A Unified Framework for Salient Structure Detection by Contour-Guided Visual Search

Kai-Fu Yang, Hui Li, Chao-Yi Li, and Yong-Jie Li, *Member, IEEE*

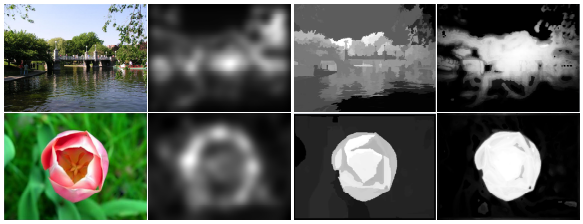
***Abstract*—We define the task of salient structure (SS) detection to unify saliency-related tasks such as fixation prediction, salient object detection, and detection of other structures of interest in cluttered environments. To solve such SS detection tasks, a unified framework inspired by the two-pathway-based search strategy of biological vision is proposed in the present study. First, a contour-based spatial prior (CBSP) is extracted based on the layout of edges in the given scene along a fast non-selective pathway, which provides a rough, task-irrelevant and robust estimation of the locations where the potential SSs are present. Second, another flow of local feature extraction is executed in parallel along the selective pathway. Finally, Bayesian inference is used to auto-weight and integrate local cues guided by CBSP and to predict the exact locations of SSs. This model is invariant to the size and features of objects. Experimental results on six large datasets (three fixation prediction datasets and three salient object datasets) demonstrate that our system achieves competitive performance for SS detection (i.e., both the tasks of fixation prediction and salient object detection) compared to the state-of-the-art methods. In addition, our system also performs well for salient object construction from saliency maps and can be easily extended for salient edge detection.**

***Index Terms*—visual search, fixation, salient object, salient edge, salient structure detection, Bayesian inference**

I. Introduction

VISUAL search is necessary for rapid scene analysis in daily life because information processing in the visual system is limited to one or a few targets or regions at one time [1]. To select potential regions or objects of interest rapidly in a task-independent manner, the so-called "visual saliency" is important for reducing the complexity of scene analysis. From the perspective of engineering, modeling visual saliency usually facilitates subsequent higher visual processing, such as image re-targeting [2], image compression [3], and object recognition [4].

This paper introduces several saliency-related concepts that need to be first clarified. (1) *Fixations* are usually related to human fixating points recorded by eye-tracker. Human fixations are usually used as the ground truth to benchmark fixation prediction methods. (2) *Regions of Interest (ROIs)* represent some regions containing interesting information, in which clear objects cannot be easily segregated from others. Although benchmarked on human fixations, fixation prediction methods usually obtain smoothed ROIs of scene, and 'Regions of Interest' and 'Saliency Map' are interchangeable when describing the output of fixation prediction methods [5]. (3) *Salient Objects* are specific dominant objects (e.g., animals, people, cars, etc.) in natural scenes, and, in general, salient object detection requires labeling pixel-accurate object silhouettes [6]. (4) *Salient Edges* are only the edges of salient objects [7]. In contrast, we propose a new model for *Salient Structure (SS) detection*, a more general term defined as the task of detecting the accurate regions containing structures of interest (e.g., ROIs, salient objects, salient edges, etc.) in a scene. This means that such a task aims to detect the ROIs in cluttered scenes and identify dominant objects in simple scenes.



(a) Image (b) IT (c) HS (d) CGVS

Fig. 1. Compared with the tasks of fixation prediction (b) and salient object detection (c), our salient structure detection (d) aims to extract interesting structures from both complex and simple scenes (a).

Fig. 1 shows two examples of SS detection. Fixation prediction methods (e.g., IT [5]) usually focus on high-contrast boundaries, but ignore object surfaces and shapes (Fig. 1b). In contrast, salient object detection models (e.g., HS [8]) may be inefficient for ROI detection in complex scenes without dominant objects (Fig. 1c). The proposed method is designed to extract SSs for both simple and complex scenes (Fig. 1d).

It is widely accepted that visual attention is influenced by both the *bottom-up* mechanism that processes the input visual scene and the *top-down* mechanism that introduces the visual knowledge about the scene and (or) the target [9]. Among the various theories explaining the biological mechanisms of integration between bottom-up and top-down information [9], [10], Feature Integration Theory(FIT) [11], [12], Guided Search Theory(GST) [1], [13] and Biased Competition Theory(BCT) [14] are three influential theories.

Explicitly or implicitly inspired by these visual attention

Manuscript received July 28, 2015; revised December 20, 2015 and May 17, 2016; accepted May 21, 2015. This work was supported by the Major State Basic Research Program under Grant 2013CB329401, and the Natural Science Foundations of China under Grant 61375115, 91420105. *(Corresponding author: Yong-Jie Li)*

K.-F. Yang, H. Li, and Y.-J. Li are with the Key Laboratory of Neuroinformation of Ministry of Education, University of Electronic Science and Technology of China, Chengdu 610054, China (email:yang_kf@163.com, li_hui_2015@163.com, liyj@uestc.edu.cn).

C.-Y. Li is School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China, and the Center for Life Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China (email:cyli@sibs.ac.cn)

theories, computational modeling of visual searching is gaining increasing attention, mainly aiming for fixation prediction [15] or salient object detection [16]. Note that in this paper we use the term "fixation prediction" to unify the computational tasks of attention models and eye movement prediction models because both of these models are often validated against eye movement data although the two types of models have slight differences in scope and approaches [15].

Most of the existing fixation prediction models are built on the biologically-inspired architecture based on the famous *FIT* [11], [12]. For instance, Itti *et al.* proposed a famous FIT-inspired saliency model that computes the saliency map with local contrast in multiple feature dimensions, such as color, orientation, etc. [5], [17]. However, FIT-based methods may risk being immersed in local saliency (e.g., object boundaries) because they employ local contrast of features in limited regions and ignore global information.

In contrast to FIT, GST [1], [13] suggests visual mechanisms to search for ROIs or objects with the guidance from scene layout or top-down sources. The recent version of GST claims that the visual system searches for the objects of interest along two parallel pathways: the non-selective pathway and the selective pathway [1]. This strategy allows observers to extract spatial layout (or gist) information rapidly from the entire scene via the non-selective pathway. This global information of scene acts as top-down modulation to guide the salient object search in the selective pathway. This two-pathway-based search strategy provides parallel processing of global and local information for rapid visual search.

In our two-pathway-based *contour-guided visual search (CGVS)* framework, the contour-represented layout in the non-selective pathway is used as the initial guidance to estimate the location and sizes of ROIs and the relative importance of low-level local cues. In contrast, the local cues (e.g., color, luminance, texture, etc.) are extracted in parallel along the selective pathway. Finally, Bayesian inference is used to integrate the contour-based spatial prior (CBSP) and the local cues to predict the saliency of each pixel. The salient structures are further enhanced via iterative processing to refine the prior guidance as the final prediction.

The proposed system attempts to bridge the gap between the two highly related tasks of human fixation prediction [15] and salient object detection [16], with a general framework. Extensive evaluations demonstrate that our model can handle both tasks well without specific tuning. In addition, the proposed method can efficiently transform the saliency maps of fixation prediction methods to salient objects. Furthermore, we apply our system for other saliency related tasks, such as salient edge detection. Our experiments show that the proposed system can be flexibly extended for related tasks.

To summarize the above, this work draws its inspiration directly from the biological visual search theory, and the contributions of the proposed model are as follows. (1) We define a new task called salient structure (SS) detection to unify various saliency-related tasks, including fixation prediction, salient object detection, and salient edge extraction. (2) We propose a unified framework for SS detection inspired by *Guided Search Theory*. (3) A simple yet efficient filling-in operation is implemented to create the global layout used for location prior of potential salient structures. (4) Simple yet robust methods are proposed to automatically define the sizes of potential salient structures and estimate the weight of each bottom-up feature. (5) The proposed system also provides an efficient way for transforming saliency maps (outputted from any model) to salient objects and multi-object searches.

II. Related Work

A. Fixation Prediction

As mentioned above, the existing fixation prediction models aim to compute 'Saliency Maps' to indicate the ROIs where human fixations locate [15], [18]. Ever since Koch and Ullman [12] proposed the concept of the *saliency map* and Itti *et al.* [5] proposed their model, many fixation prediction methods have been proposed to predict human fixations with a bottom-up framework [15], [19]. Some typical methods along this line include Graph-based (GB) [20], Information Maximization (AIM) [21], Image Signature (SIG) [22], Adaptive Whitening Saliency (AWS) [23], Local and Global Patch Rarities [24], and Earth Mover's Distance-Based Saliency Measurement and Nonlinear Feature Combination for static and dynamic saliency maps [25], [26]. Frequency domain based models include Spectral Residual (SR) [27], Phase spectrum of Quaternion Fourier Transform (PQFT) [28], [29], and Hypercomplex Fourier Transform (HFT) [30].

In addition, machine learning techniques have also been introduced to improve the performance of fixation prediction. In these models, both bottom-up and top-down visual features are learnt to predict salient locations [31], [32]. In general, interesting objects (such as humans, faces, cars, text, animals, etc.) convey more information in a scene, and they usually attract more human gaze [33]–[36]. Task-related top-down information is also commonly used [37]–[40]. Some models also learn optimal weights for channel combination in the bottom-up architecture [41], and nonparametric saliency models learn directly from human eye movement data [42].

In addition to scene context, the observer's current task also exerts dominant top-down control of visual attention [43]. This makes the investigation on the role of top-down guidance along two lines: 1) goal directed, task-driven control and 2) scene contextual information directed, task-irrelevant control. Zhang *et al.* [44] proposed a Bayesian framework for saliency computation using the statistics of natural images. Torralba *et al.* [45], [46] used global features to guide object search by summarizing the probability regions of presence of target objects in the scene. Itti and Baldi [47] proposed the Bayesian definition of surprise by measuring the difference between posterior and prior beliefs of the observer. Lu *et al*. computed saliency map from image Co-Occurrence histograms [48].

Fixation prediction models usually provide smoothing regions of interest rather than uniform regions highlighting the entire salient objects [16]. Though fixation points indicate the location information of the potential objects but miss some object-related information (e.g., object shapes and surfaces) that is necessary for further high-level tasks such as object detection and recognition.

B. Salient Object Detection

To accurately extract the dominant objects from natural scenes, Liu *et al.* [49] formulated the salient object detection as a binary labeling problem. Achanta *et al.* [6] further claimed that salient object detection requires labeling pixel-accurate object silhouettes. Most of the existing methods attempt to detect the most salient object based on local or global region contrast [8], [50]–[52]. For example, Cheng *et al.* [50] proposed a region-based method for salient object detection by measuring the global contrast between the target and other regions. Other methods include center-surround contrasts with Kullback-Leiblar [53], background prior [54], [55], etc. Salient object detection is also highly related to another task called *object proposal*, which attempts to generate a set of all objects in the scene, regardless of the specific saliency of these objects [56]–[58]. Recent advances reveal that the state-of-the-art models can produce excellent results when evaluated using the traditional benchmark images containing a clear, unambiguous object of interest [16]. However, when facing randomly-selected images (e.g., from the Internet), there is still a strong need to develop more robust methods [50].

For attentional modeling, Bayesian inference seems a reasonable tool for combining visual evidence with prior constraints by taking advantage of the statistics of natural scenes or other features that attract attention. In fact, many physiological experiments and computational analyses strongly suggest that attentional modulation in the biological visual system may arise as a consequence of Bayesian inference in cortical networks [10], [44], [47], [59], [60]. The biological plausibility of Bayesian inference inspires us to adopt it to combine the contour-based guidance and bottom-up features.

A related model is proposed by Xie *et al.* [61], who employ Bayesian inference for saliency computation, with the help of prior information estimated using the convex hull of Harris points and a Laplacian sparse subspace clustering algorithm. Unlike their method, our method obtains the spatial prior information in a much simpler way. The contour-based information in our model is used to identify the sizes of potential objects and the importance of local cues. In addition, our model provides a unified framework for various saliency-related tasks (including fixation prediction, salient object detection and salient edge extraction), termed "Salient Structure Detection" in the present study (see Fig. 1).

C. Bridging the Two Tasks

More recently, several authors attempted to bridge the gap between the two tasks of fixation prediction and salient object detection. Typically, Goferman *et al.* [2] proposed a context-aware saliency algorithm to detect both prominent objects and the parts of the background that convey the context. Li *et al.* [18] trained a random regression forest to extract salient regions based on an image segmentation method. In comparison to [2], our method can obtain more reasonable salient structures with accurate object silhouettes and object surfaces. Compared with [18], the proposed model is capable of handling the task of salient structure detection in both simple and complex scenes without foregone computation such as image segmentation.

A method based on *Boolean Map* was originally proposed for fixation prediction [62] but could also be used for salient object detection by adding specific tuning and post-processing. The recent method proposed by Cheng *et al.* [50], which targets salient object detection, is also capable of obtaining acceptable performance for fixation prediction by adjusting their model's implementation. In contrast, our method can achieve both tasks without the need for specific tuning.

III. Contour-Guided Visual Search Model

In general, contours or boundaries help to segment an image into various perceptional regions (before the perception of specific objects) that may be used by the visual system to rapidly construct a rough sketch of the image structure in space [63]. In addition, contours (shapes) lead to perceptual saliency of different geometrical properties, which have been strongly proven to contribute to early global topological perception [64]. Furthermore, physiological evidence shows that the global contours delineating the outlines of visual objects may respond quite early (perhaps via a special pathway) through the neurons of high cortexes, which, although producing only a crude signal about the position and shape of the objects, can provide sufficient feedback modulation that enhances contour-related responses at lower levels and suppresses irrelevant background input [65]. These facts inspired us to detect salient stuctures guided by dominant edges.

The flowchart of the proposed method is summarized in Fig. 2. The possible locations of potential salient structures are estimated with the distribution of dominant edges in the non-selective pathway. Meanwhile, the local features such as color, luminance, and texture are extracted from the given scene in the selective pathway. The contour-based spatial prior (CBSP) information is fed into the Bayesian framework for feature integration and salient structure prediction. Finally, the output of the Bayesian framework is used as new spatial prior to refine the salient structures with an iterative procedure.

A. Contour-based Spatial Prior (CBSP)

In the non-selective pathway, we compute the rough spatial weights of saliency based on the distribution of the dominant edges. In fact, edge information has been widely used for saliency computation [5], [66]. However, it is difficult to use these methods to provide regional information (e.g., object surfaces), and some isolated and high-contrast edges (e.g., the boundary between two large surfaces) may be incorrectly evaluated as high saliency.

In this paper, we roughly identify the potential salient regions with a visual filling-in-like operation [67] based on the dominant contours. Some authors have proposed that boundaries are a type of useful information that can be used to block the lateral spreading or diffusion in visual perception and achieve filling-in of surfaces [68], [69]. This inspires us to build a new implementation of filling-in for the computation of *Contour-based Spatial Prior (CBSP)* based on the dominant contours.

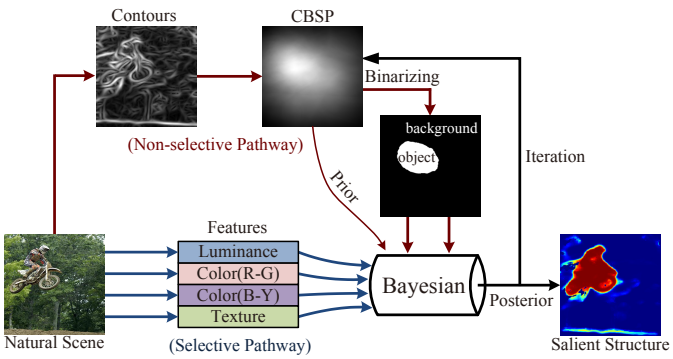


Fig. 2. The flowchart of the proposed system with selective and non-selective pathways.

In detail, we first extract the edge responses and the corresponding orientations using the edge detector proposed in [70], which is a biologically-inspired method and can efficiently detect both color- and brightness-defined dominant boundaries from cluttered scenes. Fig. 3 shows an example of reconstructing CBSP. For each edge pixel, we compute the average edge response (AER) in the "left" and "right" half disk around it. The disk is defined by the orientation of each edge pixel, with a radius of d_r. Considering the fact that radius d_r would be strongly dependent on the image size, viewing distance, etc., we select d_r as a relative value of image size instead of an absolute value. In this work, we experimentally set $d_r = min(W, H)/3$, where W and H indicate the width and height of the given image, respectively. Then, all of the pixels within the half disk having stronger AER between two half disks are voted 1, and the pixels in the other half are voted 0. For each pixel, its saliency weight is represented by the number of votes when all of the edge pixels finish their voting. We only scan the dominant edges (i.e., the ridges, red pixels in Fig. 3(middle)) to speed up the computation.

We denote the rough spatial weights of saliency as S_e. In addition, we also consider the widely-used center-bias weighting [31], [71] (denoted by S_c), which is simply modeled by a Gaussian mask with the standard deviation as $\sigma_c = min(W, H)/3$, based on a similar consideration with that of d_r. S_e and S_c are linearly normalized to the range of $[0, 1]$. Then, the final CBSP is given by

$$S_w = S_e + S_c \tag{1}$$

S_w is also linearly normalized to the range of $[0, 1]$ for later used as the prior probability.

B. Low-level Feature Extraction

In the selective pathway, basic low-level features including color, luminance and texture are extracted in parallel. With r, g, and b denoting the red, green, and blue components of the input image, the luminance (f_{lum}) and the two color-opponent channels (f_{rg} and f_{by}) are obtained as

$$f_{lum} = (r + g + b)/3 \tag{2}$$

$$f_{rg} = r - g \tag{3}$$

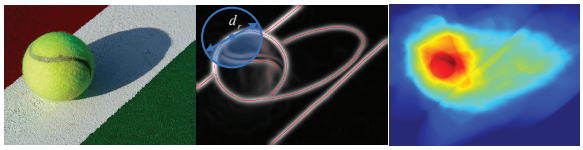


Fig. 3. Example of reconstructing the potential saliency regions based on the dominant edges. **Left**: the input image. **Middle**: the dominant edges shown in red lines. For each edge pixel, the averaged edge responses in the "left" and "right" half disks around it are compared to decide which half is located in the salient region. **Right**: the potential saliency regions.

$$f_{by} = b - (r + g)/2 \tag{4}$$

Note that these three basic channels are smoothed with a Gaussian filter with the same scale as that used in our edge detector [70] to remove noise.

In addition, considering that texture can reflect more complicated properties of image regions, we added a texture channel for a better representation of potential salient structures. In this work, the texture channel (f_{ed}) is represented by the information of edge density (ED), which reflects the contrast of local patches, because some authors have claimed that local luminance and luminance contrast are independent or weakly dependent with each other in the early visual system and in natural scenes [72], [73], although inconsistent conclusions have also been reported [74]. f_{ed} is computed by smoothing the edge responses (the same as used in the non-selective pathway, Section III-A) with an average filter of 11×11 pixels.

C. Bayesian Inference with Contour Guidance

In this paper, we employ the tool of Bayesian inference to adaptively integrate the global CBSP and the local features, simulating the interaction of top-down and bottom-up information processing flows in the selective visual attention.

With Bayesian inference, the possibility of a pixel at x belonging to a salient structure s (posterior probability), $p(s|x)$, can be computed as

$$p(s|x) = \frac{p(s)p(x|s)}{p(s)p(x|s) + p(b)p(x|b)} \tag{5}$$

where $p(s)$ and $p(b) = 1 - p(s)$ are the prior probabilities of a pixel at x belonging to a salient structure and the background, respectively. $p(x|s)$ and $p(x|b)$ are likelihood functions based on the observed salient structure and the background, respectively. In this work, we set the CBSP as the initial prior probability, i.e., $p(s) = S_w$. The observation likelihood $p(x|s)$ and $p(x|b)$ will be evaluated according to each scene context, including the possible sizes of salient structures and the relative importance of each feature. The implementation details are as follows.

1) Predict the size of potential structure: To obtain $p(x|s)$ and $p(x|b)$, the observation likelihood of the observed objects and background, we first extract the possible regions containing structures from the background. Simply, we binarize the map of prior probability ($p(s)$) with an adaptive threshold to capture rough potential regions of structures and their sizes.

We use S_{T_k} and B_{T_k} to denote the pixel sets of structure and background obtained by binarizing $p(s)$ with a threshold

T_k. The optimal threshold T_{opt} is found by searching for a possible T_k that maximizes the difference of all of the features between the structure and background pixel sets according to

$$T_{opt} = arg \max_{T_k} \sqrt{\sum\nolimits_i \left(\omega_i^0 \cdot \left(\tilde{S}_{T_k}^i - \tilde{B}_{T_k}^i\right)\right)^2} \quad (6)$$

where $\tilde{S}_{T_k}^i$ and $\tilde{B}_{T_k}^i$ denote the mean values of the structure and background pixels, respectively, in feature channel i, $i \in \{f_{lum},\ f_{rg},\ f_{by},\ f_{ed}\}$. The initial feature weight is $w_i^0 = 0.25$, which indicates the equal importance of each cue at the initial status. $T_k \in \{10\%, 12\%, 14\%, ..., 50\%\}$ indicates the percentage of pixels of the potential salient structures. This suggests a potential assumption that salient structures are usually smaller than half of the image. We ignore the regions with fine scales (<10%) to avoid fragments. This assumption is supported by a simple experiment on two popular salient object datasets: the mean sizes (percentage) of salient objects are 20.01% on ASD [6] and 28.51% on ECSSD datasets [8]. The influence on images with smaller and larger objects is discussed in Section IV-F.

2) Evaluate the importance of each feature: After finding the potential salient and background pixel sets for each feature map based on the optimal threshold T_{opt}, we re-evaluate the importance of each feature as

$$\omega_i = \frac{1}{\mu} \cdot \left|\tilde{S}_{T_{opt}}^i - \tilde{B}_{T_{opt}}^i\right| \quad (7)$$

where $\mu = \sum_i \omega_i$, $i \in \{f_{lum},\ f_{rg},\ f_{by},\ f_{ed}\}$. Equation (7) indicates that a feature will have higher importance when the difference of mean pixel values between the salient structure and background is larger in this channel.

3) Calculate the observation likelihood: We then compute the observation likelihood $p(x|s)$ and $p(x|b)$ with the potential object and background pixel sets and the weight of each feature. We assume that the four feature channels are independent to simplify the estimation of the joint probability distribution of various features. Note that this is a bold assumption. Though there are certainly interactions between the channels of texture and color contrast [75], the independence of luminance vs. contrast and luminance vs. color are still controversial [72]–[74], [76]. Nevertheless, this bold assumption proves acceptable for practical applications, such as the famous *Naive Bayes Classifier* in machine learning [61], [77], [78]. Thus, the observation likelihood at pixel x can be computed as

$$p(x|s) = \prod_{i \in \{f_{lum}, f_{rg}, f_{by}, f_{ed}\}} \left(p(x_i|S_{T_{opt}})\right)^{\omega_i} \quad (8)$$

$$p(x|b) = \prod_{i \in \{f_{lum}, f_{rg}, f_{by}, f_{ed}\}} \left(p(x_i|B_{T_{opt}})\right)^{\omega_i} \quad (9)$$

where $p(x_i|S_{T_{opt}})$ and $p(x_i|B_{T_{opt}})$ are respectively the distribution functions of each feature ($i \in \{f_{lum}, f_{rg}, f_{by}, f_{ed}\}$) in the salient structure and background sets. We simplify the computation of the observation likelihood of each feature based on the normalized histogram of pixels in salient or background sets in each feature channel. Specifically, $p(x_i|S_{T_{opt}})$ is computed as $N(x_i)/N(S_{T_{opt}})$, where $N(x_i)$ is the number of points in various disjoined bins containing feature x_i in the region of $S_{T_{opt}}$, and $N(S_{T_{opt}})$ is the total pixel number in the region of $S_{T_{opt}}$. $p(x_i|B_{T_{opt}})$ is computed similarly. ω_i indicates the contribution of the distribution function of the i^{th} feature, which will be iteratively modified as follows. Finally, $p(s|x)$ is computed using (5) as the saliency of each pixel.

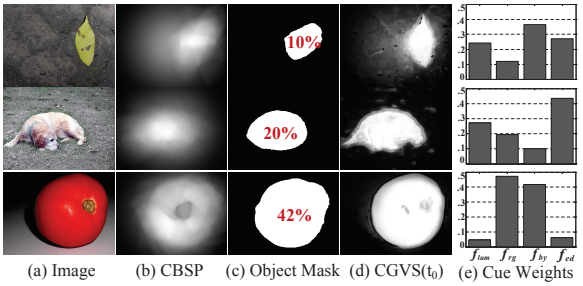


Fig. 4. Examples illustrating that our system automatically selects the size of potential salient structure and estimates the importance of each local feature. The object mask in (c) is obtained by thresholding the CBSP in (b) with the currently optimal threshold (i.e., the percentage value shown in (c)).

4) Enhance the salient structure by iterating: We further enhance the salient structures iteratively by re-initializing the prior function with $p(s) \leftarrow p(s|x)$ and the feature weights as $\omega_i^0 \leftarrow \omega_i$. In the experiment, we re-initialize the prior function with the smoothed version of $p(s|x)$ (by median filtering with a size of 21×21 pixels) to remove small fragments. Finally, we denote $CGVS(t)$, $t = t_0, t_1, ..., t_n$ as our contour-guided visual search model with various iterations (t), and the $CGVS(t_0)$ is the first step's output without iteration.

IV. Experiments

We first show the basic properties of our system in scene analysis. Then the proposed method is evaluated on both fixation prediction datasets [30], [31], [79] and salient object detection datasets [6], [8], [18]. In addition, we demonstrate that with the proposed system, fixation prediction methods can be significantly improved for the task of salient object detection. Our system is also tested to demonstrate that the proposed method can be easily extended for general salient structure detection. Finally, we exploit the influence of model parameters on detection performance.

A. Basic Property of the Proposed Model

To clearly demonstrate how the proposed system works, we first show several of its basic properties, including: (1) automatic selection of the sizes of potential salient structures and the relative importance (i.e., weight) of each feature, (2) the ability to search for objects in multi-object scenes, and (3) the contribution of iterative processing.

Fig. 4 shows three examples including objects with different spatial scales. With (6), our system can automatically select the spatial sizes of potential objects in the given scene and roughly evaluate and identify the pixels of salient regions and background (Fig. 4(c)) by thresholding the CBSP (Fig. 4(b)) with certain threshold values. Then, the weight of each feature is computed with (7), and the observed likelihood functions of salient regions and background are evaluated with (8) and

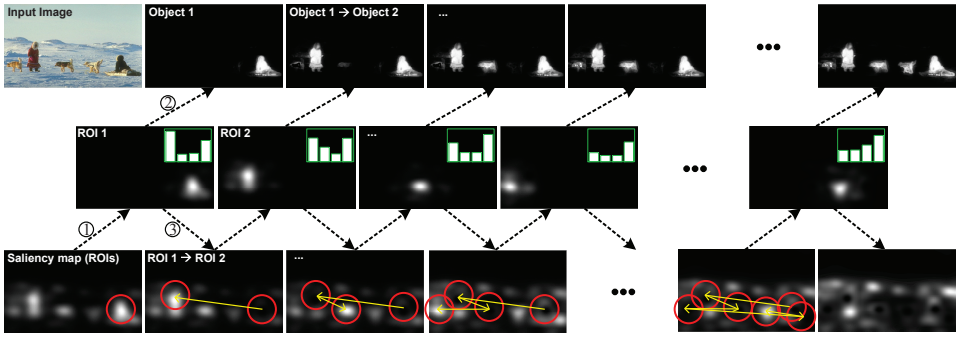


Fig. 5. The procedure of object searching in a multi-object scene. For an input image, we first obtain the smoothed saliency map (e.g., using Itti's method [5]), which contains several regions of interest (ROIs). The most salient location (ROI 1) is extract using "winner-take-all" mechanisms (step ①), and then the first object is detected with the proposed system (step ②) with ROI 1 as the spatial prior. The "inhibition-of-return" mechanism is employed to search for the next salient location (ROI 2), i.e., removing the scanned locations (step ③). Repeating these procedures, our system can extract all of the objects one by one (the top row) from each attended location instead of the classical shift of focus in Itti's model [5].

(9). Finally, the possibility of a pixel belonging to the salient structure (Fig. 4(d)) is obtained using Bayesian inference. Fig. 4(e) lists the weights of all of the features. We can clearly see that our system obtains reasonable evaluations about the size of salient structures and the importance of features of the input scene, which are important in searching for task-free interesting structures in cluttered scenes. In fact, this *auto-weighting* of different features is quite consistent with the guided search theory [1], [13], which proposes that visual search can be biased toward ROIs by modulating the relative weights through which different types of features contribute to attention adaptively.

An additional experiment was executed to model the process of object searching in multiple object scenes. Fig. 5 shows the object searching procedure beginning with an initial saliency map. Based on the previous work of Itti *et al.* [5], the mechanisms of "winner-take-all" and "inhibition-of-return" were employed to search for the salient locations. The bottom row of Fig. 5 shows the classical shift of focus of attention modeled using the method of Itti *et al.* [5]. Compared to Itti's model, our system further extracts the full structure (the top row) from each attended location (the middle row). The objects were found one by one over the time course. In addition, our method provides the importance of each feature for each object, which is indicated by the histogram shown in Fig. 5 (the middle row). We believe that these extracted features with auto-defined weights are also useful for further computer vision applications such as object recognition.

In addition, our model further re-evaluates the salient structures with new prior and feature weights to improve the confidence of salient structures. Fig. 6 shows two examples that illustrate that our system can always correctly identify the salient objects, although the initial CBSP (Fig. 6(b)) is inaccurate. For example, in the *bear* image (Fig. 6(a), the top row), the most salient location is on the head of the bear, and most parts of the bear's body are missed in the initial CBSP. However, after two steps of iteration, our model detects the full bear while suppressing the background (Fig. 6(c)-(e)).

It is also worth noting the different contributions of the non-selective pathway and selective pathway. When turning down the selective pathway, our model obtains the CBSP only along the non-selective pathway, which can be used for fixation prediction, as demonstrated in the following section. In contrast, when the non-selective pathway is turned down, the proposed framework would be reduced to a simple model if the Bayesian inference is replaced by a center-surround operation and linear combination, which would be quite similar to the classical bottom-up Itti (i.e., IT) model for fixation prediction.

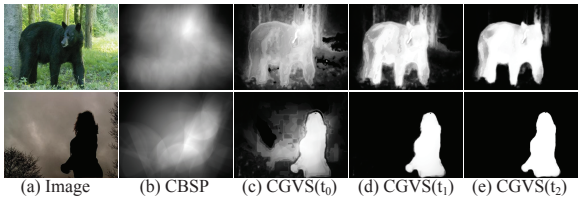


Fig. 6. Enhancing the salient structure in an iterative way. (a) Original image, (b) CBSP, (c)-(e) Results of our method at different iteration steps.

B. Fixation Prediction

Fixation prediction methods are usually benchmarked on some available human fixation datasets [30], [31]. As a common metric for fixation prediction, Natural Scanpath Saliency (NSS) aims to measure the correspondence between the salience map and scanpath [80]. Basically, both NSS and the original version of ROC reflect the combined effect of the true positive rate and false positive rate [81]. To fairly evaluate the fixation prediction ability based on the saliency map produced by a salient object detection model (such as the proposed model), it seems more important to focus mainly on the true positive rate, for which the implementation version of ROC (AUC) in [31] is an appropriate choice, because it is insensitive to false positive rate [41]. Nevertheless, in this section, we evaluated several datasets with both the NSS and the ROC (AUC) in [31] to better understand the fixation prediction performance of the proposed model. The dataset collected by Judd *et al.* [31] contains 1003 images and is widely used, whereas the dataset collected by Li *et al.* [30] contains 235 images with various sizes of regions of interest. Moreover, a recent SALICON dataset (5000 images) [79], which contains a large number of non-iconic view objects, is also considered.

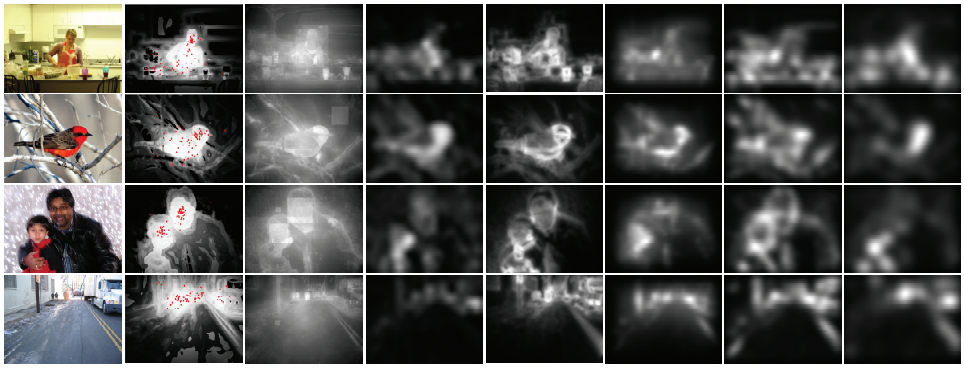


Fig. 7. Comparison of the fixation prediction results. (a) original images, (b) salient structure produced using the proposed method (CGVS) without the iterative processing, and saliency maps produced using multiple methods: (c) Judd *et al.* (Judd) [31], (d) Boolean map (BMS) [62], (e) Context-aware (CA) [2], (f) Graph-Based (GB) [20], (g) Itti *et al.* (IT) [5], and (h) Image Signature (SIG) [22]. Note that the red points on the CGVS maps indicate the human fixations (ground truth). It is clear that our CGVS generates uniformly highlighted salient structures that cover almost all of the human fixations.

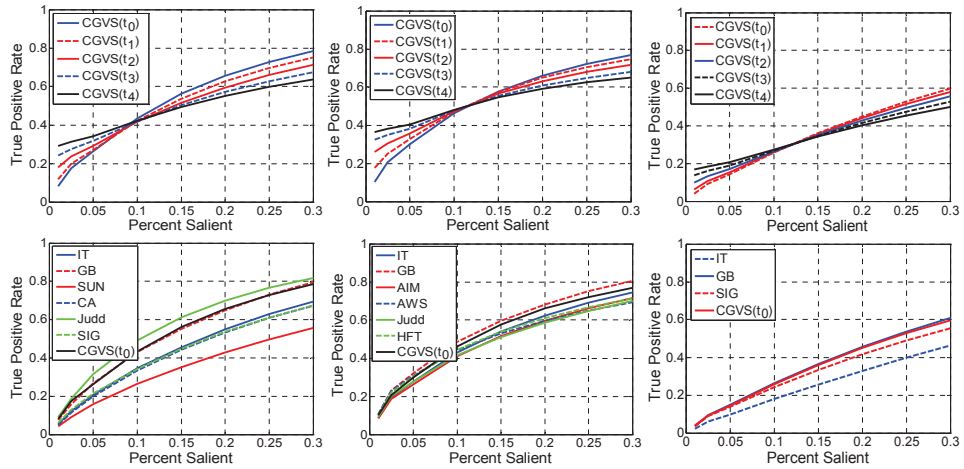


Fig. 8. Quantitative evaluation on three fixation prediction datasets. (a) Judd [31], (b) ImgSal [30], and (c) SALICON dataset [79]. **Top**: ROC curves for our method with various iteration steps, showing that $CGVS(t_0)$ obtains the best performance. **Bottom**: ROC curves for various algorithms, indicating that our model outperforms most of the methods and is comparable to GB [20]. "Percent Salient" denotes the percentage of the predicted saliency maps.

From Fig. 7, some existing methods usually provide stronger responses to regions with higher local contrasts, such as edges or boundaries of objects (e.g., CA [2]), while ignoring the surfaces of salient structures. Others models can obtain a highly blurred saliency map, which cannot provide fine shapes or structures of objects. In contrast, our CGVS method is efficient for various situations. For simple scenes with predominant objects, CGVS can respond well to full objects (Fig. 7, the first to second rows). In addition, our method is also efficient when scenes contain multiple objects (two objects in Fig. 7, the third row) or widely spread interesting regions (Fig. 7, the fourth row). In short, our CGVS contributes to saliency computation in both simple and complex scenes.

We further evaluated the performance of our CGVS for the task of fixation prediction. Fig. 8(top) shows the ROC curves of CGVS with various iterations on three datasets. In general, $CGVS(t_0)$ achieves the best performance for fixation prediction because the iterative processing makes our system focus on the most salient objects in scenes, which decreases the precision of fixation prediction. Fig. 8 (bottom) shows that our method outperforms (or at least matches) all of the considered bottom-up (low-level) methods. Note that the method proposed by Judd *et al.* [31] achieves better performance mainly because several high-level feature related operations, such as face detection and person detection, are introduced into their model. Table I lists the comparison of AUC and NSS on the Judd and SALICON datasets.

TABLE I
AUC AND NSS COMPARISONS ON TWO LARGER DATASET.

Dataset	Judd				SALICON			
	IT	GB	Judd	**CGVS**	IT	SIG	GB	**CGVS**
AUC	0.770	0.824	0.839	**0.807**	0.637	0.694	0.727	**0.703**
NSS	1.103	1.382	1.346	**1.222**	0.455	0.704	0.825	**0.745**

Note that the measure of ROC is somewhat biased when evaluating the performance of fixation prediction [81]. As indicated by Goferman *et al.* [2], incorporating a center prior to the final saliency estimation can remarkably improve quantitative evaluation, but make the saliency map look less visually convincing. Fig. 9 shows examples indicating that CA with

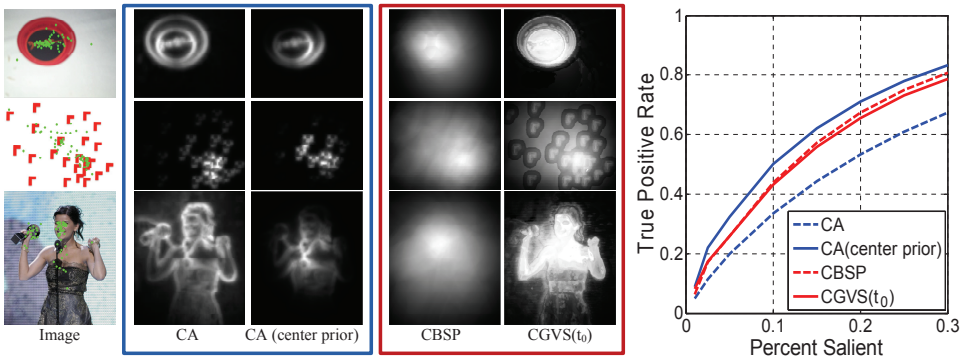


Fig. 9. Examples showing the gap between qualitative and quantitative evaluation. Quantitatively, CA provides clear improvement when introducing the center prior, but the visual assessment decreases substantially. Similarly, although quantitative evaluation of our method is a little worse than that of CA with the center prior, our CGVS obtains excellent assessment when visualizing the saliency map, which highlights almost all of the pixels of the dominant objects.

center prior obtains high performance in ROC (Fig. 9, the last column) but misses much object information when visualizing the saliency maps (Fig. 9, the third column). Conversely, CA without center prior provides better results with qualitative evaluation (Fig. 9, the second column) but a lower score on the ROC curve.

In general, two possible reasons may result in center bias. (1) In most eye-tracking experiments, image viewing is driven through a central fixation point before image presentation, and such center bias of human fixations can not be completely reduced by discarding several fixations at the beginning of the recording [31]. (2) Some photographers are apt to locate the interesting contents at the center of view when taking pictures. Actually, most of the existing methods promote higher saliency values in the center of the image plane, such as GBVS [20], Judd [31], etc. The proposed method also combines the center prior when computing CBSP. With similar observation with [2], CBSP achieves good performance on the ROC curve with the very blurred saliency region (Fig. 9, the fourth column) regardless of the structure information of objects, which demonstrates the contribution of the non-selective pathway when turning down the selective pathway. However, our final CGVS is usually capable of detecting full objects and surfaces (Fig. 9, the fifth column) and achieving high performance on the ROC at the same time (Fig. 9, the last column). Compared to CA with or without a center prior, CGVS achieves qualitatively better results, although CA with a center prior achieves higher performance on the ROC curve, because CBSP is just a rough estimate of potential salient regions, and the saliency of each pixel is eventually identified with Bayesian Inference. Therefore, the center prior introduced in CBSP will not weaken greatly the salient structures close to the border of images.

In addition, there are several other metrics for saliency evaluation, and some of them (e.g., shuffled AUC [44]) are expected to tackle the influence of the center prior [81]. However, all of these metrics for evaluating fixation prediction are computed against human fixations that are extremely sparse. Some salient structures such as large object surfaces extracted using our method may be incorrectly treated as false alarms by these metrics, which will unfairly evaluate the ability of our method to extract objects' regional information.

C. Salient Object Detection

Salient object detection methods are commonly benchmarked by binary pixel-accurate object masks [16]. In this experiment, we first used the standard F-measure (P-R curve and F-score) for performance evaluation on two popular datasets with different peculiarities: ASD [6] includes simple scenes, whereas ECSSD [8] includes more complex scenes. We also evaluated various methods on the PASCAL-S dataset [18], which was designed to be less biased. In addition, considering the analysis of Margolin *et al.* [83] showing that the F-measure does not always provide a reliable evaluation for salient object detection, we also employed the amended F-measure (called Weighted F-score) proposed in [83] and the measure termed Mean Absolute Error (MAE) [51], [55], [84].

Fig. 10 lists the performance of our methods with both the standard P-R curves and the Weighted F-scores on the ASD and ECSSD datasets. In contrast to the fixation prediction in Section IV-B, the iterative process can further enhance the regions of objects and improve the performance because the benchmark used in this experiment has few salient objects. In general, our CGVS can obtain a stable performance with only two steps of iteration (i.e., $CGVS(t_2)$).

Fig. 11 shows the *P-R Curve*, *F-score*, *Weighted F-score*, and *MAE* for various state-of-the-art salient object methods on the three datasets. From the standard P-R curve and F-score, our $CGVS(t_2)$ outperforms most of the considered algorithms except HS [8] and RF [18]. However, our method achieves high F-scores across a large range of thresholds (Fig. 11, the second column) on all of the considered datasets. This result indicates that the proposed model is capable of obtaining salient objects with high confidence. In addition, our method significantly outperforms all of the considered models on the measures of weighted F-score and MAE (except RF on PASCAL-S). Totally, our system achieves quite competitive performance compared to the state-of-the-art methods in simple scenes. In addition, considering the fact that most of the methods compared here do not employ the center bias, we also show the performance of our CGVS without center bias (denoted by $CGVS$") for fair comparison, as shown in Fig.11.

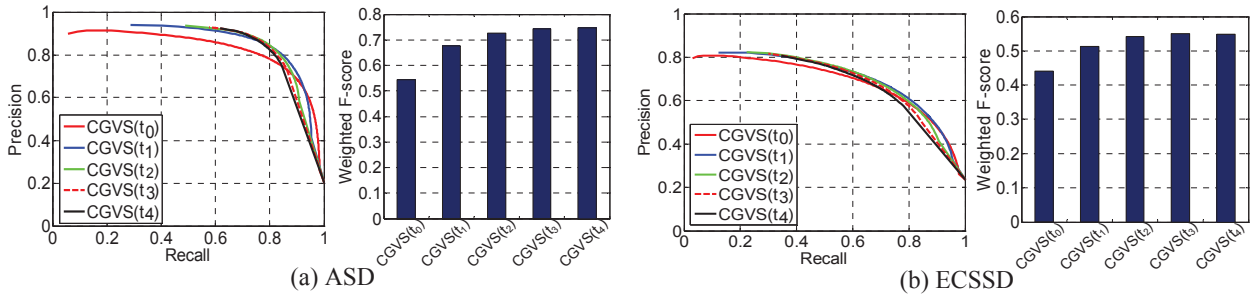


Fig. 10. Quantitative evaluation on two salient object datasets (a) ASD [6] and (b) ECSSD [8]. **Left**: P-R curves for our method with various iteration steps. **Right**: Weighted F-score for our method with various iteration steps. $CGVS(t_2)$ achieves stable performance with salient object detection.

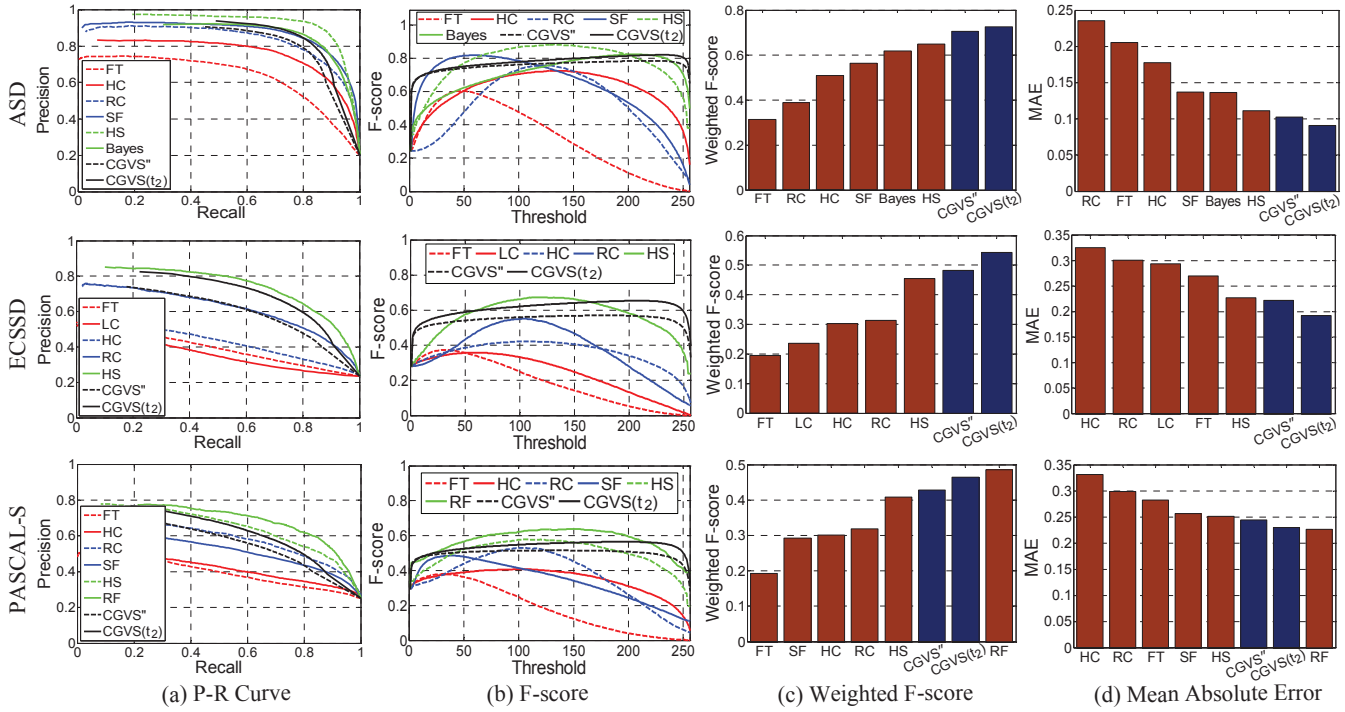


Fig. 11. Comparison of different saliency detection methods on three datasets with four metrics. (a) P-R Curve, (b) F-score, (c) Weighted F-score, and (d) Mean Absolute Error (MAE). The methods considered include LC [82], FT [6], HC [50], RC [50], SF [51], HS [8], Bayes [61]. $CGVS$" denotes the version of our method without center bias.

It is clear that, although slightly worse than that of $CGVS(t_2)$, $CGVS$" obtains acceptable performance, especially in terms of the weighted F-score and MAE. Fig. 12 shows several example results of salient object detection.

It should be noted that there are several recent methods that achieve better performance than the proposed system for the specific task of salient object detection [50], [55]. For example, the SaliencyCut method in [50] uses the detected saliency map (e.g., RC [50]) to initialize a novel iterative version of GrabCut for high quality salient object segmentation. We believe that by introducing specific post-processing similar to these methods, the performance of our model can also be improved on various metrics.

D. Between Saliency Map and Salient Object

In general, fixation prediction methods obtain poor performance when used for salient object detection. This is because almost all of the fixation prediction methods ignore the object surface and shape information and provide only a few fixation points. In this experiment, we demonstrated how to bridge the gap between the two tasks by simply transforming the saliency map to the salient object. In detail, when the saliency map is computed using certain fixation prediction methods, important information is usually distributed within or around the saliency regions. We employ the commonly used center prior to weaken the effect of saliency regions close to the image border. Then, we fit the global distribution of the saliency map with a 2-Dimension Gaussian function. The fitted result is used as the initial CBSP (Equivalent to S_w in (1)) in the non-selective pathway. Then, our system can transform the saliency map to salient objects with (5)$\sim$(9).

Fig. 13 shows two examples. We first computed the saliency map with a certain fixation prediction method (e.g., IT [5] with center prior, shown in Fig. 13(b)). The fitted Gaussian function is shown in Fig. 13(c). Then the salient object (Fig. 13(d)) is obtained by the proposed system with the Gaussian-fitted

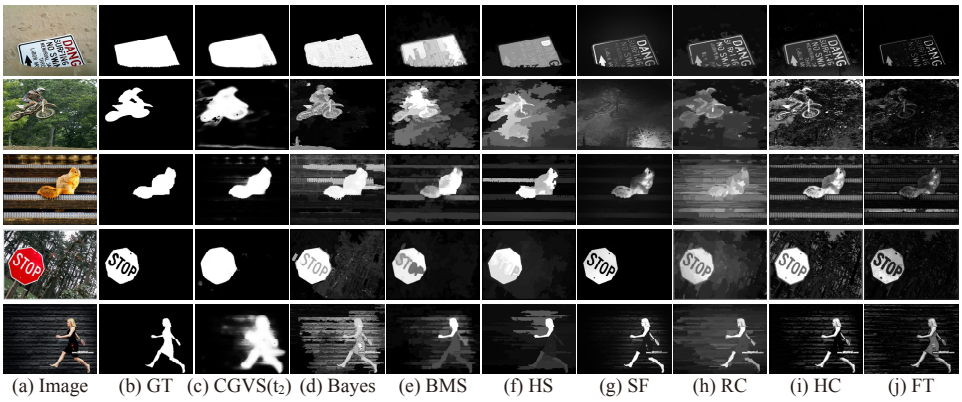


Fig. 12. Visual comparison of salient object detection. (a) original images, (b) human-marked ground truth, results of salient object detection with various methods: (c) the proposed method with two steps of iteration, (d) Bayesian saliency (Bayes) [61], (e) Boolean map (BMS) [62], (f) Hierarchical saliency (HS) [8], (g) Saliency filters (SF) [51], (h) region-based contrast (RC) [50], (i) histogram-based contrast (HC) [50], and (j) Frequency-tuned (FT) [6].

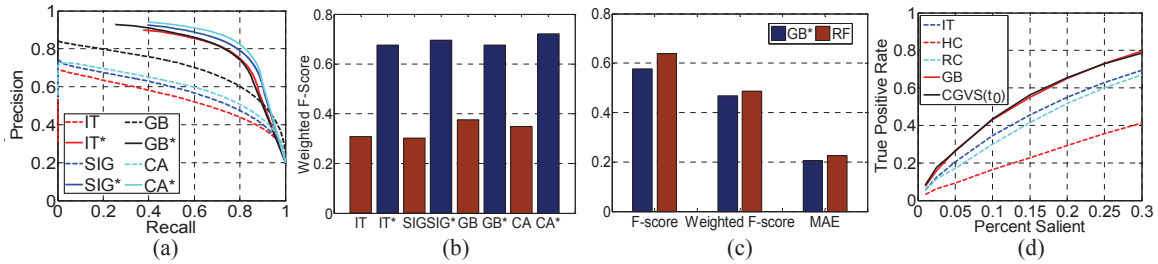


Fig. 14. Between fixation prediction and salient object detection. Evaluating the performances of several fixation prediction methods and their modified versions ($*$) for salient structure detection with the metrics of P-R Curve (a) and weighted F-score (b) on the ASD dataset [6]. (c) The comparison with RF [18], on the new PASCAL-S dataset [18]. (d) The performance of some salient object detection methods used for fixation prediction on the Judd dataset [31].

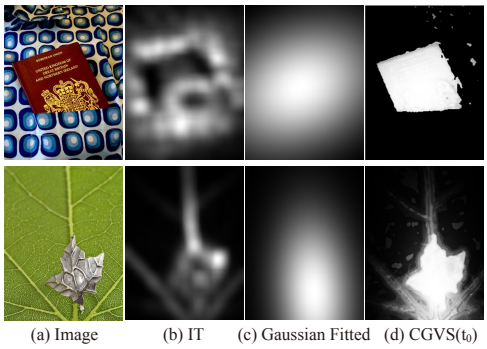


Fig. 13. The steps of transforming the saliency map to the salient object with the proposed system.

saliency map as the initial CBSP.

Fig. 14(a)-(b) show that the performances of several fixation prediction methods are significantly improved for salient object detection on both metrics of P-R Curve and weighted F-score on ASD dataset when their fixation prediction maps are inputted as the initial CBSP of our model. Fig. 14(c) shows the comparison with RF [18], which conducted a similar task (i.e., detecting salient objects based on the GB fixation prediction method) on the PASCAL-S dataset [18]. Fig. 14(c) shows that our GB^* (one version of our method that is also based on GB) is slightly worse than RF in terms of the F-score and Weighted F-score but slightly better than RF on MAE.

Fig. 14(d) shows that the models for salient object detection perform poorly at fixation prediction on the Judd dataset [31]. This is mainly because most pixels within objects detected by a salient object detection method are treated as false alarms when benchmarking on sparse human fixations.

E. Extended Saliency-Related Applications

In this experiment, we extended our CGVS system to the task of salient edge detection. To implement this task, we compute the prior edges by multiplying the CBSP in (1) with edge responses. We also replace the four low-level features with three gradients in luminance and two color-opponent channels. Thus, the final output of our system is salient edges. Fig. 15 shows two examples of salient edge detection compared to the results of [7]. Fig. 16 shows that, with iterative processing, our CGVS system can search for some specific edges such as texts in natural scenes. This example reveals the potential application of our system in the task of text detection.

F. Robustness to Parameters

In the proposed system, the observed object mask and cue weights are important for the evaluation of likelihood functions in Section III-C. Fortunately, our system is capable of automatically predicting the sizes of potential structures and the relative weight of each feature according to CBSP. Therefore, we only tested the robustness of our method to the parameters of d_r and σ_c in the range of $\{1/2, 1/3, 1/4, 1/5, 1/6\} \cdot min(W, H)$ used in the computation of CBSP in Section III-A.

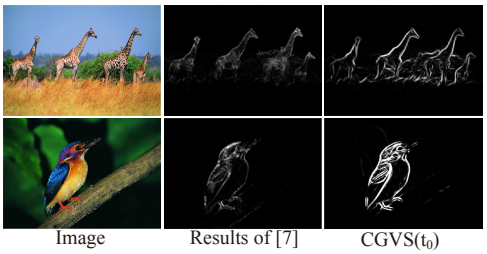


Fig. 15. Two examples of salient edge detection compared to the results of salient edges by [7].

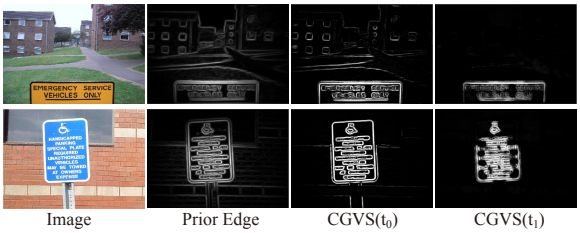


Fig. 16. The CGVS system can search for some specific edges such as texts in natural scenes with iterative processing.

Fig. 17(top) shows the F-scores and the weighted F-scores on the entire ASD dataset when varying these parameters. It can be seen that our system is very robust to these parameters. The same conclusion can be drawn from the fixation prediction experiment on Judd dataset with AUC and NSS measures (Fig. 17(bottom)). Taken together, Fig. 17 indicates that our method would not benefit much from learning an optimal parameter setting, and manual parameter selection is enough to robustly obtain quite acceptable performance.

Fig. 18 shows that setting the threshold between 10% to 50% (Section III-C) may affect the final salient structure detection. For example, for smaller objects (<10%), high thresholds (>10%) lead to many background pixels being incorrectly divided into object sets (Fig. 18(top)). In contrast, for larger objects, a low threshold may cause the algorithm to respond only to a partial region of the object (Fig. 18(bottom)). Therefore, a more robust strategy for object size selection is expected in the future to adapt to various complex scenes.

V. Discussion

In this paper, we proposed a contour-guided visual search (CGVS) system inspired by the guided search theory (GST) of the biological vision. Different from the classical FIT theory [11] and the popular model proposed by Itti *et al.* [5], our method searches for salient structures (SSs) with Bayesian inference guided by contour information, such as the location and size of SS, importance of features, etc. Although many recent models attempt to employ various image segmentation algorithms as a pre-processing step to divide the images into superpixels and obtain high quality salient object segmentation (e.g., [50], [55]), our model highlights the dominant objects with accurate shapes from relatively simple scenes with as few as two or three steps of iteration, without the requirement of segmenting input image into superpixels.

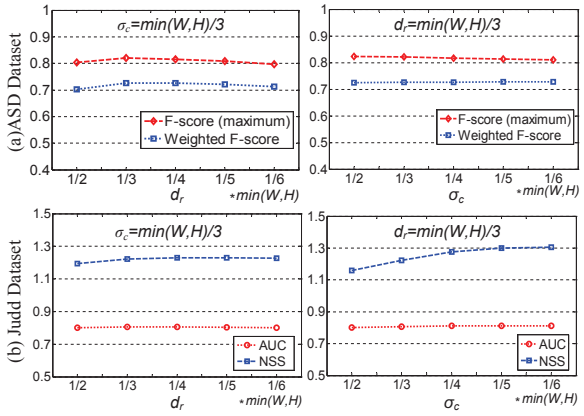


Fig. 17. Robustness to parameters on salient object dataset (ASD) and fixation prediction dataset (Judd). **Top**: Testing our $CGVS(t_2)$ with $\sigma_c = min(W,H)/3$ and various d_r (left), and with $d_r = min(W,H)/3$ and various σ_c (right). **Bottom**: similar testing of $CGVS(t_0)$ on the Judd dataset.

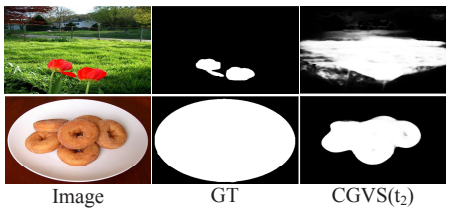


Fig. 18. Examples of detecting quite small or large objects. **Top**: incorrect estimation for a small object. **Bottom**: response to part of a larger object.

The flexibility and robustness of the proposed system come from the following novel strategies, in comparison to other Bayesian inference based saliency detection models: (1) A simple yet efficient edge-based filling-in operation is implemented to create the robust global gist used for top-down control. This edge-based filling-in can always capture the interesting regions where an unambiguous object of interest is contained. (2) Simple yet robust methods are designed to automatically and robustly define the sizes of potential salient structures and estimate the weight of each bottom-up feature, which enables the salient objects to have clear shapes and filled surfaces or highlights the interesting regions (without clear object) within a filled but limited area.

Extracting contextual (or coarse) information from scenes rapidly and then guiding the processing of fine information is an efficient strategy (i.e., coarse-to-fine) for visual perception and scene analysis without involving specific tasks. From this standpoint, the proposed model can be regarded as a general framework for guided search, and this system can be easily extended, e.g., by introducing high-level object-related global features in the non-selective pathway for a specific object search task. Meanwhile, adding more local features (e.g., depth, motion, etc.) in the selective pathway may also extend our system to more applications such as video processing. In addition, some unsupervised learning methods [85] can be employed for rapid scene analysis and finding the structured context in the non-selective pathway.

It is clear that the proposed approach is purely image signal driven. In particular, the dominant edges in the non-

selective pathway of this model act as a type of non-selective structural information to provide the "where" prior of potential objects, whereas the detailed object features (i.e., the "what" information) are extracted in the selective pathway. However, it seems difficult to closely relate our two-pathway based approach to the widely recognised work by Ungerleider on ventral and dorsal pathways [86]. As indicated by Wolfe *et al*. [1], more studies are still required before the 'selective' and 'non-selective' pathways can be properly related to the 'what' and 'where' pathways in terms of neurophysiology, though the corresponding pathways in these two theories seem to be similar in terms of specific functional roles.

To conclude, the result of this work is a single computationally efficient system that provides dual use. When given a cluttered scene without any dominant object, the proposed system will work as a fixation prediction model. Alternatively, when given a simple scene, the proposed system will determine the dominant objects contained in the scene.

REFERENCES

[1] J. M. Wolfe, M. L.-H. Võ, K. K. Evans, and M. R. Greene, "Visual search in scenes involves selective and nonselective pathways," *Trends in Cognitive Sciences*, vol. 15, no. 2, pp. 77–84, 2011.

[2] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, 2012.

[3] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The jpeg2000 still image coding system: an overview," *IEEE Trans. Consumer Electronics*, vol. 46, no. 4, pp. 1103–1127, 2000.

[4] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE CVPR*, vol. 2, 2004, pp. II–37–II–44.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 11, pp. 1254–1259, 1998.

[6] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, 2009, pp. 1597–1604.

[7] M. Holtzman-Gazit, L. Zelnik-Manor, and I. Yavneh, "Salient edges: A multi scale approach," in *Proc. ECCV Workshop*, 2010.

[8] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE CVPR*, 2013, pp. 1155–1162.

[9] M. V. Peelen and S. Kastner, "Attention in the real world: toward understanding its neural basis," *Trends in Cognitive Sciences*, vol. 18, no. 5, pp. 242–250, 2014.

[10] L. Elazary and L. Itti, "A bayesian model for efficient visual search and recognition," *Vision Research*, vol. 50, no. 14, pp. 1338–1352, 2010.

[11] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[12] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*, 1987, pp. 115–141.

[13] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[14] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Ann. Rev. of Neurosci.*, vol. 18, no. 1, pp. 193–222, 1995.

[15] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, 2013.

[16] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *arXiv preprint arXiv:1411.5878*, 2014.

[17] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[18] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE CVPR*, 2014, pp. 280–287.

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

[20] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2006, pp. 545–552.

[21] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. NIPS*, 2005, pp. 155–162.

[22] 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, 2012.

[23] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and distinctiveness provide with human-like saliency," in *Advanced Concepts for Intelligent Vision Systems*, 2009, pp. 343–354.

[24] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE CVPR*, 2012, pp. 478–485.

[25] Y. Lin, B. Fang, and Y. Tang, "A computational model for saliency maps by using local entropy." in *Proc. AAAI*, 2010.

[26] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth mover's distance-based saliency measurement and nonlinear feature combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 314–328, 2013.

[27] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[28] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. IEEE CVPR*, 2008, pp. 1–8.

[29] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, 2010.

[30] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, 2013.

[31] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE ICCV*, 2009, pp. 2106–2113.

[32] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE CVPR*, 2012, pp. 438–445.

[33] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, pp. 1–26, 2008.

[34] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of Vision*, vol. 8, no. 3, pp. 1–15, 2008.

[35] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. NIPS*, 2008, pp. 241–248.

[36] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. ECCV*, 2010, pp. 30–43.

[37] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. ICIP*, vol. 1, 2003, pp. I–253.

[38] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[39] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. IEEE CVPR*, vol. 2, 2006, pp. 2049–2056.

[40] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. IEEE CVPR*, 2007, pp. 1–8.

[41] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, pp. 1–15, 2011.

[42] W. Kienzle, F. A. Wichmann, M. O. Franz, and B. Schölkopf, "A nonparametric approach to bottom-up visual saliency," in *Proc. NIPS*, 2006, pp. 689–696.

[43] C.-C. Wu, F. A. Wick, and M. Pomplun, "Guidance of visual attention by semantic information in real-world scenes," *Frontiers in Psychology*, vol. 5, no. 54, 2014.

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

[45] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Amer. A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[46] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[47] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Proc. NIPS*, 2005, pp. 547–554.

[48] S. Lu, C. S. Tan, and J.-H. Lim, "Robust and efficient saliency modeling from image co-occurrence histograms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 195–201, 2014.

[49] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011.

[50] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.
[51] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE CVPR*, 2012, pp. 733–740.
[52] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE CVPR*, 2013, pp. 1139–1146.
[53] D. Klein, S. Frintrop *et al.*, "Center-surround divergence of feature statistics for salient object detection," in *Proc. IEEE ICCV*, 2011, pp. 2214–2219.
[54] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, 2012, pp. 29–42.
[55] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE CVPR*, 2014, pp. 2814–2821.
[56] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, 2012.
[57] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 222–234, 2014.
[58] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE CVPR*, 2014, pp. 3286–3293.
[59] R. P. Rao, "Bayesian inference and attentional modulation in the visual cortex," *Neuroreport*, vol. 16, no. 16, pp. 1843–1848, 2005.
[60] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A bayesian inference theory of attention," *Vision Research*, vol. 50, no. 22, pp. 2233–2247, 2010.
[61] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, 2013.
[62] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. IEEE ICCV*, 2013, pp. 153–160.
[63] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.
[64] L. Chen, "The topological approach to perceptual organization," *Visual Cognition*, vol. 12, no. 4, pp. 553–637, 2005.
[65] M. Chen, Y. Yan, X. Gong, C. D. Gilbert, H. Liang, and W. Li, "Incremental integration of global contours through interplay between visual cortical areas," *Neuron*, vol. 82, no. 3, pp. 682–694, 2014.
[66] P. L. Rosin, "A simple method for detecting salient regions," *Pattern Recognition*, vol. 42, no. 11, pp. 2363–2371, 2009.
[67] S. Anstis, "Visual filling-in," *Current Biology*, vol. 20, no. 16, pp. R664–R666, 2010.
[68] H. Neumann, L. Pessoa, and T. Hansen, "Visual filling-in for computing perceptual surface properties," *Biological Cybernetics*, vol. 85, no. 5, pp. 355–369, 2001.
[69] S. Grossberg, E. Mingolla, and J. Williamson, "Synthetic aperture radar processing by a multiple scale neural system for boundary and surface representation," *Neural Networks*, vol. 8, no. 7, pp. 1005–1028, 1995.
[70] K. Yang, S. Gao, C. Li, and Y. Li, "Efficient color boundary detection with color-opponent mechanisms," in *Proc. IEEE CVPR*, 2013, pp. 2810–2817.
[71] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 1–17, 2007.
[72] V. Mante, R. A. Frazor, V. Bonin, W. S. Geisler, and M. Carandini, "Independence of luminance and contrast in natural scenes and in the early visual system," *Nat. Neurosci.*, vol. 8, no. 12, pp. 1690–1697, 2005.
[73] R. A. Frazor and W. S. Geisler, "Local luminance and contrast in natural images," *Vision Research*, vol. 46, no. 10, pp. 1585–1598, 2006.
[74] J. T. Lindgren, J. Hurri, and A. Hyvärinen, "Spatial dependencies between local luminance and contrast in natural images," *Journal of Vision*, vol. 8, no. 12, p. 6, 2008.
[75] E. Switkes and M. A. Crognale, "Comparison of color and luminance contrast: apples versus oranges?" *Vision Research*, vol. 39, no. 10, pp. 1823–1831, 1999.
[76] S. Clery, M. Bloj, and J. M. Harris, "Interactions between luminance and color signals: Effects on shape," *Journal of Vision*, vol. 13, no. 5, pp. 16–16, 2013.
[77] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.
[78] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. ECCV*, 2010, pp. 366–379.
[79] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proc. CVPR*, 2015, pp. 1072–1080.
[80] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.
[81] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proc. IEEE ICCV*, 2013, pp. 921–928.
[82] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM MM*, 2006, pp. 815–824.
[83] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE CVPR*, 2014, pp. 248–255.
[84] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE ICCV*, 2013, pp. 1529–1536.
[85] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in *Proc. CVPR*, 2013, pp. 430–437.
[86] L. G. Ungerleider and J. V. Haxby, "whatand wherein the human brain," *Current Opinion in Neurobiology*, vol. 4, no. 2, pp. 157–165, 1994.

Kai-Fu Yang received the B.Sc. and M.Sc. degrees in biomedical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009 and 2012, respectively, where he is currently pursuing the Ph.D. degree. His research interests include visual mechanism modeling and image processing.

Hui Li received a B.Sc. degree in Biomedical engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2013, where she is currently pursuing the M.Sc. degree. Her research interests include visual mechanism modeling and image processing.

Chao-Yi Li received the degrees from Chinese Medical University, Shenyang, China, in 1956, and Fudan University, Shanghai, China, in 1961. He is currently a Professor with the University of Electronic Science and Technology of China, Chengdu, China, and the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. His research interest is mainly in visual neurophysiology.

Yong-Jie Li (M'14) received the Ph.D. degree in biomedical engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2004. He is currently a Professor with the Key Laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, UESTC. His research interests include visual mechanism modeling, image processing, and intelligent computation.