large-scale database [24-28, 44, 38, 39]. In this paper, we utilized an eye-tracking database[1] developed by MIT AI lab [19] to learn sparse codes. It consists of eye-tracking data from 15 different viewers across 1,003 images randomly selected from Flickr and LabelMe. In this dataset, fixation locations were generated by using an eye tracker to record viewers' gaze path as they watch images. The eye-tracking data indicates where viewers actually look in images. Learning sparse codes specifically on these eye fixation patches can facilitate us to discover which subset of features is more attractive to humans. This can certainly benefit the inference of visual saliency detection task. In our implementation, we obtained a large-scale collection of eye-fixation patches from this eye-tracking dataset, where each is of size of 7×7 and centered at a fixation location. The ICA algorithm is utilized to learn 147 (7×7×3) basis functions based on these selected patches. Finally, given the image patch $\mathbf{I}_i(x,y)$, 147 coefficients calculated according to Eq. (2) are used as features $\mathbf{F}_i = \{f_i^j\}, j = 1,...,147$ to detect visual saliency.

Fig. 2 shows the 147 basis functions learned from eye-fixation patches. As reported in literatures [24, 25, 35, 48], some of basis functions resemble Gabor filters at various positions, orientations, spatial frequencies and phases, and some others look like low-pass filters that present two opposite colors. The work [35] has provided a quantitative estimation of the distribution of basis functions in space, orientation, and scale. Essentially, each basis function represents a type of structural primitive, which might be devoted to reconstructing geometrical structures in images. As stated in [35, 48], features yielded via these basis functions resemble simple-cell receptive fields. They intuitively contain much richer information than typical pixel color. With this set of sophisticated features, we may discover more types of contrast rather than only intensity or orientation contrast used by traditional saliency models. Moreover, our basis functions are learned from eye fixation patches, which reflect specific image structures or features that are more likely to draw human attention. Accordingly, we presume the use of sparse coding based features enables us to

[1] http://people.csail.mit.edu/tjudd/WherePeopleLook/interactiveWebsite/seeFixations.html

improve the quality of saliency detection.

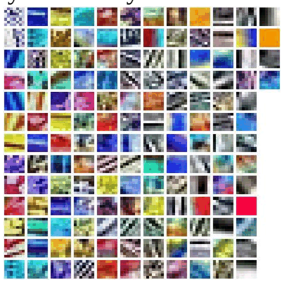


Fig. 2 Basis functions learned from eye-fixation patches.

III. OBJECT-ORIENTED SALIENCY MAP

Previous approaches [24-28, 44] using the sparse coding principle mainly concentrated on modeling contrast information of individual pixels or small areas. However, they ignored the information that a pixel belongs to the object, which leads to difficulties in uniformly finding whole salient objects with well-defined boundaries. To tackle this weakness, this paper develops a probabilistic framework to compute the saliency map by taking both pixel rarity and objectness into consideration simultaneously. The rarity is characterized by the global contrast. Three attributes reflecting salient objectness of a pixel called compactness, continuity, and center bias are measured using contextual information.

The inference of the saliency of a pixel is formulated as follows. Let $\mathbf{z}_i = (x_i, y_i)$ denote a pixel and $\mathbf{F}_i = \{f_i^j\}, j = 1,...,147$ denote its corresponding feature vector based on sparse coding representations. We assume the binary random variable r_i indicates whether the pixel stands out from its surroundings or not, and assume the binary random variable o_i denotes whether the pixel belongs to an object or not. They are formalized as:

$$r_i = \begin{cases} 1 & \text{if } \mathbf{z}_i \text{ is distinctive,} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$o_i = \begin{cases} 1 & \text{if } \mathbf{z}_i \text{ belongs to an object,} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Instead of taking only pixel rarity into account as in the earlier works [26, 15], the proposed approach intends to integrate rarity with objectness in a probabilistic framework. The saliency value S_i of $\mathbf{z}_i$ is defined as a joint posterior probability as below:

$$S_i = p(o_i = 1, r_i = 1 | \mathbf{F}_i) \tag{5}$$

It is reasonable to assume r_i and o_i are conditionally independent given $\mathbf{F}_i$. Therefore, Eq. (5) can be rewritten as:

$$\begin{aligned} S_i &= p(o_i = 1, r_i = 1 | \mathbf{F}_i) \\ &= p(o_i = 1 | \mathbf{F}_i) p(r_i = 1 | \mathbf{F}_i) \end{aligned} \tag{6}$$

where $p(o_i = 1 | \mathbf{F}_i)$ is called the "objectness probability" and $p(r_i = 1 | \mathbf{F}_i)$ is the "rarity probability".

A. *Rarity probability*

Intuitively, the rarity probability reflects how much a location is distinctive from background. In this paper, we adopt global contrast to measure the rarity. Similar to [26], according to the Bayesian rule, $p(r_i = 1 | \mathbf{F}_i)$ can be calculated by:

$$p(r_i = 1 | \mathbf{F}_i) = \underbrace{\frac{1}{p(\mathbf{F}_i)}}_{\text{bottom-up saliency}} \underbrace{p(\mathbf{F}_i | r_i = 1)}_{\text{top-down knowledge}} \underbrace{p(r_i = 1)}_{\text{Prior}} \tag{7}$$

$$\log p(r_i = 1 | \mathbf{F}_i) = -\log p(\mathbf{F}_i) + \log(\mathbf{F}_i | r_i = 1) + \text{const} \tag{8}$$

There are three terms on the right side of Eq. (7). The first item measures the bottom-up saliency. The second item corresponds to the top-down knowledge. The last one is the prior. Under the scenario of free viewing, only the first item needs to be considered, which means:

$$\log p(r_i = 1 | \mathbf{F}_i) \propto -\log p(\mathbf{F}_i) \tag{9}$$

In the proposed algorithm, $\mathbf{F}_i = \{f_i^j\}, j = 1,...,147$ is a random variable vector consisting of all filter responses. $p(\mathbf{F}_i)$ is the joint probability of filter responses. Since filters learned using ICA are approximately independent, the joint probability is simplified to the product of probability of each filter response:

$$\begin{aligned} p(\mathbf{F}_i) &= \prod_{j=1}^{147} p(f_i^j) \\ \log p(\mathbf{F}_i) &= \sum_{j=1}^{147} \log p(f_i^j) \end{aligned} \tag{10}$$

We utilize GMMs with M components to estimate the distribution of each filter response according to:

$$p(f_i^j) = \sum_{c=1}^{M} \pi_c N(f_i^j | \mu_c, \sigma_c^2) \tag{11}$$

where parameters π_c, μ_c, σ_c can be inferred using expectation-maximization (EM) algorithm. Alternatively, the generalized Gaussian distribution used in [26] also can be applied to estimate $p(f_i^j)$.

B. *Objectness probability*

Instead of training object detectors for specific classes, for example, faces, cars, or buildings, this paper develops a set of measurements of objectness generic over classes. It indicates how likely it is for a pixel to belong to an object. In contrast to object detectors extensively trained from a large number of samples, our measurements are relatively "weak" and easy to obtain, but they are effective to salient object detection. In this paper, the objectness of a salient object is characterized using three measurements: compactness, continuity, and center bias. The first two measurements are inherent properties of an object. The third measurement models a high-level attribute for a salient object, which accounts for the fact that an object closer to the centre of the image is more likely to attract interest.

Intrinsically, the objectness is a property of a group of pixels. It is meaningless to estimate the objectness using every individual pixel alone. As shown in Eq. (11), we utilize GMMs $\{\pi_c, \mu_c, \sigma_c^2\}_{c=1}^{M}$ to model each filter's responses. The components of the GMMs are regarded as the basic units to calculate the objectness. The objectness of a pixel is predicted by a probabilistic combination of various components.

1) Compactness

The compactness describes the global distribution of an object. Following Eq. (11), given the j th filter response map, every pixel $\mathbf{z}_i = (x_i, y_i)$ with the corresponding feature f_i^j is assigned to a component with the probability:

$$p(c \mid f_i^j) = \frac{\pi_c N(f_i^j \mid \mu_c, \sigma_c^2)}{\sum_{c=1}^{M} \pi_c N(f_i^j \mid \mu_c, \sigma_c^2)} \tag{12}$$

Inspired by [40], this paper proposes to measure the compactness of a component as follows. At first, the location variance of a component c is calculated by:

$$\bar{x} = \frac{\sum_i p(c \mid f_i^j) \cdot x_i}{\sum_i p(c \mid f_i^j)}, \quad \bar{y} = \frac{\sum_i p(c \mid f_i^j) \cdot y_i}{\sum_i p(c \mid f_i^j)} \tag{13}$$

$$V_T = \sum_i \left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 \right] \tag{14}$$

Afterwards, values of $p(c \mid f_i^j)$ are quantized into K non-overlapping ranges equally. Pixels are assigned to these K labels to form various class-maps based on their corresponding $p(c \mid f_i^j)$. As mentioned in [40], the class-map can be viewed as a sort of texture composition. The total variance of pixels belonging to the same class is computed as:

$$\bar{x}_m = \frac{\sum_{(x_i, y_i) \in CM_m} p(c \mid f_i^j) \cdot x_i}{\sum_{(x_i, y_i) \in CM_m} p(c \mid f_i^j)}, \quad \bar{y}_m = \frac{\sum_{(x_i, y_i) \in CM_m} p(c \mid f_i^j) \cdot y_i}{\sum_{(x_i, y_i) \in CM_m} p(c \mid f_i^j)}, \ m = \{1, 2, ..., K\} \tag{15}$$

$$V_W = \sum_{m=1}^{K} \bar{p}_m \Big(\sum_{(x_i, y_i) \in CM_m} \left[(x_i - \bar{x}_m)^2 + (y_i - \bar{y}_m)^2 \right] \Big) \tag{16}$$

Here, CM denotes the class-map and $\bar{p}_m$ is the mean of $p(c \mid f_i^j)$ for pixels belonging to each class-map.

The compactness of the component c in the j th filter response map is finally defined as:

$$CP_c^j = \frac{V_W}{V_T - V_W} \tag{17}$$

The motivation of the compactness calculation is originally from the Fisher's multi-class discriminant [40]. Its value is large when all pixels of various classes uniformly distributed over the entire image. Otherwise, its value tends to be small.

2) Continuity

Objects normally appear to be continuous individuals over space. Spatial continuity is a powerful determinant of object persistence. Boundary information is a visual feature that can indicate the object continuity. Accordingly, this paper measures continuity based on gradients. Given a component c in the j th filter response map, its continuity is calculated by:

$$CT_c^j = \sum_i [p(c \mid f_i^j) \sqrt{(\frac{\partial f_i^j}{\partial x_i})^2 + (\frac{\partial f_i^j}{\partial y_i})^2}] \tag{18}$$

As shown in Fig. 3, CT tends to be small when pixels belonging to its corresponding component are spatially continuous.

3) Center bias

Generally speaking, objects closer to the center are more likely to attract human attention. Center bias is an effective factor to detect visual saliency. We compute the center bias of a component c in the j th filter response map as follows:

$$CB_c^j = \sum_{(x_i, y_i)} [p(c \mid f_i^j) \bullet \sqrt{(x_i - \hat{x})^2 + (y_i - \hat{y})^2}] \Big/ \sum_{(x_i, y_i)} p(c \mid f_i^j) \tag{19}$$

Fig. 3 displays an example where six components and their associated three objectness measures are indicated.

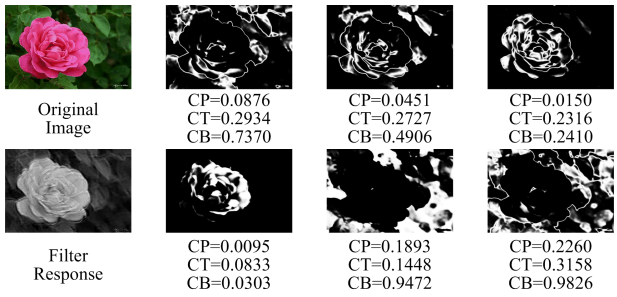


Fig. 3 An example showing six GMM components and the associated three objectness measurements. In the first column, the top image is the original image and the bottom image is the corresponding response to a filter. In other columns, six images show various objects formed by various components with their compactness (CP), continuity (CT), and center bias (CB).

After the above three objectness measurements are obtained, the objectness probability of a pixel can be derived by the probabilistic combination of objectness measurement of every component in every filter response map as follows:

$$p(o_i = 1 \mid \mathbf{F}_i) = \frac{1}{Z_1} \sum_{j=1}^{147} p(o_i = 1 \mid f_i^j) \tag{20}$$

Here, Z_1 is a normalizer. As mentioned, our approach tends to use the objectness of components to indicate the objectness of a pixel. Following this idea, $p(o_i = 1 \mid f_i^j)$ can be estimated as:

$$p(o_i = 1 \mid f_i^j) = \sum_{c=1}^{M} p(o_i = 1 \mid c) p(c \mid f_i^j) \tag{21}$$

$p(o_i = 1 \mid c)$ represents the likelihood of a component forming an object. According to those three objectness measurements, it can be formalized as an exponential distribution:

$$p(o_i = 1 \mid c) = \frac{1}{Z_2} \exp\left(-\frac{\overline{CP_c^j}^2 + \overline{CT_c^j}^2 + \overline{CB_c^j}^2}{\lambda^2}\right) \tag{22}$$

$$\overline{CP_c^j} = \frac{CP_c^j}{\sum_{j=1}^{147} \sum_{c=1}^{M} CP_c^j}, \quad \overline{CT_c^j} = \frac{CT_c^j}{\sum_{j=1}^{147} \sum_{c=1}^{M} CT_c^j}, \quad \overline{CB_c^j} = \frac{CB_c^j}{\sum_{j=1}^{147} \sum_{c=1}^{M} CB_c^j} \tag{23}$$

Here, Z_2 is a normalizer and the parameter λ can be regarded as a scale controller. It controls the shape of exponential functions and thus implies the importance of objectness measurements to the overall saliency detection. Fig. 4 displays the impact of λ on saliency maps. As can be seen from the figure, the results obtained by using a very small value of λ generally distribute quite compactly, whereas they cannot form a whole object. In contrast, results obtained by using too large a value of λ may contain redundant points from background.

This paper empirically determines an optimal value for λ, which will be described in the later experiment section.

The overall objectness probability is computed as follows:

$$p(o_i = 1 \mid \mathbf{F}_i) = \frac{1}{Z_1 Z_2} \sum_{j=1}^{147} \sum_{c=1}^{M} [\exp(-\frac{\overline{CP_c^j}^2 + \overline{CT_c^j}^2 + \overline{CB_c^j}^2}{\lambda^2}) \cdot \frac{\pi_c N(f_i^j \mid \mu_c, \sigma_c^2)}{\sum_{c=1}^{M} \pi_c N(f_i^j \mid \mu_c, \sigma_c^2)}] \quad (24)$$

Substantively, the calculation of the objectness probability comprises the selection procedure of filter response map and component, which leads to the discovery of a subset of appropriate feature spaces to compose salient objects.

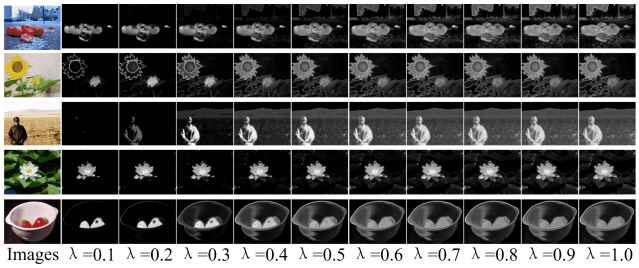


Fig. 4 Examples illustrating the impact of λ on saliency maps. The first column shows original images and the second to the eleventh column show the corresponding saliency maps generated by setting λ to values from 0.1 to 1.

IV. SALIENT OBJECT DETECTION

The last step in the saliency detection framework is to detect salient objects based on the saliency map. Although a few efforts [2, 12, 23] have attempted to segment accurate object boundaries, they are not robust to complicated images with cluttered background. Most other approaches simply found bounding boxes that can cover most of salient points based on further analyzing the saliency map. This detection strategy is also utilized by the proposed framework.

Luo et al. [30] presented an efficient algorithm that predicts bounding boxes with maximum saliency density (MSD). It formulated the problem as follows. Given an image **IM** and the corresponding saliency map S, the objective is to find a sub-image $\mathbf{W}$ to locate area of maximum saliency density where $\mathbf{W} \subseteq \mathbf{IM}$. This can be mathematically formalized as:

$$\mathbf{W}^* = \arg\max_{\mathbf{W} \subseteq \mathbf{I}} h(\mathbf{W}) \quad (25)$$

$$h(\mathbf{W}) = \frac{\sum_{(x,y) \in \mathbf{W}} S(x,y)}{\sum_{(x,y) \in \mathbf{I}} S(x,y)} + \frac{\sum_{(x,y) \in \mathbf{W}} S(x,y)}{D + Area(\mathbf{W})} \quad (26)$$

where $\mathbf{W}^*$ is the optimal sub-window and D is a positive constant to balance the area of $\mathbf{W}$. In spite of good performance reported in [30], it has a drawback that the value the free parameter D has to be determined empirically, which consists of a tedious procedure of parameter tuning. This may reduce the generality of the algorithm. Alternatively, this paper proposes an elegant algorithm which removes this parameter while achieving comparable performance. We tend to solve the problem by presenting an alternative objective function below:

$$h(\mathbf{W}) = \frac{\sum_{(x,y) \in \mathbf{W}} S(x,y)}{\sum_{(x,y) \in \mathbf{I}} S(x,y)} - \frac{\sum_{(x,y) \in \mathbf{W}} \overline{\overline{S(x,y)}}}{\sum_{(x,y) \in \mathbf{I}} \overline{\overline{S(x,y)}}} \quad (27)$$

Here, $\overline{\overline{S(x,y)}} = \max(S(x,y)) - S(x,y)$, which represents the impact of background pixels. The first term in $h(\mathbf{W})$ ensures that $\mathbf{W}$ contains more salient points in a similar manner to [30]. The second term ensures that $\mathbf{W}$ contains fewer background pixels. The maximization of these two terms simultaneously can achieve good performance. Afterwards, the optimization of the objective function follows the branch-and-bound search method described by [45, 30]. The basic idea of the optimization [45, 30] is to hierarchically split the set of all possible rectangles into disjoint subsets. An upper bound is calculated based on the objective function for each candidate rectangle set. The next search over candidate rectangle sets works in a best-first manner, which preferentially examines the most promising candidate in terms of its upper bound. The search is terminated when the most promising candidate contains only a single rectangle, which guarantees that a global maximum can be achieved. The branch-and-bound search avoids the extensive search in a large number of rectangle candidates whose upper bounds tell that they are not promising. Therefore, comparing with the exhaustive search, it can find the optimal solution with the less cost.

V. EXPERIMENTAL RESULTS

We construct experiments to demonstrate the performance of the proposed framework, which mainly includes 1) evaluation of the proposed saliency map and comparison with state-of-the-art algorithms; 2) evaluation of the scheme of sparse code learning from eye-fixation patches; 3) evaluation of the proposed salient object detection algorithm.

A. *Experimental settings*

In this paper, two publicly available benchmark datasets called MSRA dataset [18, 23] and Bruce dataset [24] are used for evaluations. The first dataset consists of 1,000 images with manually labeled ground truth [18, 23]. To our best knowledge, this dataset may be one of the largest test sets for saliency detection whose ground truth is in the form of manually labeled accurate object-contours instead of rough bounding boxes as in [18]. The benchmark dataset has been widely utilized by a variety of up-to-date saliency detection approaches to test their performance such as [18], [23], [17], [20], [29], and [26]. Details of this benchmark dataset can be found in [23] and [18]. The second dataset is an eye-fixation dataset provided by [24]. It consists of 120 images with ground truth generated by eye tracking data from 20 different subjects.

Following [13], [15], [17], [19], [21], [23-24], [26-27], [29], and [31-32], Receive Operator Characteristic (ROC) curves and the areas under ROC (AUC) are used as the metrics to quantitatively measure the performance. ROC and AUC are generated by classifying the pixels in a saliency map into salience or non-salience by varying the quantization threshold

within the range [0, 255]. The resulting false positive rate versus hit rate at each threshold value forms the ROC curve.

B. *Evaluation of the saliency map*

1) **Parameters and components analysis of the proposed model**

In this section, we analyze the effect of parameters and components of the proposed model. The evaluations were performed on the MSRA dataset. In our model, the scale parameter λ is a free parameter. The estimation of λ in principle is a non-trivial problem. This paper estimated it empirically. We generated the saliency map using the proposed approaches by varying λ between 0.1 and 1.0. Fig. 5 illustrates AUCs associated with different values of λ.

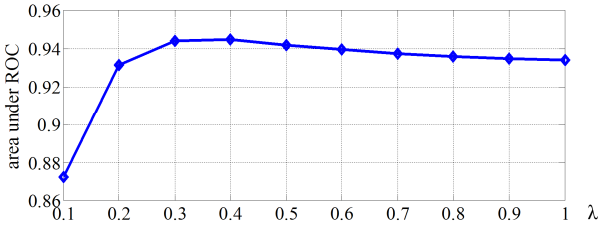


Fig. 5 The AUC with different λ values.

As can be seen, the proposed algorithm is reasonably sensitive to λ and setting λ between 0.2 and 1.0 can yield generally similar accuracy. In our implementation, λ was fixed at 0.3. Another free parameter in our model is the number M of components in the GMM model. In the current implementation, M was set to 6 empirically and we additionally found setting M to the value between 3 and 5 did not degrade the performance significantly in the experiments. It is worth mentioning that the proposed algorithm works well on all 1,000 test images using a fixed set of parameter values and without any parameter tuning on individual images, which indicates the robustness of the algorithm.

In our model, rarity probability and objectness probability are two major factors. Objectness probability further relies on three components: compactness, continuity, and center bias. To test the effect of each component, we quantitatively calculated the saliency detection performance by using rarity probability only and using rarity probability combined with each individual objectness component, respectively. Fig. 6 illustrates the AUCs associated with each combination, where "R" indicates the use of rarity probability only, and "R+CP", "R+CT", and "R+CB" indicate the combination of rarity probability and compactness, continuity, and center bias, respectively, and "R+All" indicates the combination of rarity probability and all three components. It is easy to observe that the integration of objectness measurement is certain to benefit saliency detection significantly, which obtains the improvement of 0.085 (8.5%) in terms of AUC. The components of compactness and center bias basically contribute to our model equally. The component of continuity contributes less than other two components. The integration of rarity probability and objectness probability with all three components achieves the best performance.

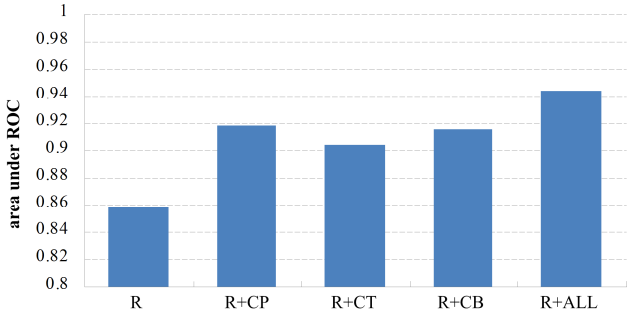


Fig. 6 Quantitative evaluation of effect of each component to the proposed model in terms of AUCs.

2) Comparisons with state-of-the-art approaches on benchmark databases

To demonstrate the effectiveness of the proposed algorithm in yielding the saliency map, we compared it with 18 state-of-the-art approaches. These approaches are selected for comparison mainly because 1) they were published in recent a few years; 2) they were published in major computer vision/machine learning conferences or journals, for example CVPR, ICCV, NIPS, and IEEE PAMI; 3) their source codes or executable codes and parameter settings were provided by the authors themselves. The selected 18 state-of-the-art approaches are AWS [41], FSDA [33], FT [23], GBVS [31], HC [47], ICL [25], IM [24], IS [34], ITTI [10], MSS [42], PWHL [19], RC [47], SDSR [15], SER [27], SIM [32], SR [21], SRDS [12], and SUN [26]. Notice that CA [29] is also a good saliency detection approach. However, it mainly aims to extract salient locations and meaningful context. Most other methods including ours are to detect salient locations only. In addition, there lacks an appropriate database to fairly compare these two different types of approaches so far. Therefore, we did not compare our method with CA in this paper.

Generally, the quality of saliency computation relies on image content. In [29], Goferman et al. categorized image content into three cases: single salient object over uninteresting background, salient object over salient context, and images of complex scenes. Since this paper does not aim for extracting context of salient objects, we basically consider two cases: one obvious salient object over "clean" backgrounds and images of complex scenes, which have multiple salient objects with small size, complex appearance, and complex background. Images of the MSRA dataset basically correspond to the first case and images of the Bruce dataset correspond to the second case. Therefore, we evaluate the saliency detection algorithms on these two datasets, respectively. Fig. 7 displays a number of results generated by the proposed method and other state-of-the-art algorithms. From the left to the right, the first six examples were from the MSRA dataset and the rest four of examples were from the Bruce dataset. The subjective evaluations by comparing with the ground truth suggest that the proposed method can yield saliency maps correctly and robustly in both cases. Our saliency detector generally can produce saliency maps with full resolution and be used to segment salient objects with well-defined boundary.

The first quantitative comparison was performed on the MSRA dataset. Fig. 8 shows the ROC curves and Fig. 9 lists AUCs of various approaches. Cheng et al. [47] proposed two excellent approaches called HC and RC. Especially, the RC algorithm incorporates image segmentation techniques into the contrast measurement, which can improve the performance. As can be seen from comparison results, our method is slightly worse than RC while outperforming other 17 algorithms in terms of AUC. The AUC difference between our method and the RC is about 0.018 (1.8%). One possible interpretation is that the image segmentation used in RC works remarkably on the MSRA data that contains a simple salient object and clean backgrounds. This point was also mentioned by [46]. It is interesting to observe that the ROC curve of the proposed work intersects with the curve of GBVS. Comparing with GBVS, our method shows higher accuracy for low false rates (<0.25). This is because our detected salient pixels fall well in true salient regions, have near uniform values, and form accurate boundary, but sometimes do not cover the entire object. In contrast, although the detected salient pixels of GBVS are not very accurate and do not have uniform values within salient regions, they can cover the entire object. In most cases, salient regions detected by GBVS are larger than the true objects. Therefore, it detects more true salient pixels when false rates become higher. However, in terms of AUC that quantifies the average quality of saliency maps, our method is better than GBVS.

Another quantitative comparison was performed on the Bruce dataset. Fig. 10 shows the ROC curves and Fig. 11 lists AUCs of various approaches. It can be seen that the proposed approach is better than 18 existing approaches. Especially, our method outperforms RC [47] by 0.09 (9%) in terms of AUC. RC performs much worse on the Bruce database compared with the MSRA database. As pointed out in [46], the major explanation is that image content of the Bruce database basically is much more complex and image segmentation algorithm may fail for this case, which results in the significant decrease of its performance. However, since our method adopts features sparsely coded using eye fixation data which embody rich information and generic objectness rules that do not rely on image segmentation, it achieves good performance in both databases. In summary, our evaluations and comparisons on two benchmark databases have demonstrated that the proposed model works effectively on images with simple content and images with relatively complex content.

The average time costs taken by various algorithms are listed on Table 1. It was estimated based on computing saliency maps of 100 randomly selected images from the MSRA database with the resolution of 400×300. All algorithms were tested on a 24-core Lenovo Server with Intel Xeon CPU of 2.8 GHz. As can be seen, our algorithm has moderate computational cost.

3) Comparisons with state-of-the-art approaches on categorized images

In the last subsection, we categorized image content in terms of the complexity of salient objects and images. It is also interesting to define image content in terms of semantic category and measure the performance of various approaches in each image category. To setup the experiment environment, we asked three participants to manually categorize 1,000 images of the MSRA dataset into 12 semantic classes according to image content, which are traffic sign, car, animal, fruit, flower, egg, building, human, dessert, leaf, toy, and others. Every category contains several tens of images.

Fig. 12 displays a few samples of each category. Table 2 lists the AUCs achieved by various approaches in different image categories. As can be seen, some approaches such as RC and our method can generally obtain consistent good performance in all categories, whereas the performance of some other methods for example, IM, SER, and SIM, appears to be inconsistent across categories.

C. *Evaluation of sparse code learning scheme*

This experiment evaluates the performance of the scheme for learning sparse codes from eye-fixation patches by comparing it against the commonly used scheme for learning sparse codes from random image patches. The fixation-based training set consists of a large number of selected eye-fixation patches from 1,003 images in the eye tracking database [19]. In addition, more than 200,000 random patches from the same database were collected to form another training set. These two training sets were applied to learn sparse codes respectively, and comparison results are shown in Fig. 13 (subjective evaluations) and Fig. 14 (quantitative evaluations). The results demonstrate that the proposed learning scheme is effective.

As explained in section II, sparse codes learned from eye-tracking data are expected to be able to find particular image features or attributes that are more likely to attract the viewers' attention. This essentially account for the observation that our proposed approach can achieve a better saliency map compared with the approach of learning sparse codes using random patches. An experiment was constructed to quantitatively demonstrate this point. Since our sparse codes were trained using the MIT eye tracking benchmark database [19], we utilized the MSRA benchmark database [18, 23] as the test data for the purpose of cross-validation. At first, all salient pixels were collected from the saliency ground truth of 1,000 images [23]. Afterwards, for each salient pixel, we yielded a quantized histogram to approximate the probability distribution of its responses to filters (f_i^j in Eq. (2)) derived from the learned sparse codes. Finally, the Shannon Entropy is calculated based on the histogram as:

$$H = -\sum_i p_i \log p_i \tag{28}$$

Here, p_i indicates the probability of the i th bin in the histogram. In this way, we can obtain the mean entropy of all salient pixels. For the purpose of comparison, we computed the average entropy corresponding to sparse codes learned from eye-tracking data and random patches, respectively.

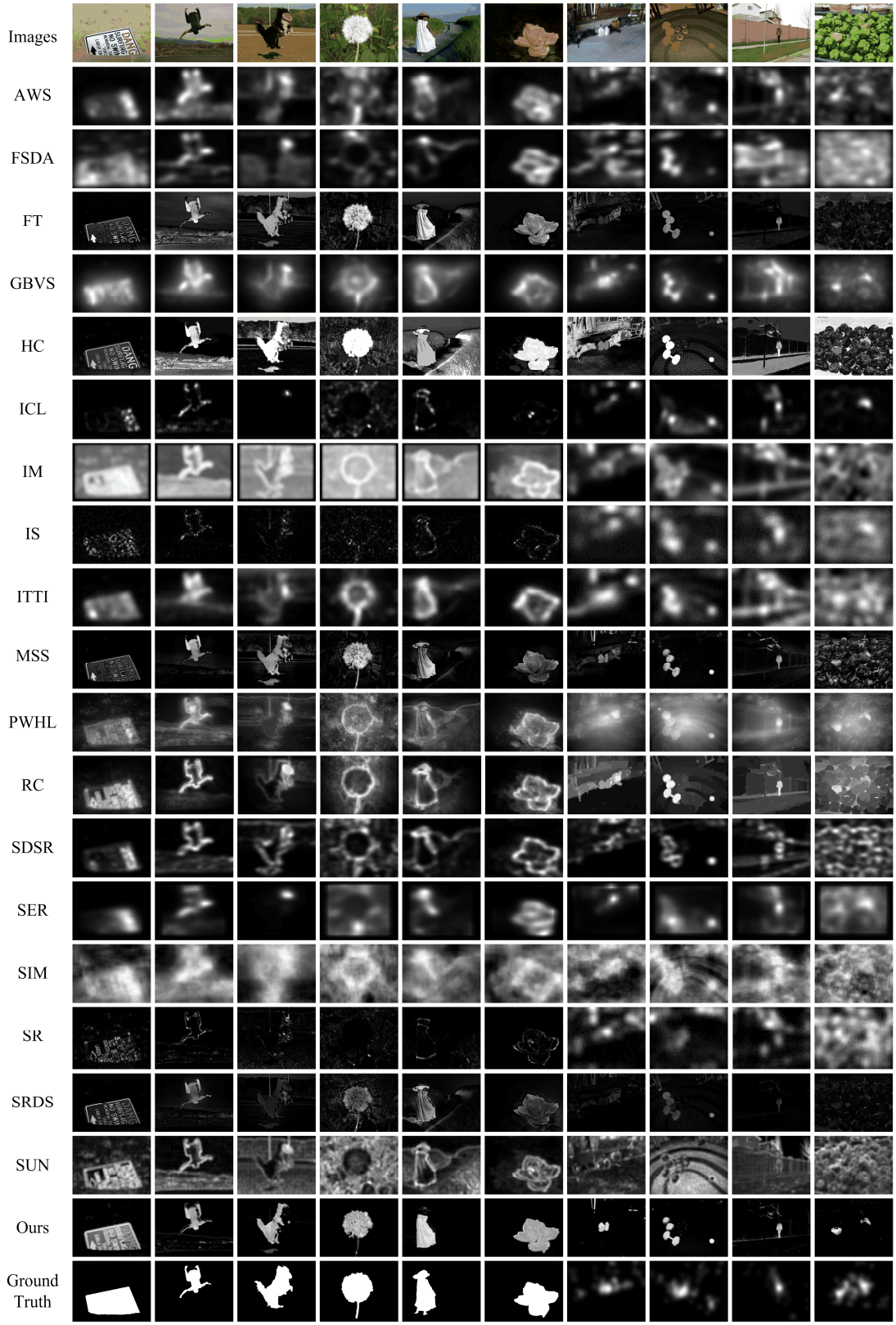


Fig. 7 A number of comparison results of 18 state-of-the-art approaches, ours, and the ground truth.

Their values are 6.78 and 6.08, respectively. A larger entropy generally indicates its uncertainty is higher. In our case, it also implies that more learned sparse codes are sensitive to particular features of attentive objects. Compared with the scheme of learning sparse codes with random patches, the proposed scheme of learning sparse codes with eye-tracking can achieve around an 11% improvement in entropy averagely, which demonstrates the sparse codes learned from eye-fixation

patches are more appropriate for accounting for particular features that are more likely to attract human attention.

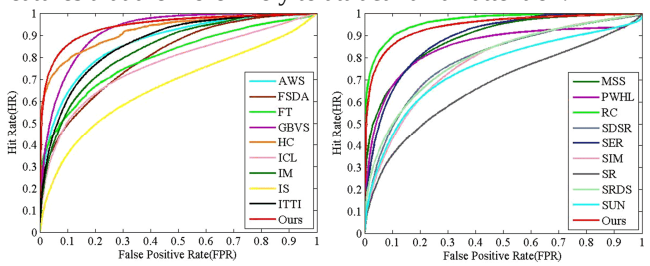


Fig. 8 Saliency map quantitative comparison of the proposed algorithm with 18 state-of-the-art approaches using ROC curves on the MSRA dataset.

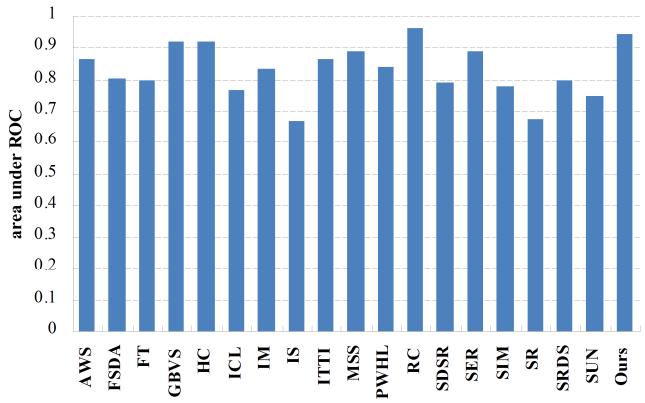


Fig. 9 Saliency map quantitative comparison of the proposed algorithm with 18 state-of-art approaches using AUC on the MSRA dataset.

D. *Evaluation of salient object detection*

In order to demonstrate the effectiveness of the adaptive approach for salient object detection proposed in section V, we compared it with [30] on saliency maps generated by using the proposed work. All 1000 images from [18] were used as the test dataset. Its ground truth, which labels detected results using bounding boxes, was provided by MSRA [18]. Similar to [30], given the rectangle-like binary mask G_a detected by the algorithm and the binary mask G_g by the ground truth, the precision, recall, and F-measure were applied to calculate the performance, which are defined as:

$$precision = \frac{\sum G_a \times G_g}{\sum G_a}, \quad recall = \frac{\sum G_a \times G_g}{\sum G_g} \tag{29}$$

$$F-measure = \frac{(1+\beta)\times precision \times recall}{\beta \times precision + recall} \tag{30}$$

Our implementation takes $\beta = 0.5$ as suggested by [30]. Our experiments followed all settings in [30] to empirically determine the value of D. Fig. 15 illustrates the relationships between D and the F-measure based on our test dataset and using the proposed saliency map. As can be seen from Fig. 15, $D = 1.1$ leads to the best performance. Accordingly, when we compared the proposed algorithm with MSD [30], we set $D = 1.1$. Fig. 16 displays some sample results of the comparison. Our method performs better than [30] on the first five examples (from the left to the right) and worse on the last three examples. We also have done quantitative comparison evaluations on our test dataset. The F-measure values of the proposed algorithm and MSD [30] are 0.84 and 0.83, respectively. As can be seen from the comparison results, the proposed algorithm can slightly improve on the performance of the MSD algorithm [30]. More importantly, the proposed algorithm can eliminate the free parameter D in MSD and remove the tedious procedure of parameter tuning.

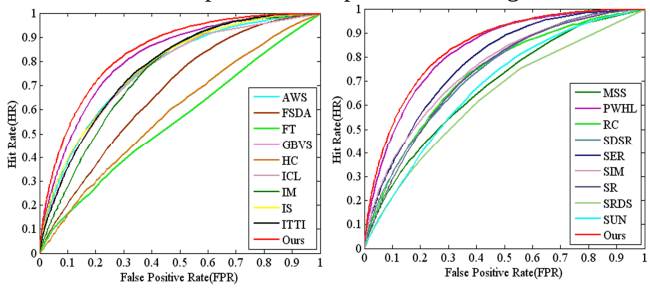


Fig. 10 Saliency map quantitative comparison of the proposed algorithm with 18 state-of-art approaches using ROC on the Bruce dataset.

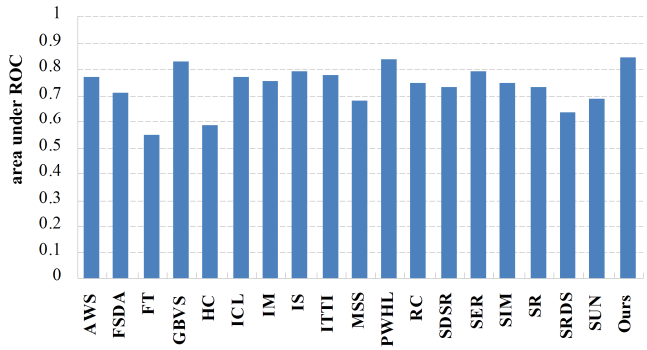


Fig. 11 Saliency map quantitative comparison of the proposed algorithm with 18 state-of-art approaches using AUC on the Bruce dataset.

VI. CONCLUSIONS

In this paper, we have reported a probabilistic framework for visual saliency detection using sparse coding representations. Two key contributions that distinguish the proposed work from most previous works are summarized as follows: 1) A probabilistic formalization is developed to integrate appearance rarity and objectness likelihood. Three generic measurements are proposed to estimate the objectness likelihood; 2) Sparse codes are learned from eye-fixation patches instead of randomly selected patches. Comprehensive evaluations and comparisons with 18 state-of-the-art methods on publicly available benchmark datasets have demonstrated the effectiveness of the proposed framework.

Fig. 12 Samples from various image categories.

TABLE 1
AVERAGE TIME TAKEN TO COMPUTE A SALIENCY MAP USING VARIOUS APPROACHES

Method	**AWS**	**FSDA**	**FT**	**GBVS**	**HC**	**ICL**	**IM**	**IS**	**ITTI**
Time (s)	4.0970	4.2653	0.011	1.4719	0.014	0.2376	8.6450	0.0071	0.2611
Code	Matlab	Matlab	C++	Matlab and C++	C++	Matlab	Matlab	Matlab	Matlab and C++
Method	**MSS**	**RC**	**SDSR**	**SER**	**SIM**	**SR**	**SRDS**	**SUN**	**Ours**
Time (s)	0.0721	0.177	2.0238	37.9458	1.3718	0.0105	0.0855	1.9329	1.2106
Code	C++	C++	Matlab	C++	Matlab	Matlab	C++	Matlab	Matlab and C++

TABLE 2
AUCs ACHIEVED BY VARIOUS APPROACHES IN DIFFERENT IMAGE CATEGORIES

	Traffic sign	Car	Animal	Fruit	Flower	Egg	Building	Human	Dessert	Leaf	Toy	Others
AWS	0.879	0.913	0.872	0.813	0.855	0.781	0.939	0.918	0.844	0.832	0.913	0.895
FSDA	0.811	0.840	0.810	0.838	0.825	0.809	0.851	0.868	0.835	0.817	0.836	0.813
FT	0.770	0.743	0.828	0.757	0.855	0.792	0.803	0.819	0.730	0.774	0.815	0.808
GBVS	0.883	0.937	0.944	0.883	0.927	0.903	0.937	0.947	0.935	0.894	0.934	0.930
HC	0.904	0.841	0.900	0.878	0.964	0.903	0.913	0.893	0.838	0.933	0.936	0.908
ICL	0.803	0.823	0.764	0.734	0.804	0.675	0.819	0.810	0.699	0.767	0.805	0.769
IM	0.860	0.914	0.879	0.802	0.832	0.734	0.924	0.895	0.833	0.794	0.900	0.856
IS	0.686	0.755	0.703	0.620	0.658	0.582	0.735	0.734	0.645	0.622	0.725	0.685
ITTI	0.876	0.905	0.906	0.795	0.844	0.826	0.916	0.915	0.822	0.813	0.908	0.881
MSS	0.857	0.864	0.893	0.868	0.908	0.882	0.878	0.898	0.854	0.858	0.915	0.892
PWHL	0.853	0.934	0.931	0.851	0.868	0.848	0.954	0.936	0.903	0.861	0.923	0.905
RC	0.946	0.956	0.962	0.958	0.974	0.958	0.953	0.962	0.926	0.961	0.969	0.965
SDSR	0.843	0.878	0.807	0.717	0.740	0.607	0.853	0.879	0.743	0.721	0.826	0.801
SER	0.923	0.937	0.865	0.821	0.908	0.799	0.911	0.921	0.808	0.864	0.919	0.886
SIM	0.820	0.881	0.840	0.686	0.732	0.613	0.827	0.893	0.634	0.724	0.810	0.790
SR	0.631	0.796	0.712	0.642	0.737	0.546	0.774	0.687	0.662	0.646	0.750	0.678
SRDS	0.835	0.838	0.803	0.742	0.763	0.699	0.818	0.860	0.710	0.762	0.824	0.795
SUN	0.724	0.838	0.715	0.730	0.800	0.589	0.798	0.762	0.661	0.697	0.794	0.718
Ours	0.912	0.922	0.935	0.939	0.977	0.968	0.941	0.918	0.943	0.918	0.948	0.938

In our current algorithm, inferring GMMs for each of 147 filter responses took most of computation time. One extension to this work is to selectively build GMMs for useful filter responses when generating the saliency map. This should reduce computational complexity. Another potential future work is to incorporate reliable semantic-based object detectors into the proposed work, which is presumed to improve the performance. In addition, we intend to apply the outcome of our work, which automatically predicts locations of interest of human perception to improve the performance of many current challenging problems like image object segmentation, image object retrieval and browsing, and image object categorization.

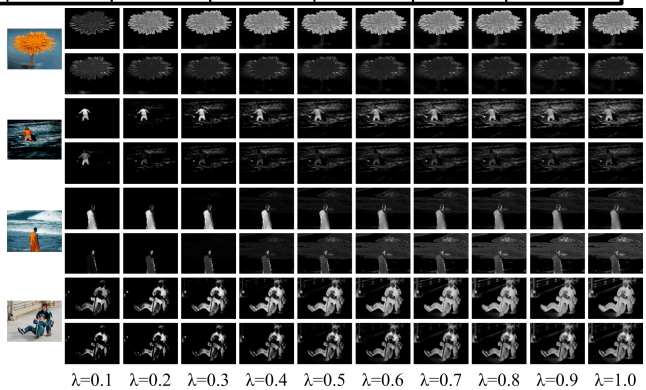


Fig. 13 A number of examples obtained by two different schemes for learning sparse codes. The top row shows the results based on sparse codes learned from eye-fixation patches with varying λ from 0.1 to 1, while the bottom row is the results based on sparse codes learned from random patches.

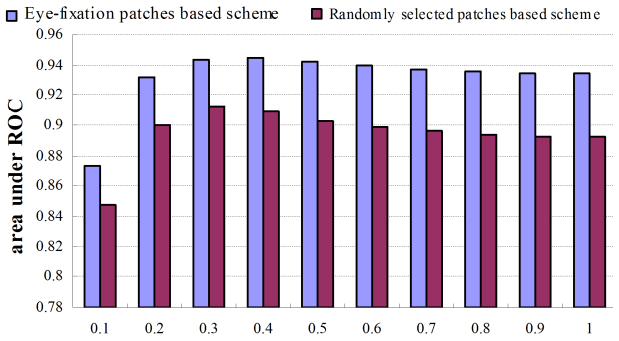


Fig. 14 The AUC comparisons of two different schemes of learning sparse codes by varying λ from 0.1 to 1.

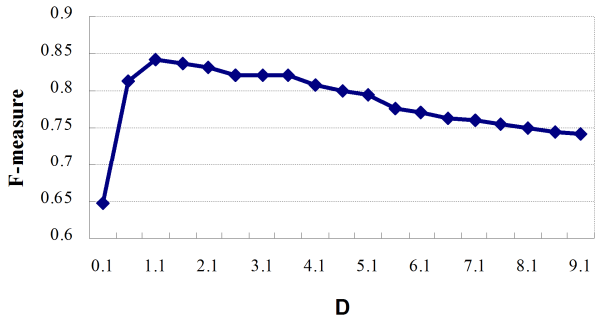


Fig. 15 Evaluation D value on our test dataset. The x-coordinate is D value measured in unit $10^4 \times ImageSize$ and the y-coordinate is F-measure.

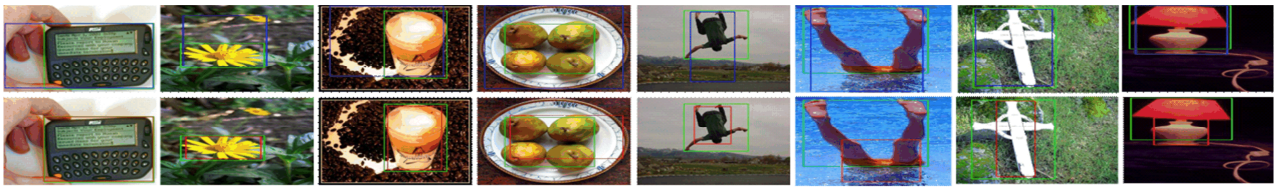


Fig. 16 Some comparison samples. In each result, the red rectangle is detected results by our method, the blue one is detected results by using MSD [30], and the green rectangle is the ground truth.

REFERENCES

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiol*, vol. 4, pp. 219-227, 1985.

[2] J. Han, K. Ngan, M. Li, and H. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 16, no.1, pp. 141-145, Jan. 2006.

[3] M. Kim, J. Choi, D. Kim, H. Lee, M. Lee, C. Ahn, Y. Ho, "A VOP generation tool: automatic segmentation of moving objects in image sequences based on spatio-temporal information," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 9, no. 8, pp. 1216-1226, Dec. 1999.

[4] F. Liu and M. Gleicher, "Video Retargeting: Automating Pan and Scan," in *Proc. ACM International Conference on Multimedia*, New York, 2006, pp. 241–250.

[5] C. Gu, M. Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 8, no. 5, pp. 572-584, Sep.1998.

[6] P. Jiang and X. Qin, "Keyframe-Based Video Summary Using Visual Attention Clues," *IEEE Multimedia*, vol. 17,no. 2,pp. 64-73, 2010.

[7] S. Feng, D. Xu and X. Yang, "Attention-driven salient edge(s) and region(s) extraction with application to CBIR," *Signal Processing*, vol. 90, pp. 1–15, Jan. 2010.

[8] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang and H. Shum, "Picture Collage," *IEEE Trans. on Multimedia*, vol. 11, no. 7, pp.1225-1239, May. 2009.

[9] Z. Chen, J. Han and K. Ngan, "Dynamic Bit Allocation for Multiple Video Object Coding," *IEEE Trans. on Multimedia*, vol. 8, no. 6, pp. 1117-1124, Dec. 2006.

[10] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[11] Y. Ma and H. Zhang, "Contrast-based Image Attention Analysis by Using Fuzzy Growing," in *Proc. ACM International Conference on Multimedia*, 2003, pp. 374-381,

[12] R. Achanta, F. J. Estrada, P. Wils and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. ICVS*, pp. 66–75. May. 2008

[13] D. Gao, V. Mahadevan and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp.1-18, Jun. 2008.

[14] O. Le Meur, P. Le Callet, D. Barba and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell* , vol. 28, no. 5, pp.802-817, May. 2006.

[15] H. Seo and P. Milanfar, "Static and Space-time Visual Saliency Detection by Self-Resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 1-27, Nov. 2009.

[16] V. Gopalakrishnan, Y. Hu and D. Rajan, "Salient Region Detection by Modeling Distributions of Color and Orientation," *IEEE Trans. on multimedia*, vol. 11, no. 5, pp. 892-905, Aug. 2009.

[17] W. Zhang, Q. Wu, G. Wang and H. Yin, "An Adaptive Computational Model for Salient Object Detection," *IEEE Trans. on Multimedia*, vol. 12, no. 4, pp. 300-316, Jun. 2010.

[18] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang and H. Shum, "Learning to Detect a Salient Object," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 33, no. 2, pp. 353-367, Feb. 2011.

[19] T. Judd, K. Ehinger, F. Durand and A. Torralba, "Learning to Predict Where Humans Look," in *Proc..IEEE Int. Conf. Computer Vision*, 2009, pp. 2106-2113.

[20] P. Khuwuthyakorn, A. Robles-Kelly and J. Zhou, "Object of Interest Detection by Saliency Learning," in *Proc. European Conference on Computer Vision*, Greece, 2010, pp. 636-649.

[21] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2007, pp.1-8.

[22] C. Guo, Q. Ma and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, pp. 1-8. 2008.

[23] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1597-1604.

[24] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Conf. on Advances in neural information processing systems*, 2006, pp. 155-162.

[25] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Conf. on Advances in neural information processing systems*, 2008, pp. 681-688.

[26] L. Zhang, M. Tong, T. Marks, H. Shan and G. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 1-20, Dec. 2008.

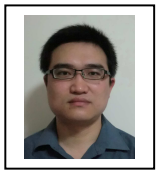
[27] W. Wang, Y. Wang, Q. Huang and W. Gao, "Measuring Visual Saliency by Site Entropy Rate," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2010. pp. 2368–2375.

[28] S. He, J. Han, X. Hu, M. Xu, L. Guo, and T. Liu, "A biologically inspired computational model for image saliency detection," in *Proc. ACM International Conf. on Multimedia*, New York, 2011, pp. 1465-1468.

[29] S. Goferman, L. Zelnik-Manor and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 2376–2383.
[30] Y. Luo, J. Yuan, P. Xue and Q. Tian, "Saliency Density Maximization for Efficient Visual Objects Discovery," *IEEE Trans. on Circuits Syst. Video Techn.*, vol. 21, no. 12, pp. 1822–1834, Dec. 2011.
[31] J. Harel, C. Koch and P. Perona, "Graph-Based Visual Saliency," in *Proc. Conf. on Advances in neural information processing systems*, 2007, pp. 545-553.
[32] N. Murray, M. Vanrell, X. Otazu and C. Parraga, "Saliency Estimation Using a Non-Parametric Low-Level Vision Model," in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 433-440.
[33] J. Li, M. Levine, X. An, and H. He, "Saliency Detection Based on Frequency and Spatial Domain Analyses," in *Proc. British Machine Vision Conference*, UK, 2011, pp. 86.1-86.11.
[34] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell*, Vol. 34, no. 1, pp. 194-201, Jan. 2012.
[35] B. Olshausen, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607-609, Jun. 1996.
[36] J. Reynolds, T. Pasternak and R. Desimone, "Attention Increases Sensitivity of V4 Neurons," *Neuron*, vol. 26, no. 3, pp. 703–714, Jun. 2000.
[37] J. Maunsell and S. Treue, "Feature-based attention in visual cortex," *Trends in Neurosciences*, vol. 29, no.6, pp. 317–322, Jun. 2006.
[38] J. Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," in *Proc. of The Royal Society*, 1998, pp. 359-366.
[39] M. Wainwright, O. Schwartz, and E. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons," In R. Rao, B. Olshausen, & M. Lewicki, (Eds.), Probabilistic models of the brain: Perception and neural function. MIT Press.
[40] Y. Deng and B. S. Manjunath, "Unsupervised Segmentation of Color-Texture Regions in Images and Video," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 23, no. 8, pp. 800-810, Aug. 2001.
[41] A. García-Díaz, X. Fernández-Vidal, X. Pardo, and R. Dosil, "Decorrelation and Distinctiveness Provide with Human-Like Saliency," in *Proc. of ACIVS*, 2009, pp 343-354.
[42] R. Achanta and S. Susstrunk, "Saliency Detection using Maximum Symmetric Surround," in *Proc. IEEE Int. Conf. Image Processing*, 2010, pp. 33-40.
[43] D. Klein and S. Frintrop, "Center-surround Divergence of Feature Statistics for Salient Object Detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 2204–2219.
[44] X. Sun, H. Yao, R. Ji, P. Xu, X. Liu and S. Liu, "Saliency Detection Based on Short-term Sparse Representation," in *Proc. IEEE Int. Conf. Image Processing*, 2010, pp. 1101-1104.
[45] C. Lampert, M. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2129–2142, 2009.
[46] J. Li, D. Xu, and W. Gao, "Removing Label Ambiguity in Learning-based Visual Saliency Estimation," *IEEE Trans. on Image Processing*, vol. 21, no. 4, pp. 1513-1525, April. 2012.
[47] M. Cheng, G. Zhang, N. Mitra, X. Huang, S. Hu, "Global Contrast based Salient Region Detection," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 409–416.
[48] P. Hoyer, A. Hyvärinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network*, vol. 11, no. 3, pp. 191-210, 2000.
[49] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit Rate Coding of Image Sequence using Adaptive Regions of Interest," *IEEE Tran. on Circuits and Systems for Video Technology*, vol. 8, no. 8, pp. 928-934, Dec. 1998.

Junwei Han received his Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2003. He is currently a professor with Northwestern Polytechnical University. His research interests include computer vision and multimedia processing.

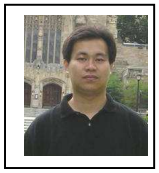
Sheng He received his B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2009. His research interest is computer vision.



Xiaoliang Qian is currently a Ph.D. student in Northwestern Polytechnical University, Xi'an, China. His research interests are computer vision and pattern recognition.

Dongyang Wang is currently a Master student in Northwestern Polytechnical University, Xi'an, China. His research interest is multimedia information retrieval.

Lei Guo received his Ph.D. degree from Xidian University, Xi'an, China, in 1994. He is currently a professor in Northwestern Polytechnical University, China. His research interests include computer vision, pattern recognition, and medical image processing.



Tianming Liu received his Ph.D. degree in computer science from Shanghai Jiaotong University in 2002. He is currently an Assistant Professor of Computer Science at The University of Georgia. His research focuses on computational brain imaging. Before he moved to UGA, he was a faculty member of Weill Medical College of Cornell University and Harvard Medical School. He is the recipient of the Microsoft Fellowship Award and the NIH NIBIBK01 Career Award.