Neurocomputing 103 (2013) 63–74

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Visual saliency detection with center shift

Weibin Yang [a,*], Yuan Yan Tang [a,b], Bin Fang [a], Zhaowei Shang [a], Yuewei Lin [c]

[a] *School of Computer Science, Chongqing University, Chongqing 400030, China*
[b] *Department of Computer and Information Science, University of Macau 999078, Macau*
[c] *Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA*

ARTICLE INFO

Article history:
Received 3 January 2012
Received in revised form
28 July 2012
Accepted 1 August 2012
Communicated by Pingkun Yan
Available online 17 October 2012

Keywords:
Visual saliency
Probabilistic model
Center shift
Multiscale analysis
Salient object
Human fixation

ABSTRACT

This paper proposes a novel method for visual saliency detection based on an universal probabilistic model, which measures the saliency by combining low level features and location prior. We view the task of estimating visual saliency as searching the most conspicuous parts in an image and extract the saliency map by computing the dissimilarity between different regions. We simulate the moving of the center of human visual field, and describe how the center shift process works on visual saliency. Furthermore, multiscale analysis is adopted for improving the robustness of our model. Experimental results on three public image datasets show that the proposed approach outperforms 18 state-of-the-art methods for both salient object detection and human eye fixation prediction.

1. Introduction

Human visual system has a remarkable ability to pay more attention to some conspicuous regions or objects in natural complex scenes, due to the fact that it cannot fully process the tremendous amount of input visual information [1]. This ability, viewed as visual attention, plays an essential role in extracting and analyzing the important visual stimuli in many engineering applications. In order to find out an analogous model mimicking human selective attention mechanism, researchers in physiology, psychology and computer vision have been making a good effort for a long time and proposed many computational models. Generally, a saliency map is used for illustrating salient regions with higher values, in contrast to background regions with lower values. As a consequence, the saliency maps calculated by various models are widely used in many computer vision and pattern recognition applications, such as image segmentation [2], object detection [3], object recognition [4,5], image or video quality assessment [6,7], image fusion [8], video compression [9] and object tracking [10].

Over the past decade, many different methods have been proposed to estimate visual saliency, which typically transform a given input image into a scalar-valued map [11]. They can be broadly classified as biologically inspired [12], purely computational [13,14], and a combination of both [15]; as frequency domain based [16], spatial domain based [17], and both considered [18]; as for static images [19], for dynamic video [20] and for both [21,22]; and as parametric [23] and non-parametric [24]. Recently, a lot of saliency models tried to extract saliency maps based on mathematical or statistical principles that address the purpose of the computation [14,25–28]. These principles come from information theory, mathematical tools, and statistical analysis, and describe consistent properties as the definition from neural and psychophysical experiments. Furthermore, by defining visual saliency at each location as the dissimilarity between itself and its local neighborhood or global counterparts, many state-of-the-art models estimate visual saliency in terms of block or region [17,29–36]. These methods are efficient because they present the dissimilarity of a block or region by various meaningful weighted distances, which may more accord with the human visual system. Generally speaking, an actual region with similar features can attract more attention than a single pixel, and we are more likely to notice a large region than a small one with the same conspicuity, which is similar to the "large scale bias" presented in [15].

The work proposed in [12,37,38] suggested that human visual system scans the scene both in a rapid, bottom-up, saliency-driven, and task-independent manner as well as in a slower, top-down, volition-controlled, and task-dependent manner. Most research papers estimate visual saliency in a bottom-up way, which usually utilizes different low level visual features. The most

[*] Corresponding author.
E-mail address: wbyang.cn@gmail.com (W. Yang).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.08.029

influential computational framework for estimating visual saliency was proposed by Itti et al. [12], which implemented and further developed the physiologically inspired saliency-based model of visual attention introduced by Koch and Ullman [37]. Under the hypothesis that visual attention is attracted by various local features, Itti et al. proposed a unified framework with three steps: feature extraction, contrast computation and map combination. Different features like color, intensity, and orientation at different scales were first extracted respectively, and a single conspicuous map was then formed by applying the center-surround operation across scales to compute the contrast value. Finally, conspicuous maps over different scales in different feature space were summed to obtain the master map. Many following saliency models used the same or similar architecture [13,15,39,40]. Walther et al. [39] further developed Itti's model, and proposed a biologically plausible model of forming and attending to proto-objects in natural scenes. Harel et al. [13] used the same features as in Itti's model, built fully connected graph over all locations and assigned different weights between nodes to compute saliency maps in each feature map over different scales. Valenti et al. [40] directly viewed the extracted curvature, isocenters and color edges maps over scales as saliency maps and then combined them linearly. The model of Lin et al. [15] adopted a similar framework as Itti's model, whereas it added a local entropy feature map and measured the center-surround difference using the Earth Mover's Distance (EMD) based on weighted histogram. Aziz and Mertsching [29] categorized the feature computation methods in the attention models into three classes: pixel-based, frequency domain and region-based approaches, and generated the saliency map by incorporating five feature channels using the rarity criteria. Based on the stochastic model and some observations, Avraham and Lindenbaum [30] used a graphical model approximation to evaluate which regions are more likely to be salient bystarting with a rough pre-attentive segmentation. Recently, Wang et al. [41] incorporated near-infrared clues into the detection framework to form the multi-spectral saliency detection method, and Perazzi et al. [42] estimated the complete contrast and saliency using high dimensional Gaussian filters.

In our approach, we view an image as a set of regions, and obtain visual saliency in a unified probabilistic model. We estimate visual saliency in a bottom-up, task-independent way similar to most popular methods, so our model starts from the simple assumption that visual saliency is closely related to features and locations, and builds up a probabilistic model to indicate these two key factors. We assume that these two elements are independent to each other, and construct two functions represent their conditional probabilities respectively. Moreover, based on the central bias introduced by [43], we propose a center shift process which mimics the moving of human visual field. In addition, multiscale analysis is given for improving the performance on searching the saliency at different scales. As a result, our algorithm is not only more robust than other saliency models, but also more biologically plausible.

Generally, ground truths are necessary for a unified comparison and analysis to evaluate the performance of various visual saliency models. Concluding from ground truths given by some popular datasets and the applications of current various saliency models, we categorize the basic tasks into two groups, salient object detection and human fixation prediction. Some visible examples are illustrated in Fig. 1. For salient object detection, which prefers finding out important, conspicuous and even meaningful regions in the complex natural scene, the proposed model needs to emphasize more on the whole regions of salient objects than on some isolated pixels. Meanwhile, for human fixation prediction, one may lay particular stress on highlighting any conspicuous clutter, even when no obvious object exists in the scene. Theoretically, an ideal saliency model can complete both tasks perfectly. However, most methods do not have satisfactory effectiveness on both tasks, while our proposed model performs better than 18 state-of-the-art methods on both tasks, because we view regions as prime elements and consider a center shift process. More details are discussed in Section 3.

The rest of this paper is organized as follows. Section 2 introduces the proposed framework for estimating visual saliency. Section 3 evaluates the proposed approach comparing with 18 sate-of-the-art methods in both salient object detection and human fixation prediction. Conclusions are given in Section 4.

2. The proposed approach

In this section, we describe the details of the proposed approach. First, we present an universal probabilistic model for visual saliency estimation, which, in our method, is composed of feature-based visual saliency and location-based visual saliency. After that, we simulate a shifting process of the center of the visual field, which is called center shift, and then present the multiscale analysis.

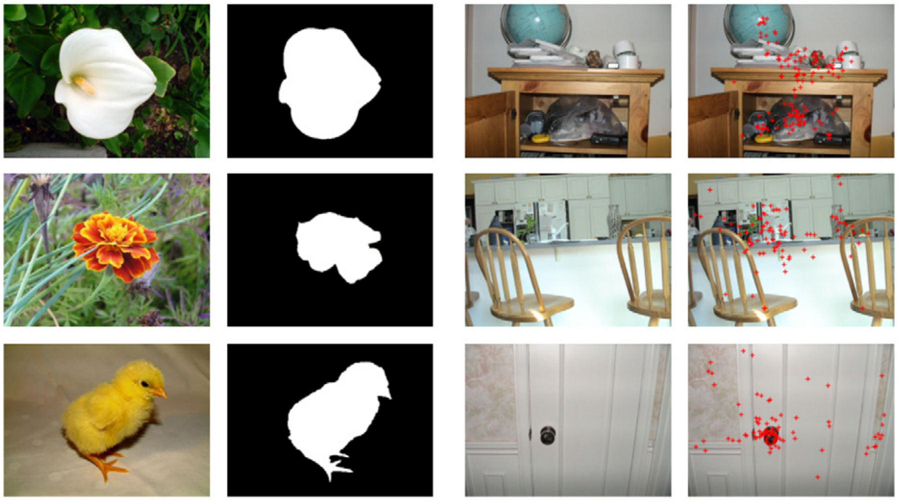

Fig. 1. Different sample images and ground truths for different tasks. (a) Ground truths for salient object detection [14]. The masks (right) are ground truths for the corresponding images (left). (b) Ground truths for human fixation prediction [25]. The red crosses (right) are ground truths for the corresponding images (left). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

2.1. The probabilistic model

Human visual system always instinctively searches the most conspicuous parts in the visual field, which is achieved by estimating the importance of regions at every location. We propose that a probability of a region, which represents its importance, is the visual saliency.

We consider an image as a set of regions R, $\{r_1,r_2,\ldots,r_N\}$, where N is the total number of regions, segmented by a graph-based image segmentation algorithm [44], which is a widely used method also used in [34,45]. For each region r_i, a saliency value S_{r_i} is estimated, and all of the values corresponding to different regions compose the whole image saliency map. Suppose that a random variable V_{r_i} denotes that region r_i is salient, and B_{r_i} is a set of factors which results in the visual saliency value S_{r_i}. Furthermore, S_{r_i} could be a conditional probability of V_{r_i} given B_{r_i}, that is, $S_{r_i}=P(V_{r_i}|B_{r_i})$. It is an universal model to estimate visual saliency. Usually, the two involving key problems are what factors influence the saliency value and how they work. Many biological and mathematical methods have launched thorough researches. In our method, we conclude from various existing saliency models and experimental results, and then consider two important factors, feature and location. We let a random variable F_{r_i} denote visual features observed at region r_i, and let a random variable L_{r_i} denote the location of region r_i. Then the saliency value of region r_i, S_{r_i}, is defined as

$$S_{r_i}=P(V_{r_i}|F_{r_i},L_{r_i}) \tag{1}$$

For the same consideration of simplicity as in [46], we assume that features and locations are independent, and then the definition can be rewritten as

$$S_{r_i}=P(V_{r_i}|F_{r_i})P(V_{r_i}|L_{r_i}) \tag{2}$$

By formalizing the saliency as Eq. (2), we suggest that the saliency value of any location in an image depends on two factors involving features and locations. As long as the probabilistic model may express the relationship between the two factors and real visual saliency, or describe the same monotonicity, the model is reasonable and efficient. To create such a model, we need to analyze how these two different factors act on the saliency.

The first term of Eq. (2), $P(V_{r_i}|F_{r_i})$, represents a conditional probability of possible saliency value given observed features F_{r_i}. Obviously, it mainly depends on what features we choose. The chosen features may be discriminative enough to distinguish the conspicuous parts from their surroundings, which is consistent with the definition of human visual conspicuity in [46]. In our method, according to this principle, we utilize the dissimilarity, measured by the weighted color contrast and spatial distance among regions, to form the description of low level features.

The second part, $P(V_{r_i}|L_{r_i})$, describes the relevance between the location and the saliency, which is called location prior [46]. It is independent of visual features and reflects the locations that the observer probably notice without any influence of image content. We adopt the central bias theory [43] to show that observers pay more attention to the regions closer to the center of the visual field.

Let $\Phi(F_{r_i})$ and $\Phi(L_{r_i})$ denote the two terms $P(V_{r_i}|F_{r_i})$ and $P(V_{r_i}|L_{r_i})$ respectively, so the Eq. (2) may be rewritten as

$$S_{r_i}=\Phi(F_{r_i})\Phi(L_{r_i}) \tag{3}$$

To generate the probabilistic model, we construct the two functions $\Phi(F_{r_i})$ and $\Phi(L_{r_i})$ respectively based on their corresponding implications. The details are discussed below.

2.1.1. Feature-based visual saliency

It is conventional to compute the saliency map using various low level features, such as color, intensity and orientation [12], and the ultimate goal is to measure the similarity or dissimilarity between different regions or pixels by local or global ways. As aforementioned, given a region r_i, we estimate the dissimilarity contrast to all the other regions by both the color and spatial distance,

$$\Phi(F_{r_i})=\frac{\sum_{j=1}^{N}\omega(r_i)D_c(r_i,r_j)D_s(r_i,r_j)}{\max\sum_{j=1}^{N}\omega(r_i)D_c(r_i,r_j)D_s(r_i,r_j)} \tag{4}$$

where N is the total number of regions, $\omega(r_i)$ is the weight of region r_i, $D_c(r_i,r_j)$ is the color dissimilarity between regions r_i and r_j, and $D_s(r_i,r_j)$ is the spatial distance between regions r_i and r_j. The denominator is for the purpose of normalization. We set $\omega(r_i)$ to be the number of pixels in region r_i to emphasize contrast to bigger regions, which follows the principle that human pays more attention to big salient objects than small ones [15].

We utilize a popular method [33,34] to measure the color contrast and the spatial distance. The definition of the color contrast $D_c(r_i,r_j)$ is as follows:

$$D_c(r_i,r_j)=\sum_{k=1}^{N_i}\sum_{l=1}^{N_j}p_{c_{i,k}}p_{c_{j,l}}\|c_{i,k}-c_{j,l}\| \tag{5}$$

where N_i and N_j is the number of colors in region r_i and r_j respectively, $c_{i,k}$ is the k-th color in region r_i, $c_{j,l}$ is the l-th color in region r_j, $p_{c_{i,k}}$ is the probability of the color $c_{i,k}$ among all N_i colors in region r_i, $p_{c_{j,l}}$ is the probability of the color $c_{j,l}$ among all N_j colors in region r_j, and $\|c_{i,k}-c_{j,l}\|$ is the Euclidean distance between the color $c_{i,k}$ and $c_{j,l}$ in CIE Lab color space. To reduce the computational complexity, a quantization operation is necessary [34]. Each channel of the RGB color space is first quantized to 12 different values, and then frequently occurring colors which cover more than 95% of the image pixels are chosen. After that, we transform RGB color space into Lab color space for further computation according to Eq. (5).

The weighted spatial distance $D_s(r_i,r_j)$ is defined as

$$D_s(r_i,r_j)=\exp\left(-\frac{\|C_{r_i}-C_{r_j}\|}{\delta}\right) \tag{6}$$

where C_{r_i} and C_{r_j} is the center of region r_i and r_j respectively, $\|C_{r_i}-C_{r_j}\|$ is the Euclidean distance between two centers, and δ controls the strength of spatial distance which is set to be 0.4 as in [33,34].

2.1.2. Location-based visual saliency

The central bias [33,43,47] proves that observers pay most attention to the center. To demonstrate the location prior and emphasize more on the center of the visual field, we employ a two dimensional anisotropic Gaussian function to describe how locations conduct the conspicuity,

$$\Phi(L_{r_i})=\exp\left\{-\left(\frac{(x_c-x_0)^2}{2\sigma_x^2}+\frac{(y_c-y_0)^2}{2\sigma_y^2}\right)\right\} \tag{7}$$

where (x_c,y_c) is the center of region r_i, (x_0,y_0) is the center of the visual field, and σ_x^2 and σ_y^2 are variances along the two directions respectively. In all of our experiments, σ_x^2 is set to $0.5W_{im}$ and σ_y^2 is set to $0.5H_{im}$, where W_{im} and H_{im} are the width and height of the image respectively. As expressed in Eq. (7), we give a formal description to draw the relationship between location and saliency, which involves only a center and two parameters σ_x^2 and σ_y^2. We introduce a center shift process to describe the effect of the moving of the center in Section 2.2.

2.2. Center shift

As argued in [43], when searching the scenes for a conspicuous target, fixation distributions are shifted from the image center to the distributions of image features. We simulate the shifting process of human fixations, and draw the fact that the center of the visual field plays an important role. We call this process center shift. Suppose that we look for salient regions in an arbitrary visible scene. We throw a glance first and naturally fix our attention more on the center of the scene, and then our attention is attracted by some distinctive image features (color, luminance, texture, and so on) rapidly, so the center of visual field shifts toward the conspicuous regions. For example, as illustrated in Fig. 2, when we take a look at this image unconsciously, we first fix our eyes on the initial center habitually. However, since the distinct black part appears on the right side, we quickly transfer our attention to the black region, so the center of this black region becomes the new center of the visual field. From Fig. 2(b) we can see that the shifted center is more meaningful and effective while calculating the location-based visual saliency by Eq. (7).

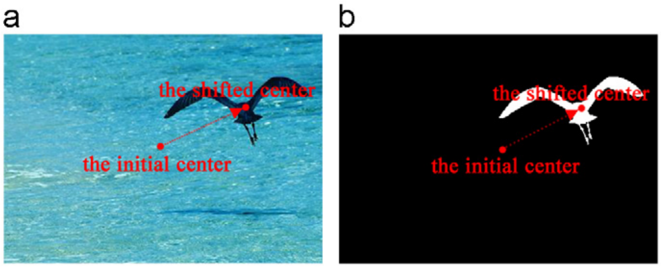

Fig. 2. An illustration of the center shift process. (a) The original image. (b) The ground truth.

Therefore, the initial center of the visual field is fixed on the center of the input image,

$$(x_0^i, y_0^i) = \left(\frac{W_{im}}{2}, \frac{H_{im}}{2}\right) \tag{8}$$

where W_{im} and H_{im} are the width and height of the image respectively. Then the initial saliency map S^i is obtained using the initial center (x_0^i, y_0^i). We consider the intensity centroid of the initial saliency map as the shifted center, which is also widely used in other applications, such as invariant keypoint detector [48]. The

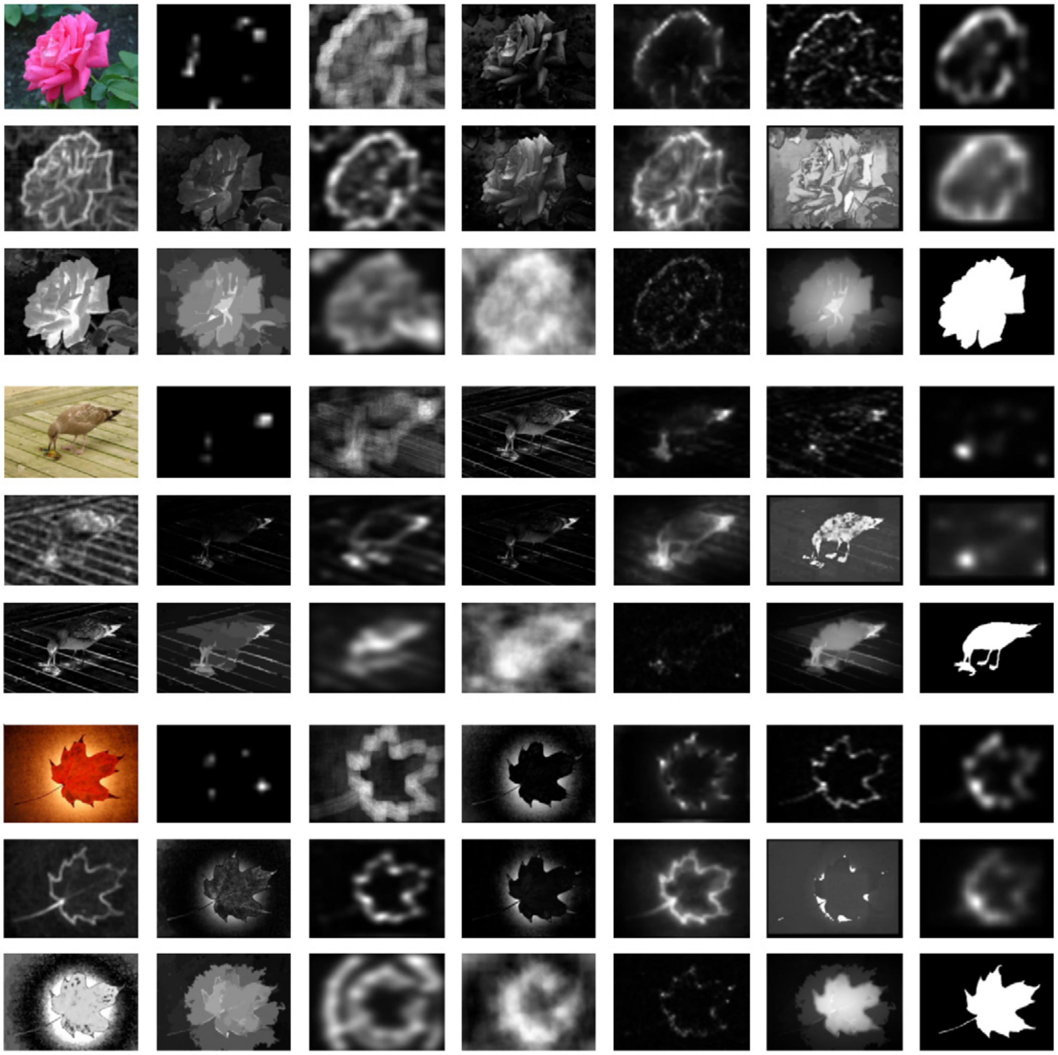

Fig. 3. Comparison of saliency estimation results for various methods with our approach on the MSRA dataset. Images are shown in the order of the original image, results on Itti98 [12], Bruce06 [25], Zhai06 [54], Harel07 [13], Hou07 [16], Hou08 [21], Zhang08 [46], Achanta09 [14], Seo09 [55], Achanta10 [50], Goferman10 [32], Rahtu10 [31], Wang10 [27], Cheng11HC [34], Cheng11RC [34], Li11 [18], Murray11 [28], Hou12 [51] and the proposed model, and the ground truth at the end.

intensity centroid is measured by moments as follows:

$$(x_0^s, y_0^s) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{9}$$

where m_{00}, m_{10} and m_{01} are defined uniformly by

$$m_{pq} = \sum_{x,y} x^p y^q S^i(x,y) \tag{10}$$

where $S^i(x,y)$ is the pixel value at (x,y) in the initial saliency map S^i. We compute S^i at the small scale, which is mentioned in Section 2.3. For simplicity and efficiency, we assume that the shifted center plays the same role as the original center in location-based visual saliency, so we keep the same parameter settings when calculating visual saliency based on the shifted center.

2.3. Multiscale analysis

Multiscale analysis is conventional and useful for estimating visual saliency, and it is widely used in many literatures [12,15,22,32]. Human visual system can adaptively capture a salient object in the visual field no matter what size the object is. In addition, attention on small scale images focuses on a whole object with same features, while attention on large scale images cares more about the local details. Particularly, estimating visual saliency at smaller scale makes an integrate object more conspicuous because it ignores some very small regions with inconsistent features by segmenting the downsampling image. Hence we extract visual saliency over different scales.

For simplicity and efficiency, we consider two different scales with scale factors $\sigma_l = 1$ and $\sigma_s = 0.2$, which means the same operations are implemented on both the original image and 1/5 size (both the height and the width) of the original image. First, the initial saliency map S^i based on the initial center of the visual field (x_0^i, y_0^i) is estimated at the small scale, and then the shifted center is computed by Eqs. (9) and (10). Finally, we measure two different saliency maps at two different scales.

It is worth noting that we only need to recompute the location prior $\Phi(L_{r_i})$ using the shifted center (x_0^s, y_0^s) when we reestimate the saliency map at the small scale, because the other part $\Phi(F_{r_i})$ may be saved from obtaining the initial saliency map S^i. Therefore, in the actual implementation, the process of calculating the shifted center increases no more computation than working out the location prior.

The ultimate saliency map, also called the master saliency map, is obtained by

$$S = \alpha S_l + (1-\alpha) S_s \tag{11}$$

where α is a weight parameter and we set it to be 0.5 in our experiments, S_l and S_s are the saliency maps calculated at both the large and the small scales.

Table 1
AUC scores of various saliency models on different datasets.

Saliency model	MSRA	YORK	MIT
Itti98 [12]	0.6301	0.5755	0.5586
Bruce06 [25]	0.6994	0.7003	0.6906
Zhai06 [54]	0.7683	0.5553	0.5463
Harel07 [13]	0.8545	0.8078	0.8146
Hou07 [16]	0.7234	0.6769	0.6668
Hou08 [21]	0.8509	0.7764	0.7595
Zhang08 [46]	0.7457	0.6653	0.6676
Achanta09 [14]	0.8631	0.5493	0.5427
Seo09 [55]	0.7937	0.7364	0.7041
Achanta10 [50]	0.8934	0.6809	0.6770
Goferman10 [32]	0.8772	0.7795	0.7586
Rahtu10 [31]	0.9324	0.7249	0.7373
Wang10 [27]	0.8944	0.7938	0.7835
Cheng11HC [34]	0.9177	0.5841	0.5772
Cheng11RC [34]	0.9636	0.7512	0.7573
Li11 [18]	0.8250	0.7128	0.7243
Murray11 [28]	0.7904	0.7470	0.7058
Hou12 [51]	0.7599	0.7021	0.6794
Proposed	**0.9744**	**0.8173**	**0.8243**

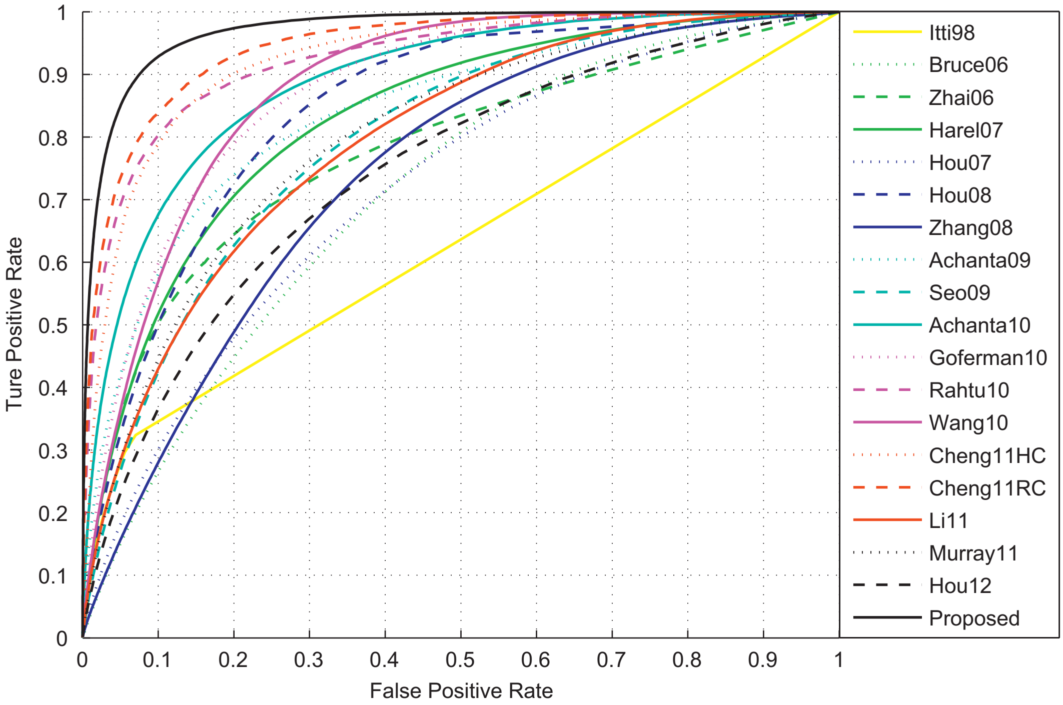

Fig. 4. ROC curves of various saliency models on the MSRA dataset.

3. Experimental results

In this section, we evaluate our approach on three public image datasets and compare with other 18 state-of-the-art saliency models. First, we introduce the public available datasets used in our experiments. Then, by using commonly used validation approaches, we give the performance of the proposed method and various saliency models for reference on two different tasks, i.e. salient object detection and human fixation prediction. Finally, the discussion is given.

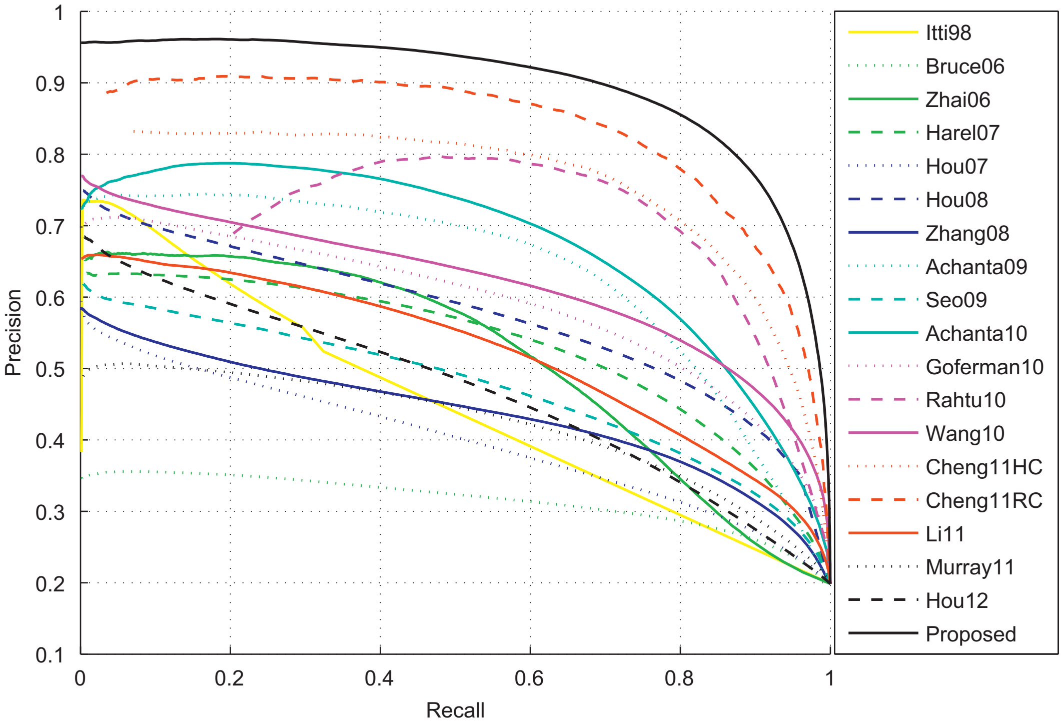

Fig. 5. Precision–recall curves of various saliency models on the MSRA dataset.

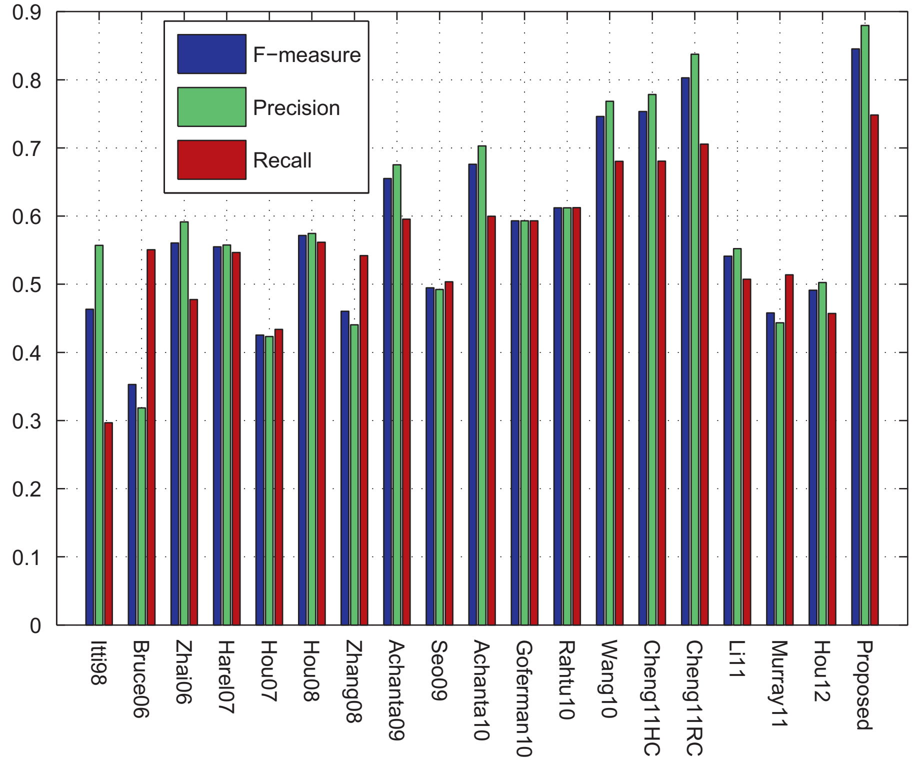

Fig. 6. Precision–recall bars for binarization of saliency maps on the MSRA dataset.

3.1. Image datasets

We apply our method on three public available image datasets to evaluate its performance, and divide the experiments into two parts based on two different research purposes.

The first dataset, named as MSRA [14], contains 1000 color images with accurate pixel-wise object-contour segmentations, which is selected from a 5000 images dataset [19] with rectangular segmented object annotations by nine observers. These images are collected mostly from image forums and image search engines, and each image contains at least one salient object or one distinctive foreground object in simple or complex scenes. These salient objects differ in category, color, shape, size, and so on. In other words, there is no more prior knowledge or constraint on these objects except that they are the most salient. For the reason that a bounding box-based ground truth is far from accurate [49], Achanta et al. [14] took a subset of 1000 images, and recommended to use binary masks which exactly described salient regions. Henceforth, many saliency models for detecting or segmenting salient object evaluate their performance on this image database [31,34,35,50].

The second one, called YORK, is introduced in [25]. There are 120 images including indoor and outdoor scenes in the dataset, some with very salient items, others with no particular regions of interest. Eye fixations of 20 subjects are recorded for each image. These subjects are positioned 0.75 m from a 21 CRT monitor and given no particular instructions except to observe the images. All the image sizes are 681×511 pixels. This dataset is widely used for predicting and tracking human eye fixation [13,21,46,51].

The last image database, denoted as MIT, is proposed by [47] for predicting where humans look. There are 1003 natural images containing different scenes and objects, collected from Flickr creative commons and LabelMe [52]. The corresponding eye tracking data from 15 users who free viewed these images are also recorded. The dimension images ranges from 405 to 1024. There are 779 landscape images and 228 portrait images. Also some recent research work pays much attention to this database [28,33,53].

For the reason of generalization and universality, we evaluate our proposed approach and compare with other 18 saliency models based on the MSRA dataset for detecting salient object, and based on the YORK and the MIT datasets for the task of human fixation prediction. We denote these 18 methods Itti98 [12], Bruce06 [25], Zhai06 [54], Harel07 [13], Hou07 [16], Hou08 [21], Zhang08 [46], Achanta09 [14], Seo09 [55], Achanta10 [50], Goferman10 [32], Rahtu10 [31], Wang10 [27], Cheng11HC [34], Cheng11RC [34], Li11 [18], Murray11 [28] and Hou12 [51], respectively. We use the same parameters of our proposed model in all experiments and all results of other methods are obtained by executing their corresponding public available softwares or codes.

3.2. Experiments

Based on the MSRA dataset, we generate the saliency maps with our approach and compare with other 18 methods and ground truths. Some results are illustrated in Fig. 3. The results of Itti98 and Hou08 lack conspicuous regions, and extract only a small part of the salient object. Bruce06, Harel07, Hou07, Zhang08, Seo09, Geforman10, Wang10 and Hou12 care more about local abrupt changes so they can only capture edges of objects. Zhai06, Achanta09, Achanta10, Li11, and Murray11 pay attention to the whole region of the salient object, but they either miss large parts of salient objects, or produce unreasonable or diffuse maps. By contrast, Rahtu10, Cheng11HC and Cheng11RC perform better, but their results involve a lot of background details. Obviously, our results are more closely similar to ground truths.

For comparing the quality of different saliency maps, we first utilize a widely used method, the receiver operating characteristics (ROC) curve. In general, ROC curve is a useful tool to visualize the performance of binary classifiers [56]. Besides, it is the most prevalent criteria for evaluating the performance of visual saliency models. Given a saliency map and a binary ground truth mask, the true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN) can be calculated as follows:

$$\begin{cases} TP=\sum_{i}^{N_{im}}\varphi(S_i,t)M_i \\ FN=\sum_{i}^{N_{im}}\varphi(t,S_i)M_i \\ FP=\sum_{i}^{N_{im}}\varphi(S_i,t)(1-M_i) \\ TN=\sum_{i}^{N_{im}}\varphi(t,S_i)(1-M_i) \end{cases} \tag{12}$$

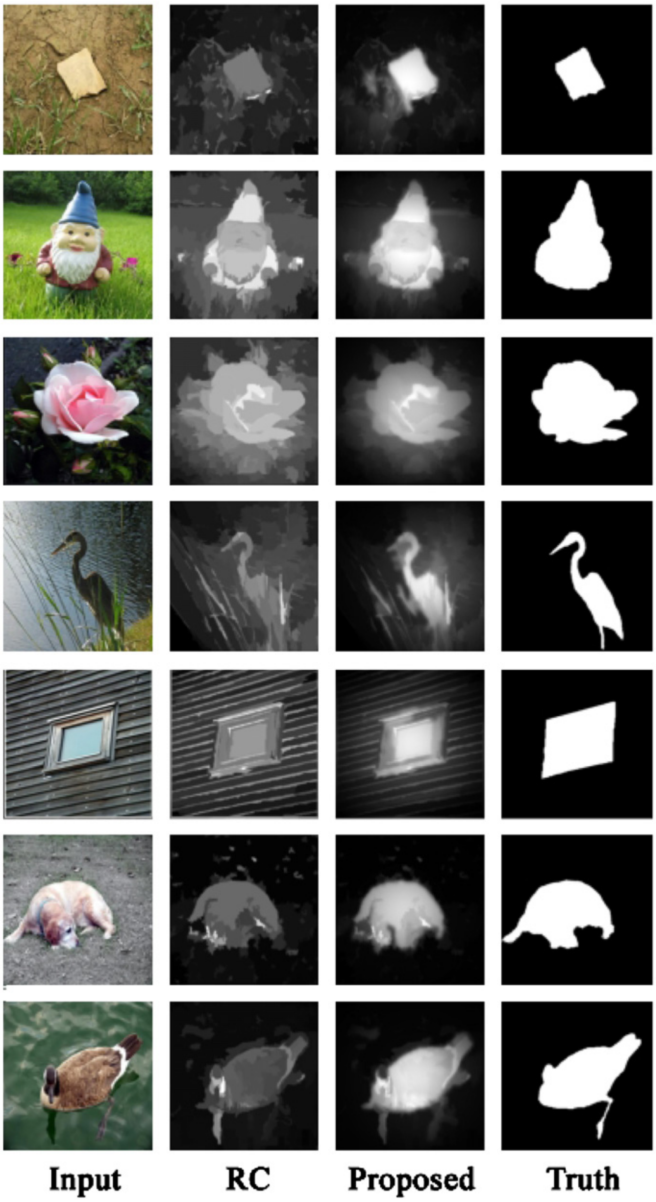

Fig. 7. Comparison with region based contrast method (RC [34]).

where N_{im} is the total number of pixels in the saliency map S, t is the threshold for binarization, M is the binary mask and the function $\varphi(\cdot)$ is defined as

$$\varphi(a,b)=\begin{cases}1, & a\geq b\\ 0, & a<b\end{cases} \tag{13}$$

Correspondingly, the false positive rate (FPR) is calculated as $FP/(FP+TN)$ and the true positive rate (TPR) is calculated as $TP/(TP+FN)$. By varying the threshold t from 0 to 255, furthermore, the ROC curve for the saliency model is plotted as the FPR versus TPR. For further quantitative comparisons, the area under the ROC curve (AUC) is also calculated. Therefore, for a given dataset, the mean FPR and TPR is computed for plotting the ROC curve for all test data, and the mean AUC is also computed to demonstrate the overall performance of the saliency model. The ROC curves of various saliency models is shown in Fig. 4, and the corresponding AUC scores are given in Table 1. Obviously, our method achieves a robust ROC curve and a high AUC score. We achieve a AUC score 0.9744, followed by the score 0.9636 obtained by Cheng11RC.

After that, we employ an another method, the precision and recall (PR) curve, to measure the performance of different saliency models. As argued in [57], a curve dominates in ROC space if and only if it dominates in PR space, simple linear interpolation is insufficient between points in PR space, and an algorithm that optimizes the area under the ROC curve is not guaranteed to optimize the area under the PR curve, we choose PR curve as a further strong evaluator. Similarly, the precision is defined as $TP/(TP+FP)$, and the recall is $TP/(TP+FN)$. Following the same experimental setting as in [14,34], we vary the threshold from 0 to 255 on a given saliency map with saliency values in the range [0,255], and compute the precision and recall at each value of the threshold. The precision and recall curves are shown in Fig. 5. It can be seen distinctly that the curve of our model demonstrates better performance than the others.

Finally, in order to automatically detect the salient object, we use an optimized threshold of each saliency model by maximizing the F-measure of each model,

$$T_{pr}=\arg\max F_{\beta} \tag{14}$$

where F_{β} is defined as

$$F_{\beta}=\frac{(1+\beta^2)\times precision\times recall}{\beta^2\times precision+recall} \tag{15}$$

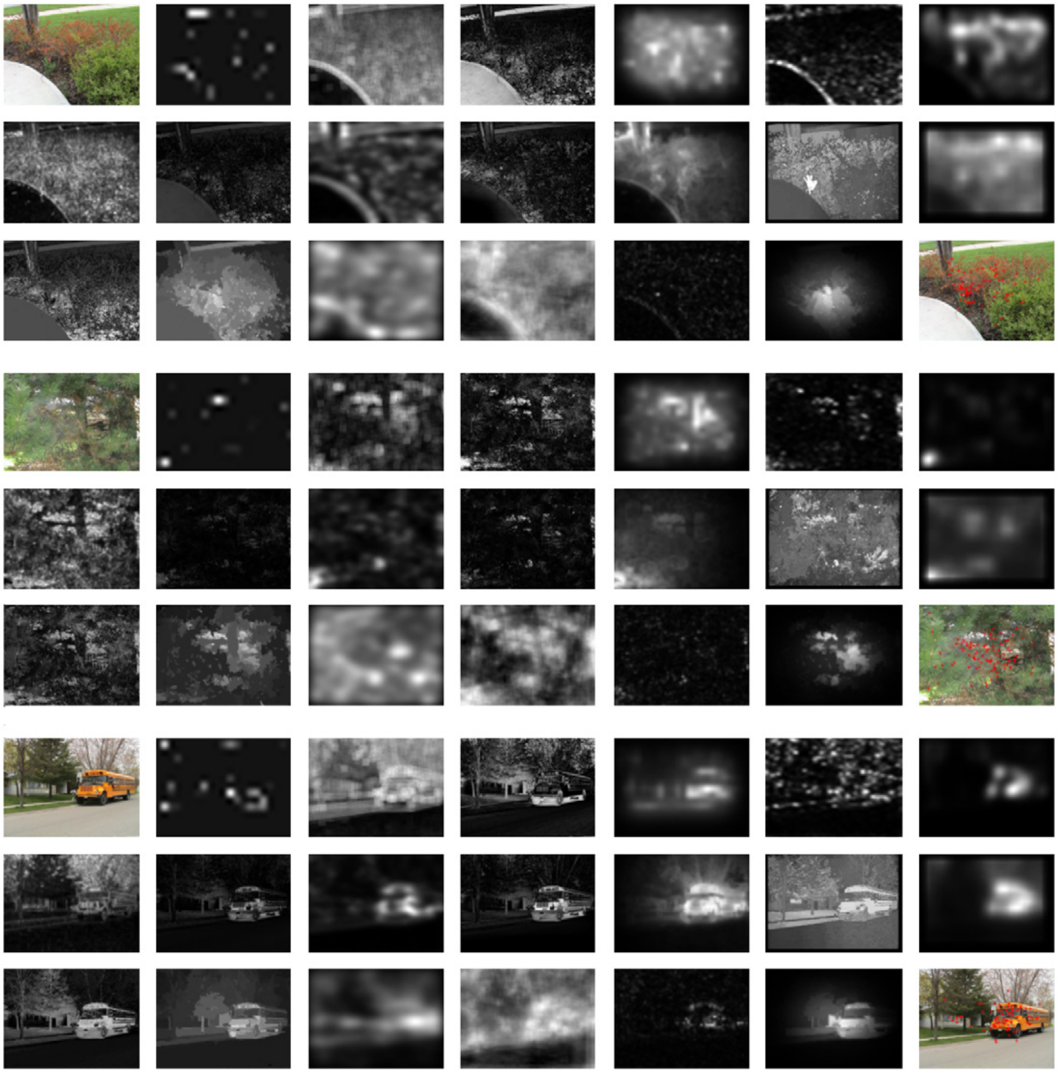

Fig. 8. Comparison of saliency estimation results for various methods with our approach on the YORK dataset. Images are shown in the order of the original image, results on Itti98 [12], Bruce06 [25], Zhai06 [54], Harel07 [13], Hou07 [16], Hou08 [21], Zhang08 [46], Achanta09 [14], Seo09 [55], Achanta10 [50], Goferman10 [32], Rahtu10 [31], Wang10 [27], Cheng11HC [34], Cheng11RC [34], Li11 [18], Murray11 [28], Hou12 [51] and the proposed model, and the ground truth at the end, where red crosses in the image are recorded human eye fixations. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

For weighing the precision more than the recall, β^2 is set to 0.3 similar to the setting in [14,34]. That is, by obtaining the optimized threshold, we compare the optimized detection result of each method without adopting any other algorithms. The comparison results are shown in Fig. 6. We can see that the optimized segmented results using our approach significantly outperform other methods with $F_\beta = 84.54\%$, *precision*=87.96%, and *recall*=74.84%, and the suboptimal result is obtained by Cheng11RC with $F_\beta = 80.28\%$, *precision*=83.74%, and *recall*=70.55%. In addition, most models obtain a higher precision than the recall because we set $\beta^2 = 0.3$ which emphasizes more on the precision.

From experimental results shown above, we can see that our model outperforms other methods. Especially, since we combine the central bias and multiscale analysis into traditional feature based method, and adopt an useful center shift process, our model provides more reasonable results. To clearly point out the advantage of our method, we compare with region based contrast method in [34], which is the most effective salient object detector in existing saliency models. Fig. 7 shows that the saliency maps of our method extract clearer objects, and discard more irrelevant details. The better performance may come from distinguishing features, central bias, center shift and multiscale analysis.

We use the eye-tracking data of the YORK and the MIT datasets to evaluate our extracted saliency maps. Some saliency maps generated by various saliency models are illustrated in Figs. 8 and 9. We can see that our method provides more consistent salient regions with human fixations. In addition, though Harel07 obtains relatively better results, it does not concentrate enough on the center of the fixations.

We also utilize the ROC curve and the AUC score as evaluators of various saliency models. From Figs. 10 and 11 we can see that our method provides better ROC curves on both the YORK and the MIT datasets. Meanwhile, as can be seen from Table 1, we achieve the highest AUC score 0.8173 and 0.8243 respectively, and the suboptimal result is obtained by Harel07 with 0.8078 and 0.8146 respectively. In general, results of these models on the two datasets are mainly consistent, which proves the reliability of the ROC curve and the AUC score.

3.3. Discussion

As mentioned above, visual saliency estimation can be evaluated by binary masks including salient objects and recorded

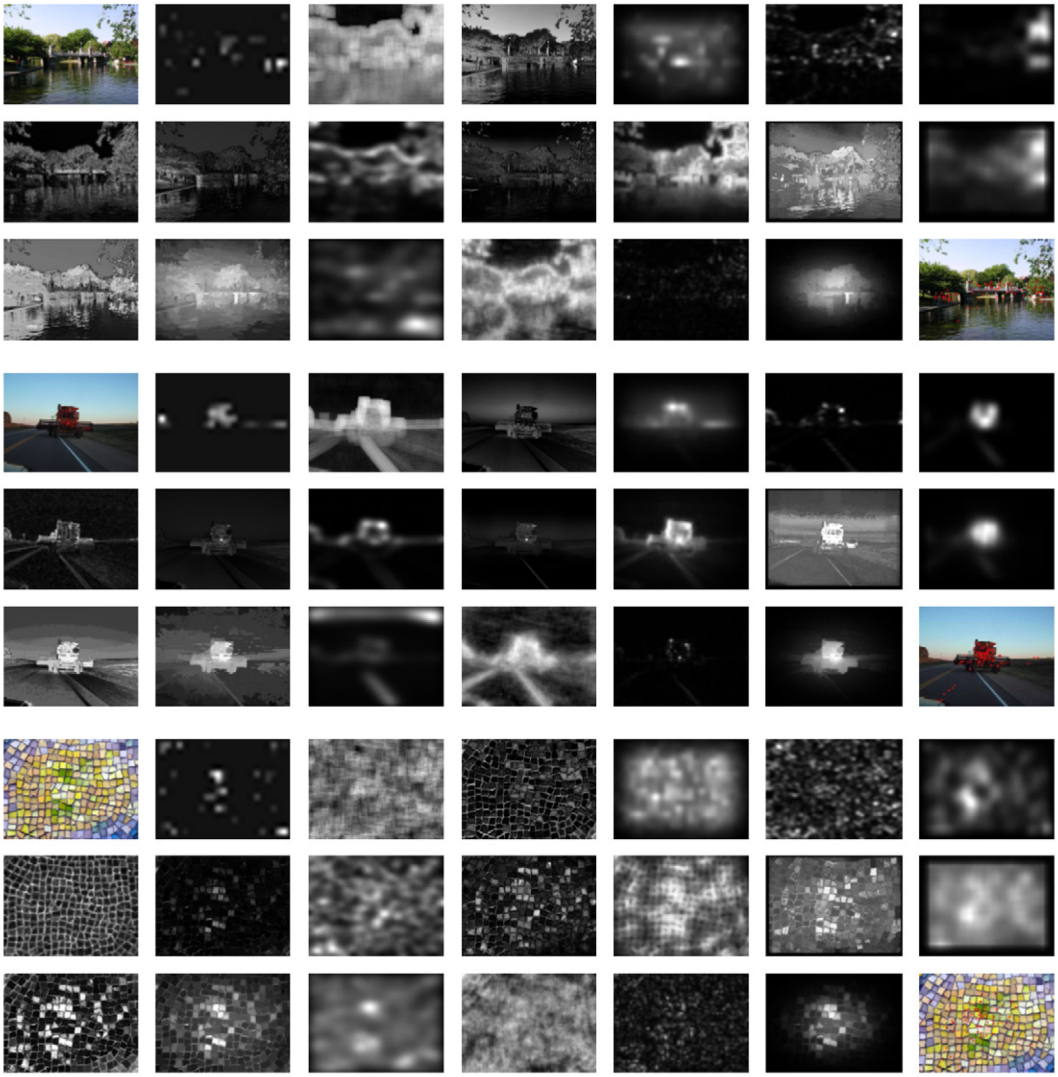

Fig. 9. Comparison of saliency estimation results for various methods with our approach on the MIT dataset. Images are shown in the order of the original image, results on Itti98 [12], Bruce06 [25], Zhai06 [54], Harel07 [13], Hou07 [16], Hou08 [21], Zhang08 [46], Achanta09 [14], Seo09 [55], Achanta10 [50], Goferman10 [32], Rahtu10 [31], Wang10 [27], Cheng11HC [34], Cheng11RC [34], Li11 [18], Murray11 [28], Hou12 [51] and the proposed model, and the ground truth at the end, where red crosses in the image are recorded human eye fixations. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

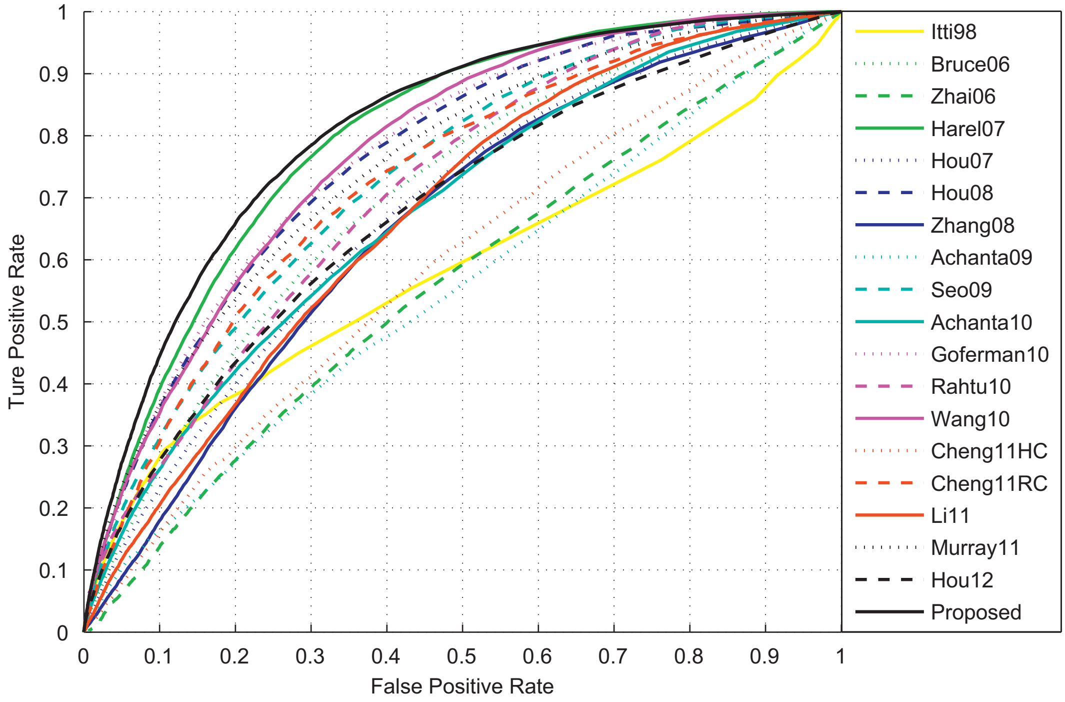

Fig. 10. ROC curves of various saliency models on the YORK dataset.

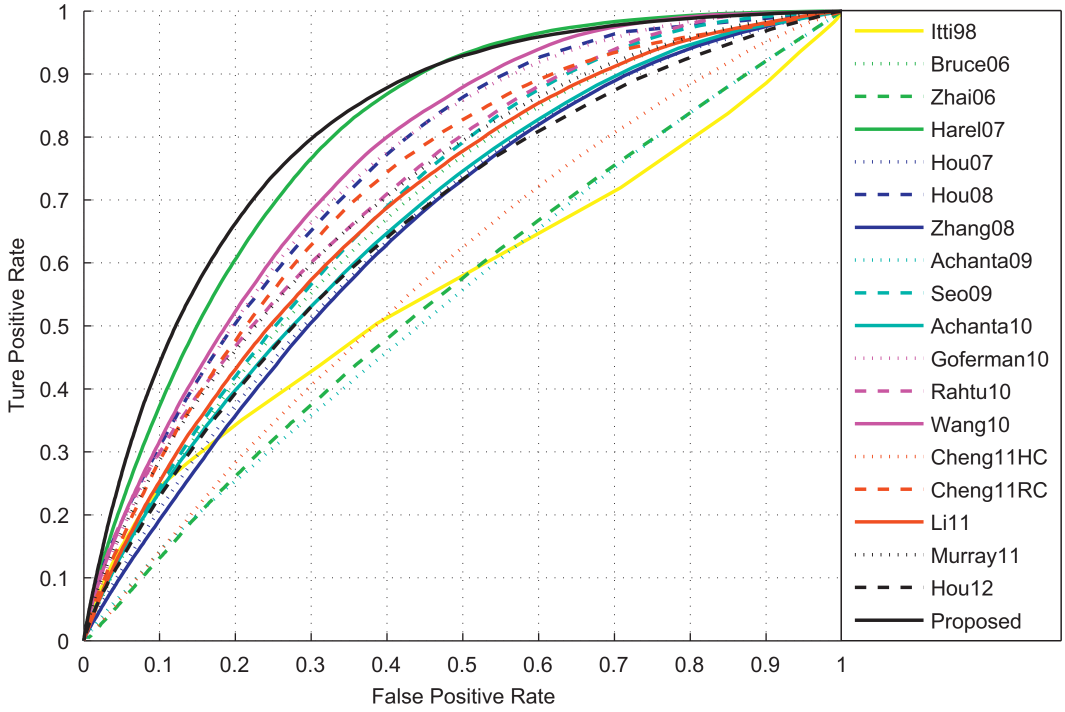

Fig. 11. ROC curves of various saliency models on the MIT dataset.

human eye fixations, so we compare our approach with other models based on both tasks. As illustrated by various experimental results, existing saliency models concentrate only either salient object detection or human fixation prediction, so they may not perform well simultaneously. For example, Cheng11RC, Cheng11HC and Rahtu10 achieve better performances for detecting salient objects than other state-of-the-art saliency models, but for predicting human eye fixations, their results are mediocre even. On the contrary, the similar phenomenon also happens on Harel07, Wang10, Goferman10 and Zhang08, which perform better on human fixation prediction. However, our method performs well on all three datasets, particularly on the MSRA dataset. Three possible reasons are as follows. (1) We view regions segmented by the graph-based image segmentation algorithm as the prime elements in the image, which means that we first coarsely partition all pixels into some meaningful regions with similar low level features. This makes the consideration that the saliency map is organized by meaningful regions with different saliency values. Obviously, it is more consistent with the way human freely views scenes. (2) If there exists a salient object in a natural image, our attention is attracted soon, while if there is no salient object, we pay attention to the center of the visual field and some non-meaningful regions with high dissimilarity form their neighborhood. Some other saliency models give equal importance to all pixels in the image, so that their saliency maps highlight not only salient regions but also parts of background. (3) The shifting process is important for visual saliency estimation. As we mentioned in Section 2.2, human visual system transfers its visual center from the center of the image to the center of distribution of features. So our model based on a center shift process outperforms other models which only emphasize the image center all the time.

4. Conclusions

The most important problems involving visual saliency estimation are what factors influence the saliency map and how they work. We choose low level features and location prior as two key factors to calculate the image saliency. Meanwhile, we propose a center shift process to simulate the shift of human visual field. We combine center shift and multiscale analysis to provide more reliable saliency maps in various complex scenes, which is a distinct difference comparing with existing saliency models. As a result, our model outperforms 18 state-of-the-art saliency models on both salient object detection and human fixation prediction. Since we view regions as prime elements and consider center shift and multiscale analysis, it is believed that our approach can be applied to other large-scale datasets as well and to other applications in pattern recognition and computer vision. However, robust visual saliency estimation in complex conditions is still a challenge. Future work may focus on finding out more meaningful physiological, psychological and mathematical evidences, and combining more effective similarity measurement methods. In addition, robust visual saliency detectors may be used in many applications, such as video compression, image segmentation, object tracking, and so on.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61173129, 61173130, 61103116), Research Grants of University of Macau (SRG010-FST11-TYY, MYRG187(Y1-L3)-FST11-TYY, RG205(Y1-L4)-FST11-TYY), and the Fundamental Research Funds for the Central University (CDJXS11182240 and CDJXS10182216).

References

[1] J. Tsotsos, Analyzing vision at the complexity level, Behav. Brain Sci. 13 (3) (1990) 423–469.

[2] K. Chang, T. Liu, S. Lai, From co-saliency to co-segmentation: an efficient and fully unsupervised energy minimization model, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 2129–2136.

[3] Z. Ding, Y. Yu, B. Wang, L. Zhang, An approach for visual attention based on biquaternion and its application for ship detection in multispectral imagery, Neurocomputing 76 (1) (2012) 9–17.

[4] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 989–1005.

[5] S. Han, N. Vasconcelos, Biologically plausible saliency mechanisms improve feedforward object recognition, Vision Res. 50 (22) (2010) 2295–2307.

[6] H. Liu, I. Heynderickx, Visual attention in objective image quality assessment: based on eye-tracking data, IEEE Trans. Circuits Syst. Video Technol. 21 (7) (2011) 971–982.

[7] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, D. Kukolj, Salient motion features for video quality assessment, IEEE Trans. Image Process. 20 (4) (2011) 948–958.

[8] G. Bhatnagar, Q. Wu, An image fusion framework based on human visual system in framelet domain, Int. J. Wavelets Multiresolution Inf. Process. 10 (1) (2012) 1–30.

[9] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image Vision Comput. 29 (1) (2011) 1–14.

[10] Q. Wang, F. Chen, W. Xu, Saliency selection for robust visual tracking, in: IEEE International Conference on Image Processing, IEEE, 2010, pp. 2785–2788.

[11] A. Toet, Computational versus psychophysical bottom-up image saliency: a comparative evaluation study, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2131–2146.

[12] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[13] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems, vol. 19, 2007, pp. 545–552.

[14] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1597–1604.

[15] Y. Lin, B. Fang, Y. Tang, A computational model for saliency maps by using local entropy, in: AAAI Conference on Artificial Intelligence, 2010, pp. 967–973.

[16] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[17] L. Duan, C. Wu, J. Miao, L. Qing, Y. Fu, Visual saliency detection by spatially weighted dissimilarity, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 473–480.

[18] J. Li, M. Levine, X. An, H. He, Saliency detection based on frequency and spatial domain analyses, in: British Machine Vision Conference, BMVA Press, 2011, pp. 86.1–86.11.

[19] T. Liu, J. Sun, N. Zheng, X. Tang, H. Shum, Learning to detect a salient object, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.

[20] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.

[21] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, in: Advances in Neural Information Processing Systems, vol. 21, 2008, pp. 681–688.

[22] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.

[23] J. Li, Y. Tian, T. Huang, W. Gao, Multi-task rank learning for visual saliency estimation, IEEE Trans. Circuits Syst. Video Technol. 21 (5) (2011) 623–636.

[24] P. Rosin, A simple method for detecting salient regions, Pattern Recognition 42 (11) (2009) 2363–2371.

[25] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, vol. 18, 2006, pp. 155–162.

[26] D. Gao, N. Vasconcelos, Bottom-up saliency is a discriminant process, in: IEEE International Conference on Computer Vision, IEEE, 2007, pp. 1–6.

[27] W. Wang, Y. Wang, Q. Huang, W. Gao, Measuring visual saliency by site entropy rate, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2368–2375.

[28] N. Murray, M. Vanrell, X. Otazu, C. Parraga, Saliency estimation using a non-parametric low-level vision model, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 433–440.

[29] M. Aziz, B. Mertsching, Fast and robust generation of feature maps for region-based visual attention, IEEE Trans. Image Process. 17 (5) (2008) 633–644.

[30] T. Avraham, M. Lindenbaum, Esaliency (extended saliency): meaningful attention using stochastic image modeling, IEEE Trans. Pattern Anal. Mach. Intell. 32 (4) (2010) 693–708.

[31] E. Rahtu, J. Kannala, M. Salo, J. Heikkilä, Segmenting salient objects from images and videos, in: European Conference on Computer Vision, 2010, pp. 366–379.

[32] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2376–2383.

[33] L. Duan, C. Wu, J. Miao, A. Bovik, Visual conspicuity index: spatial dissimilarity, distance and central bias, IEEE Signal Process. Lett. 18 (11) (2011) 690–693.

[34] M. Cheng, G. Zhang, N. Mitra, X. Huang, S. Hu, Global contrast based salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 409–416.

[35] Y. Xue, Z. Liu, R. Shi, Saliency detection using multiple region-based features, Opt. Eng. 50 (5) (2011) 057008:1–9.

[36] W. Luo, H. Li, G. Liu, K. Ngi Ngan, Global salient information maximization for saliency detection, Signal Process.: Image Commun. 27 (3) (2012) 238–248.

[37] C. Koch, S. Ullman, Shifts in selective visual attention: toward the underlying neural circuitry, Human Neurobiol. 4 (4) (1985) 219–227.

[38] A. Treisman, G. Gelade, A feature-integration theory of attention, Cognitive Psychol. 12 (1) (1980) 97–136.

[39] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395–1407.

[40] R. Valenti, N. Sebe, T. Gevers, Image saliency by isocentric curvedness and color, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 2185–2192.

[41] Q. Wang, P. Yan, Y. Yuan, X. Li, Multi-spectral saliency detection, Pattern Recognition Lett. 34 (1) (2013) 34–41.

[42] F. Perazzi, P. Krahenbul, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[43] B. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions, J. Vision 7 (14) (2007) 4,1–17.

[44] P. Felzenszwalb, D. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vision 59 (2) (2004) 167–181.

[45] B. Alexe, T. Deselaers, V. Ferrari, What is an object? in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 73–80.

[46] L. Zhang, M. Tong, T. Marks, H. Shan, G. Cottrell, SUN: A Bayesian framework for saliency using natural statistics, J. Vision 8 (7) (2008) 32,1–20.

[47] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: International Conference on Computer Vision, IEEE, 2009, pp. 2106–2113.

[48] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to SIFT or SURF, in: IEEE International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.

[49] Z. Wang, B. Li, A two-stage approach to saliency detection in images, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 965–968.

[50] R. Achanta, S. Susstrunk, Saliency detection using maximum symmetric surround, in: IEEE International Conference on Image Processing, IEEE, 2010, pp. 2653–2656.

[51] X. Hou, J. Harel, C. Koch, Image signature: highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 194–201.

[52] B. Russell, A. Torralba, K. Murphy, W. Freeman, LabelMe: a database and web-based tool for image annotation, Int. J. Comput. Vision 77 (1) (2008) 157–173.

[53] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, J. Vision 11 (3) (2011) 9,1–15.

[54] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: ACM International Conference on Multimedia, ACM, 2006, pp. 815–824.

[55] H. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vision 9 (12) (2009) 51,1–27.

[56] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Lett. 27 (8) (2006) 861–874.

[57] J. Davis, M. Goadrich, The relationship between Precision–Recall and ROC curves, in: International Conference on Machine Learning, ACM, 2006, pp. 233–240.

Weibin Yang received his B.S. degree in Computer Science from Southwest University, Chongqing, China, in 2006, M.S. degree in Computer Application Technology from Chongqing University, Chongqing, China, in 2009. He is currently a Ph.D. candidate in the Department of Computer Science of Chongqing University, Chongqing, China. His research interests include computer vision, image processing and pattern recognition.

Yuan Yan Tang received the B.S. degree in Electrical and Computer Engineering from Chongqing University, Chongqing, China, the M.S. degree in Electrical Engineering from the Beijing University of Post and Telecommunications, Beijing, China, and the Ph.D. degree in Computer Science from Concordia University, Montreal, QC, Canada.

Professor Yan Yuan Tang is a Chair Professor in Faculty of Science and Technology at University of Macau (UM). Before joining UM, Prof. Tang served as Chair Professor of Department of Computer Science in Hong Kong Baptist University, Dean of College of Computer Science in Chongqing University, China. Prof. Tang is a Fellow of IEEE, Fellow of Pattern Recognition Society (IAPR) and Chair of Technical Committee on Pattern Recognition in IEEE Systems, Man, and Cybernetics Society (IEEE SMC) for his great contributions to wavelet analysis, pattern recognition and document analysis. Recently, he was elected as one of the executive directors of the Chinese Association of Automation Council. With all his distinguished achievement, Prof. Tang is also the Founder & Editor-in-Chief of "International Journal of Wavelets, Multiresolution, and Information Processing (IJWMIP)", and Associate Editor of "International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)" and "International Journal on Frontiers of Computer Science (IJFCS)". He was presented a numerous awards such as the First Class of Natural Science Award of Technology Development Centre, Ministry of Education of the People's Republic of China in November 2005 and the Outstanding Contribution Award by the IEEE System, Man and Cybernetics Society in 2007. He has published more than 300 papers, books and book chapters.

Bin Fang received the B.S. degree in Electrical Engineering from Xi'an Jiaotong University, Xi'an, China, the M.S. degree in Electrical Engineering from Sichuan University, Chengdu, China, and the Ph.D. degree in Electrical Engineering from the University of Hong Kong, Hong Kong. He is currently a Professor with the College of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, information processing, biometrics applications, and document analysis. He has published more than 120 technical papers and is an Associate Editor of the International Journal of Pattern Recognition and Artificial Intelligence. Prof. Fang has been the Program Chair, and a Committee Member for many international conferences.

Zhaowei Shang received the B.S. degree in Computer Science from the Northwest Normal University, Lanzhou, China, in 1991, the M.S. degree from the Northwest Polytechnical University, Xi'an, China, in 1999, and the Ph.D. degree in Computer Engineering from Xi'an Jiaotong University, Xi'an, in 2005. He is currently an Associate Professor with the Department of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, image processing, and wavelet analysis.

Yuewei Lin received the B.S. degree in Optical Information Science & Technology from Sichuan University, Chengdu, China, and the M.E. degree in Optical Engineering from Chongqing University, Chongqing, China. He is currently working toward the Ph.D. degree in the Department of Computer Science & Engineering, University of South Carolina. His current research interests include computer vision, image/video processing.