J. Vis. Commun. Image R. 24 (2013) 465–476

Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

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

Saliency detection using joint spatial-color constraint and multi-scale segmentation

Linfeng Xu [a], Hongliang Li [a,*], Liaoyuan Zeng [a], King Ngi Ngan [b]

[a] School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 11 May 2012
Accepted 30 January 2013
Available online 19 February 2013

Keywords:
Visual attention
Saliency model
Region detection
Human fixation prediction
Spatial constraint
Color double-opponent
Similarity distribution
Multi-scale technique
Segmentation-based method

ABSTRACT

In this paper, a novel method is proposed to detect salient regions in images. To measure pixel-level saliency, joint spatial-color constraint is defined, i.e., spatial constraint (SC), color double-opponent (CD) constraint and similarity distribution (SD) constraint. The SC constraint is designed to produce global contrast with ability to distinguish the difference between "center and surround". The CD constraint is introduced to extract intensive contrast of red-green and blue-yellow double opponency. The SD constraint is developed to detect the salient object and its background. A two-layer structure is adopted to merge the SC, CD and SD saliency into a saliency map. In order to obtain a consistent saliency map, the region-based saliency detection is performed by incorporating a multi-scale segmentation technique. The proposed method is evaluated on two image datasets. Experimental results show that the proposed method outperforms the state-of-the-art methods on salient region detection as well as human fixation prediction.

1. Introduction

As William James, the father of American psychology, remarked, "Every one knows what attention is" [1], attention is the cognitive process of selectively concentrating on the important parts in a complex visual environment while ignoring the others. Attention is considered to be a key ability that enables creatures to find their prey or potential danger rapidly. Furthermore, it allows brain and visual system to break through the information-processing bottleneck because human visual system only efficiently processes parts of the massive sensory incoming information in detail [2]. Two mechanisms are believed for attention deployment: the bottom-up, rapid, pre-attentive and stimulus-driven manner as well as the top-down, slower, attentive and task-dependent manner [3–5]. Visual attention is of widespread interest due to a large number of applications, including adaptive image/video compression [6–9], object-of-attention image segmentation [10–13], object recognition [14,15], surveillance [16], smart image retargeting [17,18], image/video retrieval and summary [18–20].

Visual saliency is the perceptual quality that makes an object visually different to its neighborhoods and grabs our attention [21]. In the last several decades, various computational models for visual saliency detection have been proposed in physiology, neuro-psychology, cognitive science, computer vision, etc. The purpose of these models is to predict the areas which attract human attention. The result is often a saliency or master map of the input signal, which is a scalar, two-dimensional map providing higher intensities for the most prominent areas [3,22]. It is generally considered that the saliency lies in visual uniqueness, abnormality, maximization of self-information, surprise or common objects in image pairs [5,23–25]. Let us imagine the following visual phenomena: a bright red flower among green leaves in a dark background or a sudden movement in a still scene. The flower or the unexpected movement is likely to be salient. The goal of saliency detection is to rapidly highlight the abnormal regions.

In this paper, we extract the salient regions based on a bottom-up stimulus driven saliency model since the bottom-up biological visual pathway is much simpler to be understood [26]. The proposed visual saliency detection method which adopts joint spatial-color constraint and multi-scale segmentation technique can uniformly highlight salient regions with full resolution and suppress the surrounding background even in the large object case. Three components are used to estimate the pixel-level saliency for an image. The first component performs global contrast with spatial constraint to distinguish the difference between "center and surround". The second one is to extract intensive contrast of red-green and blue-yellow double opponency. The third is computed to detect the salient object and its background by using the similarity distribution of a pixel. Then the pixel-level saliency map is fused from the three components by a two-layer structure. Finally, in order to produce a uniform region-level saliency map, we present a multi-scale segmentation based technique to increase the consistency of the map. The proposed method is evaluated on

* Corresponding author. Fax: +86 028 61830064.
E-mail addresses: lfxu@uestc.edu.cn (L. Xu), hlli@uestc.edu.cn (H. Li).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.02.007

two publicly available image datasets. On the 5000 salient object benchmark images, the proposed method achieves better performance than the state-of-the-art methods in terms of precision and recall. The precision and recall are commonly-used validation criteria for salient region extraction. On the eye tracking dataset, the comparison results demonstrate that the proposed method outperforms the existing methods on predicting human fixations.

The rest of this paper is organized as follows. The related work of saliency detection is introduced in Section 2. In Section 3, the proposed method is presented from pixel to region level to extract saliency in an image. The experimental results are shown in Section 4. Finally, conclusion is drawn in Section 5.

2. Related work

In order to efficiently detect saliency in an image, lots of bottom-up saliency-driven methods have been proposed in the past few decades, such as biology-based, purely computational or an integration of the two [27]. We epitomize these methods as two main ideas: detecting saliency based on local or global contrast.

Local contrast based methods, which have overt biological supporting, explore a salient feature depending on its neighborhoods. Based on a biologically-inspired architecture proposed by Koch and Ullman [22], which is motivated from Treisman and Gelade's feature integration theory [3], Itti et al. [5] proposed a bottom-up visual attention model for rapid analysis. Itti's model combined three multi-resolution extracted local feature contrasts, i.e., luminance, chrominance and orientation, to produce a topographic saliency map. Ma and Zhang [28] generated a saliency map using local contrast analysis and extracted salient objects using a fuzzy growing method from the map. Walther and Koch [29] extended Itti's model to infer proto object regions and applied it to achieve object recognition. By merging the seed values from Itti's saliency map and some low-level features, Han et al. [10] applied a Markov random field (MRF) model to segment salient objects in color images. Harel et al. [30] created feature maps using Itti's method but formed activation maps using a graph based approach, which were then normalized to more conspicuous maps using a Markovian algorithm. Gao and Vasconcelos [26] presented the discriminant saliency detection model by maximizing the mutual information between features of the center and surround regions in an image. Liu et al. [31] optimally combined a set of local, regional and global features including multi-scale contrast, center-surround histogram and color spatial distribution to describe a salient object through Conditional Random Field (CRF) learning. A limitation of the local contrast based methods is that the saliency maps obtained by these methods usually highlight the object boundary instead of the entire object, which will be shown in the section of experimental results.

Global contrast based methods integrate the entire information features all over the visual field, i.e., using the contrast to all of the pixels in an image to determine the saliency of a pixel or region. Zhai and Shah [32] detected the spatial saliency of a pixel by using 1-D histogram of a specific color channel (e.g., red channel) to compute the contrast to all the other pixels. This method ignored the relationship between different color channels. Bruce and Tsotsos [23] used an Independent Component Analysis (ICA) to represent the probability distribution of local image patches using a large database of patches from natural images. They proposed a bottom-up saliency model based on the strategy of maximum information sampling. Hou and Zhang [33] proposed a saliency detection method in spectral domain, which extracted the saliency map from the spectral residual of the log-spectrum of an image. Achanta et al. [27] applied a frequency-tuned technology to detect pixel saliency using color and luminance features. Goferman et al. [34] combined low-level features, global considerations and visual organization rules to obtain a context-aware saliency, and then they employed high-level factors for post-processing. Recently, Cheng et al. [35] extended the spatial saliency in [32] to yield a full resolution saliency map using the global contrast. They used a histogram-based method to make the processing more efficient and employed a smoothing procedure to reduce the saliency map noise. In Cheng's work, in order to highlight the entire objects uniformly, the saliency values are assigned to the over-segmented regions to produce region-based contrast maps. However, inconsistent saliency detection may be obtained due to the single-scale segmentation.

In general, the global contrast based methods are able to produce saliency maps with full resolution, defined boundaries and uniformly highlighted regions. However, the models used by most of these methods usually sum the unweighted contrast to all the other pixels to compute the saliency of a pixel. Let us consider a simple image including two regions, each of which has a single color. A higher saliency value will be assigned to the smaller region in the image by these methods. As a result, these methods are area-depending. In other words, these methods can detect salient regions with small sizes and obvious contrast with regard to the rest. However, if the salient regions have more pixels than the background or there are noisy regions, such as the shadows, in the scene, these methods may not detect the right region accurately.

3. Saliency detection from pixel to region level

3.1. Pixel saliency detection using joint spatial-color constraint

In this section, we propose a pixel-level saliency detection method using joint spatial-color constraint to overcome the area-depending deficiency and enhance the detection performance.

3.1.1. Spatial constraint based saliency

Based on the psychological observation in saliency detection [34–36], the saliency map should consider the spatial distance between adjacent pixels and group the similar pixels together. So we hypothesize that a pixel is salient when it is spatially close to the pixels which have strong contrast to it, and the pixel is contrarily less salient when it is far away from the strong contrast pixels. Definitely, extended from [32], the saliency of a pixel in an image I is defined as the spatially weighted global contrast to all the other pixels in the image, which is expressed as

$$S(p) = \sum_{\forall q \in I} \exp\left(-\frac{\|p-q\|^2}{\sigma_{s1}^2}\right) \mathcal{D}(\mathcal{I}_p, \mathcal{I}_q) \tag{1}$$

where I_p represents the color vector of pixel p in CIE L*a*b* color space, i.e., $I_p = (L_p, a_p, b_p)^T$, $\mathcal{D}(\mathcal{I}_p, \mathcal{I}_q)$ is the Euclidean distance between the color vectors of pixels p and q, $\|p-q\|$ is the Euclidean distance between the positions of the two pixels, and σ_{s1} is the control parameter for spatial similarity. Large σ_{s1} combines the farther pixels to compute the saliency of the current pixel.

Based on the "center-surround" inhibition mechanism in the early stages of biological vision, the center region is more salient than the surrounding background. However, most of the global contrast based methods are not able to represent this center-surround difference. Although the color distance of a pixel p in the center region to a pixel q in the surround is equal to the color distance of the pixel q to the pixel p, they should be treated differently under the center-surround mechanism. Specifically, for two pixels p and q with a strong contrast, the center pixel p is more salient than the surrounding pixel q. So a larger weight should be assigned to the color distance when the saliency of pixel p is computed using pixel q. As used in [37], we introduce the Gaussian weighting function from the center of the image to weight the

color distances for global contrast computation in (1). Thus the spatial constraint (SC) based saliency of a pixel is formulated as

$$S_{SC}(p) = \sum_{\forall q \in I} w^s_{p,q} \mathcal{D}(I_p, I_q) \tag{2}$$

where $w^s_{p,q}$ is the spatial constraint factor, defined as

$$w^s_{p,q} = \frac{1}{K} \exp\left(-\frac{\|p-q\|^2}{\sigma^2_{s1}}\right) \frac{1}{\mathcal{G}}(q, I^c) \tag{3}$$

where K is a normalizing constant that guarantees $\sum_q w^s_{p,q} = 1$, and $\mathcal{G}(q, I^c)$ denotes the Gaussian function of the position of pixel q from the center of the image I^c. The reciprocal of the Gaussian function for the surrounding pixel has a greater value than the center pixel. As shown in the experimental section, the SC saliency enables the proposed method to detect salient regions effectively even for the large object detection.

3.1.2. Color double-opponent saliency

According to the measurement in human visual cortex [38], which shows the strongest response is from red-green and blue-yellow stimuli, more human attention is grabbed by red-green and blue-yellow contrast. Following the method of generating red-green and blue-yellow opponencies in [29], we define the chromatic double opponency of red-green and blue-yellow for a pixel p as

$$RG(p) = R(p) - G(p) \tag{4}$$

$$BY(p) = B(p) - \min\{R(p), G(p)\} \tag{5}$$

with $R(p), G(p)$ and $B(p)$ being the red, green and blue channels of the pixel, respectively. Then the color double-opponent (CD) saliency is presented to highlight the pixels with strong global red-green or blue-yellow contrasts which are defined as

$$S_{RG}(p) = \frac{1}{N} \sum_{\forall q \in I} |RG(p) - RG(q)| \tag{6}$$

$$S_{BY}(p) = \frac{1}{N} \sum_{\forall q \in I} |BY(p) - BY(q)| \tag{7}$$

where N is the number of pixels in the image. Summing the global contrasts of RG and BY, the CD saliency of a pixel is computed as

$$S_{CD}(p) = \frac{S_{RG}(p) + S_{BY}(p)}{\mathcal{D}_{max}(p)} \tag{8}$$

where the normalization factor $\mathcal{D}_{max}(p)$ represents the maximal value among the values of $|RG(p) - RG(q)|$ and $|BY(p) - BY(q)|$ for a pixel p with respect to all the other pixels q. An example of the CD saliency is shown in Fig. 1. The original image and the spatial constraint based saliency map are shown in Fig. 1(a) and (b), respectively. From the saliency map, we can see that the dark green leaves in the background are detected as the salient objects, while the real salient petals are not indicated correctly. Fortunately, this problem can be solved by using the CD saliency. The computed saliency map in Fig. 1(c) just highlights the salient flower. Fig. 1(d) shows the final pixel-level saliency map obtained from the two-layer saliency structure which is introduced in the subSection 3.1.4.

3.1.3. Similarity distribution based saliency

As mentioned in Section 2, detecting the salient objects with very large areas or the objects affected by noise regions remains an open problem to most global contrast based methods. This problem can be solved by foreground detection because the foreground objects are generally more salient than their backgrounds. Actually, the foreground objects and the widely spreading background can be "easily" distinguished because the foreground objects generally distribute in a focused way and the background distributes dispersedly. For a pixel inside an object, the sum of spatial distances to all the other pixels within the same object is always much smaller than the sum for a background pixel. However, this seemingly easy process is difficult to implement because it is hard to distinguish whether a pixel is within an object or not.

Assuming the pixels within an object or background have similar color values, we propose the similarity distribution (SD) of pixels to detect the saliency of foreground objects. The SD of a pixel is related to the sum of spatial distances to the other similar pixels. The SD based saliency for a pixel p is defined as

$$S_{SD}(p) = \exp\left(-\frac{SimiDist(p)}{\sigma^2_{d1}}\right) \tag{9}$$

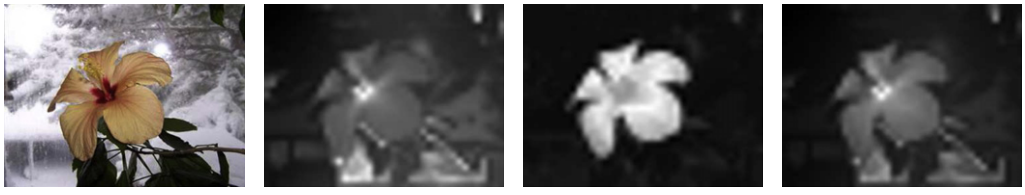
where $SimiDist(p)$ is the similarity distribution, and σ_{d1} is used to control the strength of the distribution. Large value of σ_{d1} reduces the effect of the SD saliency. $SimiDist(p)$ is formulated as

$$SimiDist(p) = \sum_{\forall q \in I} w^d_{p,q} \|p - q\|^2 \tag{10}$$

where $w^d_{p,q}$ denotes the similarity measurement between pixels p and q. We represent $w^d_{p,q}$ as

$$w^d_{p,q} = \frac{1}{K'} \exp\left(-\frac{\mathcal{D}(I_p, I_q)}{\sigma^2_{d2}}\right) \tag{11}$$

where $\mathcal{D}(I_p, I_q)$ is the color distance between the two pixels in CIE L*a*b* color space, K' is a normalizing constant that ensures $\sum_q w^d_{p,q} = 1$, and σ_{d2} is the control parameter for range similarity. Large value of σ_{d2} provides the two pixels with more similarity. For two similar pixels p and q with small color distance, the weight $w^d_{p,q}$ is approximately $\frac{1}{K'}$, and the weighted distance in (10) is close to the normalized spatial distance of the two pixels. On the contrary, if the two pixels are distinct, the corresponding weight will be close to zero. According to the hypothesis, if pixel p is within an object, $SimiDist(p)$ can be approximated as the normalized sum of spatial distances to the other pixels in the same object, and the value may be smaller than the corresponding value of a background pixel. So it is more likely for those pixels with larger SD saliency values (9)



(a) (b) (c) (d)

Fig. 1. An example of the color double-opponent (CD) saliency map: (a) original image, (b) spatial constraint (SC) based saliency map, (c) CD saliency map, (d) two-layer saliency map.

to be within a salient foreground object than those with smaller values.

3.1.4. Two-layer saliency structure

The SC, CD and SD saliency of the joint spatial-color constraint based method are proposed to detect salient pixels with strong "center-surround" contrast, intensive red-green and blue-yellow contrast, and greater chance to be within the foreground object, respectively. The three terms need to be combined to generate a saliency map in an effective manner. Using the two-layer saliency structure in [39], the final pixel-level saliency map is synthesized from two layers, the basic layer and enhancement layer, which is defined as follows.

(i) Basic layer is designed based on the spatial constraint based saliency, which obtains large saliency values when the global contrast is high.
(ii) Enhancement layer is designed based on the color double-opponent and similarity distribution saliency, which attempts to highlight the objects with bright color or more concentrated distribution when the SC saliency is relatively uniform in an image.

Based on the two layers, the final saliency value of a pixel p is defined as

$$S_m(p) = S_{SC}(p)(1 + \rho_1 S_{CD}(p) + \rho_2 S_{SD}(p)) \tag{12}$$

where $S_{SC}(p)$, $S_{CD}(p)$ and $S_{SD}(p)$ represent the normalized spatial constraint based saliency, color double-opponent saliency and similarity distribution based saliency at pixel p, respectively. The parameters ρ_1 and ρ_2 are weight factors which adjust the extent of emphasis for color double-opponent and similarity distribution saliency.

3.2. Extension to region saliency

As mentioned in [27,35,39], salient region detection is useful for many computer vision applications, such as object segmentation and recognition. In order to extract the salient object effectively, the saliency map should highlight the whole salient regions uniformly and consider the spatial relationship of pixels. Recently, based on this requirement, region based saliency models are proposed in [35,39,40], which first segment an image into multiple regions and compute the saliency value for each region using the histogram or filtering based methods.

Based on the extracted saliency map obtained from the two-layer saliency structure (12), we over-segment the image into regions to get the regional saliency. The first saliency value of each region is computed by averaging the saliency values of all the pixels in the region, which is represented as

$$S_1(r_p) = \frac{1}{N_r} \sum_{\forall q \in r_p} S_m(q) \tag{13}$$

where N_r denotes the pixel number of region r_p. Inspired by the guided image filtering in [41], the first saliency map obtained by (13) is filtered to generate the second saliency value of each region by using the original image as the guidance image. Namely, the second saliency value of region r_p is produced by the combination of weighted saliency values of the other regions, which is given by

$$S_2(r_p) = \sum_{r_q \neq r_p} w_{r_p,r_q} S_1(r_q) \tag{14}$$

where the weight w_{r_p,r_q} is related to the spatial distance and color distance of the two regions. It is defined as

$$w_{r_p,r_q} = \frac{1}{K''} \exp\left(-\frac{\|r_p^c - r_q^c\|^2}{\sigma_{r1}^2}\right) \exp\left(-\frac{\mathcal{D}(r_p, r_q)}{\sigma_{r2}^2}\right) \tag{15}$$

where K'' is the normalizing constant, r_p^c and r_q^c represent the centroids of regions r_p and r_q, respectively, and $\mathcal{D}(r_p, r_q)$ denotes the distance between the average color values of the two regions. The parameters σ_{r1} and σ_{r2} are the control parameters for the spatial and range similarity, respectively. For example, large value of σ_{r1} combines values from farther regions to obtain the saliency of the current region.

3.3. Saliency extraction using multi-scale segmentation

The ultimate goal of the region based saliency is to highlight the salient object uniformly. As shown in Fig. 2(c), if the butterfly can be completely segmented, the computed saliency map shown in Fig. 2(d), which is obtained by averaging the saliency values of the pixels in the region using (13), is almost the same with the ground truth. However, depending on different segmentation algorithms, an object may be segmented into different regions or a region may contain several objects. So the region based saliency may not exactly reflect the saliency of the object, i.e., the saliency map may not uniformly highlight the salient object or may be disturbed by the disordered backgrounds as shown in Fig. 2(e). In order to increase the consistency of the region-level saliency map, we propose a multi-scale segmentation based saliency model which uses saliency values of multiple segmentation results to generate the saliency of the image.

This model contains six steps, they are:

Step 1. The baseline regions are generated using over-segmentation.
Step 2. The original image is segmented into larger or smaller regions to form M different segmentation scales by setting the parameters of the segmentation algorithm.
Step 3. The saliency values of regions in all the scales are computed using (13).
Step 4. The saliency maps of pixels in the M segmentation scales are obtained. The saliency value of a region in a certain segmentation scale is used as the saliency values of the pixels in the region in the corresponding scale.
Step 5. The multi-scale segmentation based saliency of each baseline region can be obtained by combining the saliency values in the M segmentation scales. In the experiments, two combination schemes are adopted, i.e., maximization and average.
(i) The maximum of the M saliency values of a pixel is chosen to generate a saliency map of pixels. Then a saliency value of each baseline region is computed by using the saliency values (in the generated map) of the pixels in the region as in (13).
(ii) M saliency values of each baseline region is firstly computed by the combination of weighted saliency values of the regions in the corresponding scale as in (14). Then the M values are averaged to produce another saliency value of the baseline region.
Step 6. The two saliency values of each baseline region, in the Steps 5(i) and 5(ii), are merged into the final multi-scale segmentation based saliency.

Fig. 3 shows the flowchart of the model.

The multi-scale segmentation based saliency is formulated as follows. In the Step 5(i), the saliency value of a baseline region r_p is represented as

$$S_1^m(r_p) = \frac{1}{N_r} \sum_{\forall q \in r_p} S_{max}^m(q) \tag{16}$$

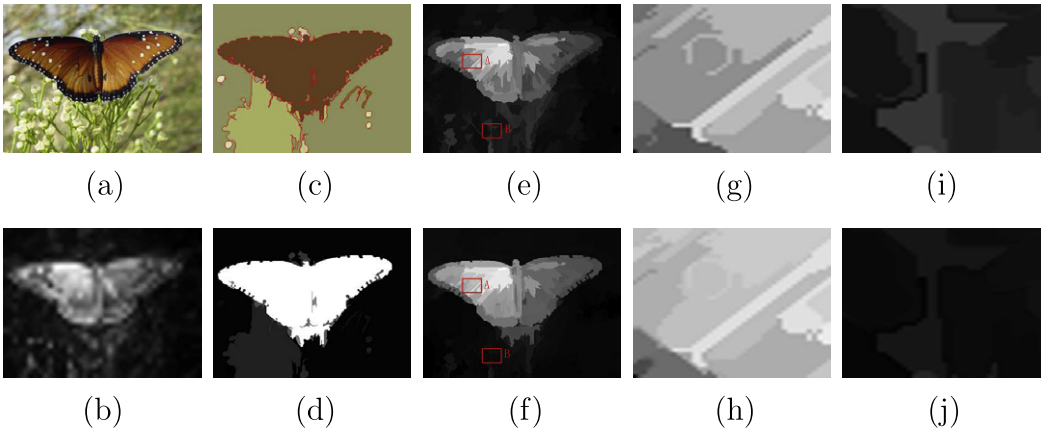


Fig. 2. An example of the multi-scale segmentation based saliency: (a) original image, (b) pixel saliency from the two-layer saliency structure, (c–d) nearly ideal segmentation and the corresponding normalized saliency map. (e, g, i) A map of region saliency from particular over-segmentation and its enlarged maps for region A and region B, respectively. (f, h, j) Multi-scale segmentation based saliency map and its enlarged maps for region A and region B, respectively.

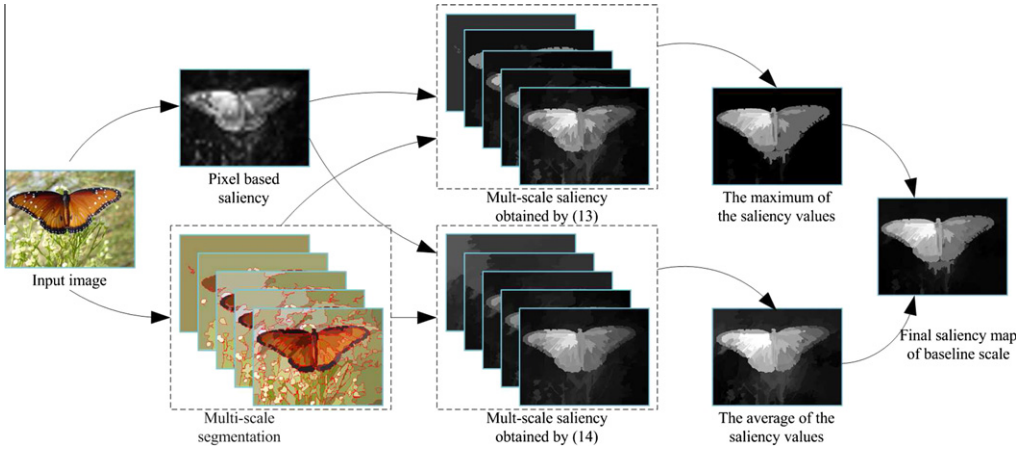


Fig. 3. The flowchart of the multi-scale segmentation model. Inputs are segmented into different regions in multiple scales to generate different saliency maps. The maximal and average counterparts of these maps are merged into the final saliency map.

where $S^m_{max}(q)$ is the maximum among the M saliency values of pixel q. In the Step 5(ii), the saliency value can be written as

$$S^m_2(r_p) = \frac{1}{M}\sum_{k=1}^{M} S^k_2(r_p) \tag{17}$$

where $S^k_2(r_p)$ is the saliency value of region r_p in the kth scale, which is obtained by the combination of weighted saliency values of the regions in the scale k following (14). As mentioned in the Step 6, the two saliency values are merged together in the experiments. Based on the visual psychological principle that humans generally pay more attention to the regions near to the image center [42], the final saliency value of region r_p in the baseline scale is defined as

$$S^m(r_p) = \exp\left(-\frac{\|r^c_p - I^c\|}{\sigma^2_m}\right)\left(\frac{1}{2}S^m_1(r_p) + \frac{1}{2}S^m_2(r_p)\right) \tag{18}$$

where I^c denotes the image center and σ_m is used to control the strength of spatial weighting. Large σ_m reduces the effect of spatial weighting.

The result of the multi-scale segmentation based saliency model is represented in Fig. 2(f). Comparing with the enlarged maps of local regions in Fig. 2(g)–(j), we can see that the multi-scale segmentation based saliency map highlights the object more uniformly and suppresses the background noise significantly.

4. Experimental results

In this section, the results of the proposed method are evaluated on two public image datasets. We compare the proposed method with the state-of-the-art saliency detection methods in subjective and objective assessments. The same parameters of the proposed method are used across the two datasets to demonstrate its robustness.

4.1. Parameters setting

In this section, the parameters setting in the experiments are introduced. When computing pixel-level saliency $S_m(p)$, the images are downsampled to compute the saliency map in order to reduce computational expense. Then the full resolution map is interpolated bilinearly. To compute the factor of SC saliency defined in (3), we set $\sigma^2_{s1} = 300$. To compute the SD saliency given in (9) and (11), the parameters are set as $\sigma^2_{d1} = 0.2$ and $\sigma^2_{d2} = 10$. We simplify the weights as $\rho_1 = \rho_2 = 1$ when we use the two-layer saliency structure in (12). These parameters are determined empirically and verified to positively contribute to the performance of saliency detection by the experiments. For the multi-scale segmentation, the graph-based segmentation algorithm [43] is used, in which the parameter k, i.e., the term for the threshold function [43], is adjusted to generate eight segmentation scales.

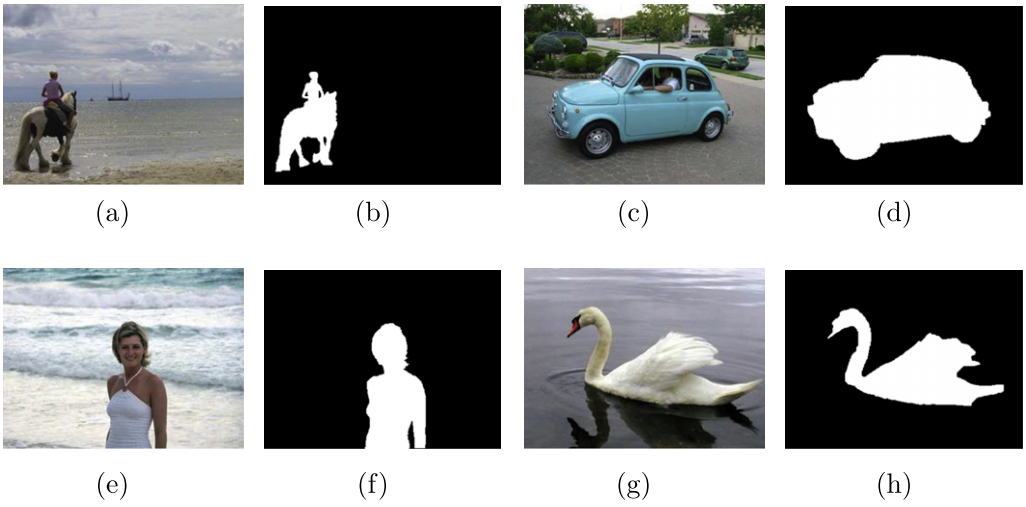


Fig. 4. Examples of MSRA image dataset and the corresponding ground truth mask: (a, c, e, g) original images; (b, d, f, h) ground truth masks.

4.2. Experiments on MSRA image dataset

In the first experiment, we evaluate the results of the proposed method on the popular MSRA image dataset used by Liu et al. [31], which comprises 5000 color images and human labeled bounding boxes indicating the most salient object. In order to make an accurate and objective assessment for the extracted saliency maps, we have manually partitioned all the images into salient objects and backgrounds according to the labeled boxes. The objects and backgrounds are labeled as one and zero in the ground-truth mask, respectively. The examples of images and the corresponding ground-truth masks from the dataset are shown in Fig. 4. All the manually labeled ground-truth can be downloaded from the website of the intelligent visual information processing and communication (IVIPC) lab in the near future.

We compare the proposed method with seven state-of-the-art saliency detection methods, i.e., IT [5], FT [27], GB [30], CA [34], MS [44], HC [35] and RC [35], which are listed in Table 1. These methods involve a variety of saliency models, such as biologically motivated, computational, frequency based, local contrast, global contrast or full resolution models. For the existing methods, we use the source codes or executable codes provided by the authors. The proposed method is implemented in Matlab.

Subjective comparison of the proposed method with the state-of-the-art methods are shown in Fig. 5. The original images are shown in Fig. 5(a), while the results of the seven state-of-the-art methods are presented in Fig. 5(b)–(h). The saliency maps obtained by the proposed method are given in Fig. 5(i), and the ground truth masks are illustrated in Fig. 5(j). The comparison results show that the proposed method can lead to improved performance for salient objects extraction from images. For the test images, in which the object of attention is small and distinct from the others, most methods are able to detect the salient object easily. For example, for the first image in Fig. 5, most maps can extract the lotus well. However, some of the methods, such as CA, HC and RC, spotlight the background regions as well, which can be solved by using joint constraint and multi-scale segmentation of the proposed method. Thus the proposed method can achieve higher precision for high recall values than the existing methods. Methods IT, GB and CA usually highlight the boundary, while the proposed method can highlight the salient regions uniformly. It is more applicable for the proposed method to extract or segment salient objects. Furthermore, the proposed method outperforms previous methods on the test images with large objects, complex backgrounds or noisy shadows. For instance, for the second to the seventh images, the saliency maps obtained by the proposed method coincide with the ground truth better.

To assess the salient region detection quality of the proposed method as well as the seven state-of-the-art methods, we accomplish an objective comparison which measures the effectiveness of the extracted saliency maps from different methods with the ground-truth mask as a criterion. Following the two binarization methods used in [27], we segment the salient region by fixed thresholding and adaptive thresholding, respectively. For a particular threshold, the precision and recall are computed based on the ground-truth mask to measure the quality and quantity of the extracted salient regions. The precision is the fraction of extracted regions that are correct, while the recall is the fraction of correct regions that are extracted.

For the fixed thresholding, the threshold is varied from 0 to 255. The curves of precision versus recall for different saliency maps are shown in Fig. 6. It is seen that the proposed method shows the highest precision for most of the recall values, which proves that the saliency maps obtained by the proposed method have the best performance in highlighting salient regions in the MSRA image dataset. As shown in Fig. 6, RC achieves the best performance among the previous methods. However, the proposed method

Table 1
The state-of-the-art methods for comparison.

Algorithm name	Reference	Implementation code
Itti's model (IT)	Itti [5]	Matlab code by Harel [30]
Frequency tuned (FT)	Achanta [27]	Matlab code by author
Graph-based visual saliency (GB)	Harel [30]	Matlab code by author
Context aware (CA)	Goferman [34]	Matlab code by author
Maximum symmetric surround (MS)	Achanta [44]	Executable code by author
Histogram-based contrast (HC)	Cheng [35]	Executable code by author
Region-based contrast (RC)	Cheng [35]	Executable code by author

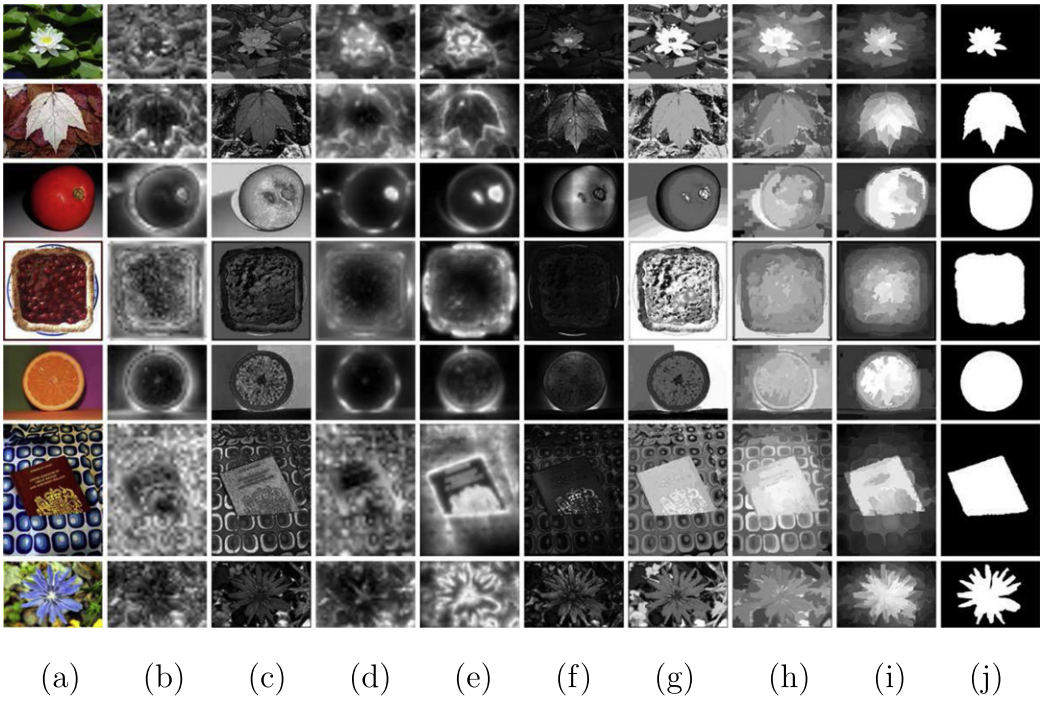


Fig. 5. Examples of saliency maps over 5000 MSRA benchmark images: (a) original images, (b)–(i) saliency maps achieved by the methods IT [5], FT [27], GB [30], CA [34], MS [44], HC [35], RC [35] and the proposed method; (j) ground truth masks.

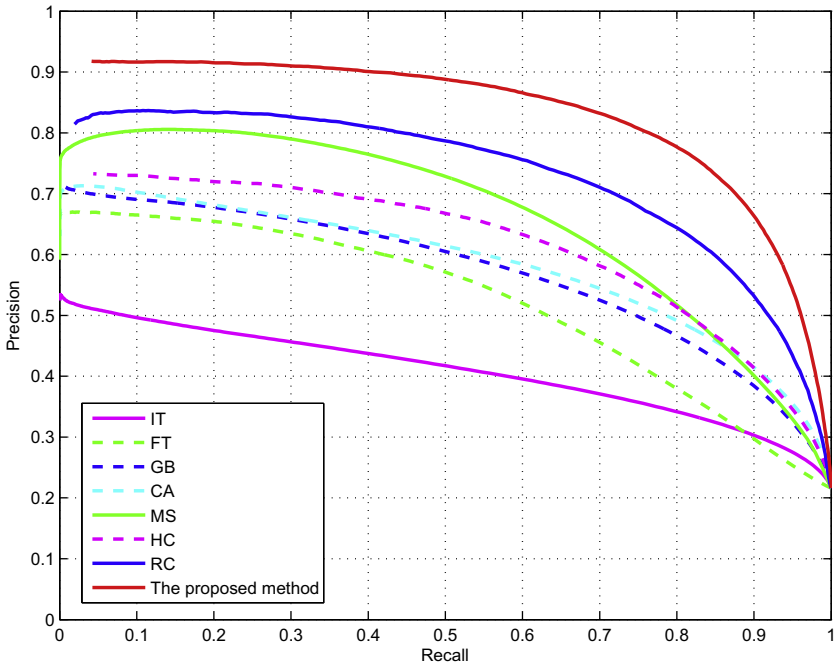


Fig. 6. Curves of precision versus recall for altering thresholds from 0 to 255 of saliency maps over 5000 MSRA benchmark images.

shows about 0.1 higher on precision for certain recall rate than RC. Furthermore, the average precision of the proposed method reaches 0.918 over a low average recall value (0.04) on the 5000 MSRA images. It means that with regard to a large thresholding, the probability to correctly detect the location of the salient object using the proposed method exceeds 90% on the MSRA dataset. Method MS achieves similar precision for low recall values with RC, but its curve drops steeply as the recall value increases. The same thing happens for the method FT, while its precision is low for small recall values. The main reason is that more non-salient regions in the backgrounds are extracted as the salient for these methods when the thresholding exceeds a certain value. The proposed method based on joint spatial-color constraint and multi-scale segmentation can solve this problem successfully.

For another objective comparison, we utilize an adaptive threshold as in [27,33] to extract the salient object. The adaptive threshold value T_a is determined as two times the mean saliency value of a saliency map, which can be represented as

$$T_a = \frac{2}{N}\sum_{\forall p \in I} Sal(p) \tag{19}$$

where $Sal(p)$ is the saliency value of pixel p in the saliency map. Using this adaptive threshold, the binary saliency maps from the compared methods are obtained. Then, over the ground-truth dataset, we obtain the average values of precision, recall and F-Measure which is defined as

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{20}$$

We set $\beta^2 = 0.3$ as in [27] to weight precision more than recall. The comparison results are shown in Fig. 7. We can see that method IT shows poor recall and relative high precision because the goal of the method is to detect salient points instead of objects. Comparable performances are achieved for the methods FT and GB, or the methods CA, MS and HC. Method RC shows very high precision but low recall comparatively, so its F_β is not high. It is visually clear in Fig. 7 that the proposed method outperforms the existing methods with the highest precision, recall and F_β on the MSRA dataset.

4.3. Experiments on eye tracking dataset

We also evaluate the proposed method on the eye tracking image dataset provided by Bruce and Tsotsos in [23]. This popular dataset is generally used to evaluate the performance of visual saliency prediction. There are 120 color images including indoor and outdoor scenes in the dataset along with eye fixation data from 20 different subjects. The proposed method is also compared with the seven state-of-the-art methods in Table 1.

Fig. 8 shows the subjective comparison between saliency maps obtained by different methods and the fixation density maps generated from the sum of all 2D Gaussians corresponding to each fixation point. The comparison results show that the proposed

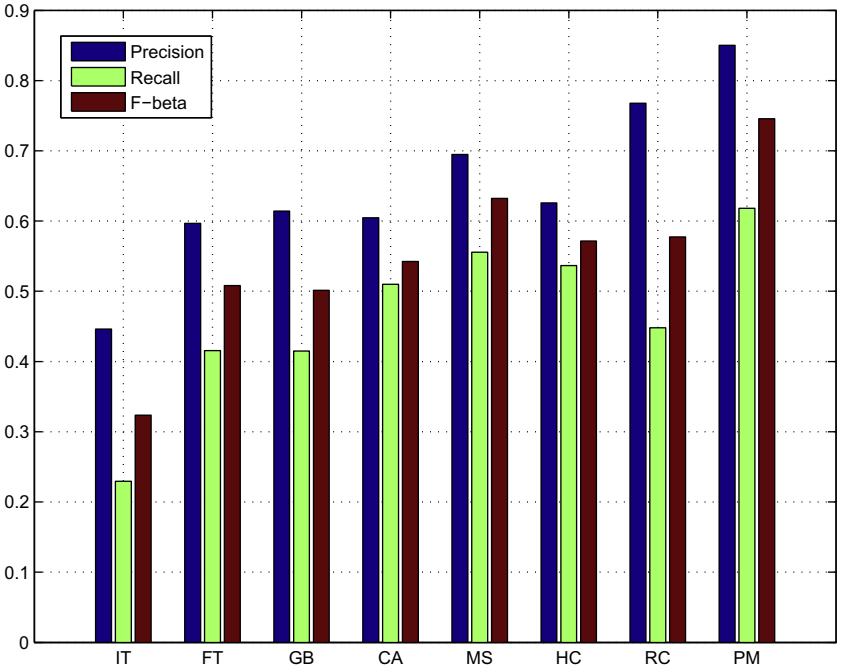


Fig. 7. Precision-recall bars of binarized saliency maps using adaptive thresholds over 5000 MSRA benchmark images (PM: the proposed method).

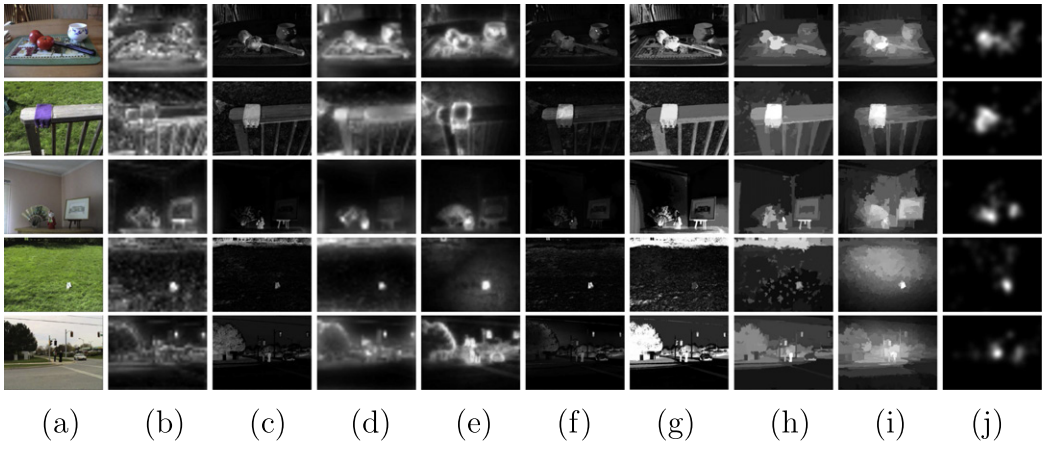


Fig. 8. Results for a subjective comparison between the proposed method and the seven state-of-the-art methods on the eye tracking dataset provided by Bruce and Tsotsos. (a) Original images. (b)-(i) Saliency maps achieved by the methods IT [5], FT [27], GB [30], CA [34], MS [44], HC [35], RC [35] and the proposed method. (j) The human fixation density maps.

method can not only detect the salient regions, but also predict visual saliency. The saliency maps from the proposed method are more consistent with the human fixation density maps than the maps from the other methods. Some of the previous methods are vulnerable to the background noise, such as the shadow in the fourth image, while the proposed method is more robust.

To compare the saliency maps with the human fixations objectively, we use the popular validation approach as in [23]. The area under the receiver operator characteristic (ROC) curve is used to evaluate the performance of visual saliency detection. We use the fixation points provided by the dataset as the ground-truth mask for all compared methods, i.e., only the points are fixations and the rest are non-fixations. The ROC areas and ROC curves are generated using the Matlab code provided by Harel et al. [30]. The results of ROC areas and curves of the compared methods are shown in Table 2 and Fig. 9, respectively. For the existing methods, methods RC and CA have the similar fixation prediction performance on this dataset, whose ROC areas are 0.6461 and 0.6307, respectively. However, the ROC area of the proposed method is about 0.10–0.12 (16–19%) larger than the two methods. We can see that the proposed method outperforms the state-of-the-art methods on predicting human fixations on this eye tracking dataset.

4.4. Discussion

In Section 4.2, we have evaluated the proposed method on the MSRA dataset of 5000 images and our hand-labeled ground-truth masks. In the field of full-resolution saliency map, the dataset of 1000 images provided in [27] is one of the most widely used. The 5000-image dataset contains all the 1000 images in order to make sure the dataset is not biased by any particular algorithm. To compare the proposed method against the published results of the existing methods [27,35], we also perform saliency detection on the public dataset [27]. The ground truth given in [27] is used to perform the objective comparison instead of our hand-labeled masks. It is seen from the curves of precision versus recall in Fig. 10(a) that the proposed method still achieves the highest precision for most of the recall values. It is also clear from the bars in Fig. 10(b) that the average precision, recall and F_β of the proposed method are the highest among all the methods. Note that the precision, recall and F_β for the precision-recall bars in Fig. 10(b) are different from the reported results in [27,35]. The discrepancy is due to the employed segmentation technique to refine the initial binary mask, such as mean-shift in [27] and saliency cut in [35]. However, in our experiments, we only use two times the mean saliency values of a saliency map to obtain a binary map. The compared methods were all evaluated on the same validation approach, so their relative performance should not be affected.

Several techniques, i.e., pixel-level saliency of SC, CD and SD, two-layer saliency structure and multi-scale segmentation, have been combined in the proposed method for saliency detection, which outperforms the existing methods. In order to further exploit the contribution from each of these techniques, we plot the precision-recall curves when using the techniques individually or combined together on the 5000 MSRA benchmark images in Fig. 11. CD saliency performs not as good as the other two pixel-level techniques. SD saliency achieves higher precision than SC saliency when the recall value is below 0.3. However, SC saliency outperforms SD saliency when the recall value grows above 0.3. When using the two-layer saliency structure to combine the three saliency components, an overt performance enhancement is obtained. For example, at the recall value of 0.2, the combined saliency gains about 5.9% performance enhancement over the SC saliency. Furthermore, when the

Table 2
The ROC areas on the eye tracking dataset provided by Bruce and Tsotsos. The number in bold shows the best method which achieves the maximal ROC area.

Method	IT [5]	FT [27]	GB [30]	CA [34]
ROC area	0.5709	0.4851	0.5237	0.6307
Method	MS [44]	HC [35]	RC [35]	PM[a]
ROC area	0.6107	0.5008	0.6461	**0.7499**

[a] PM: the proposed method.

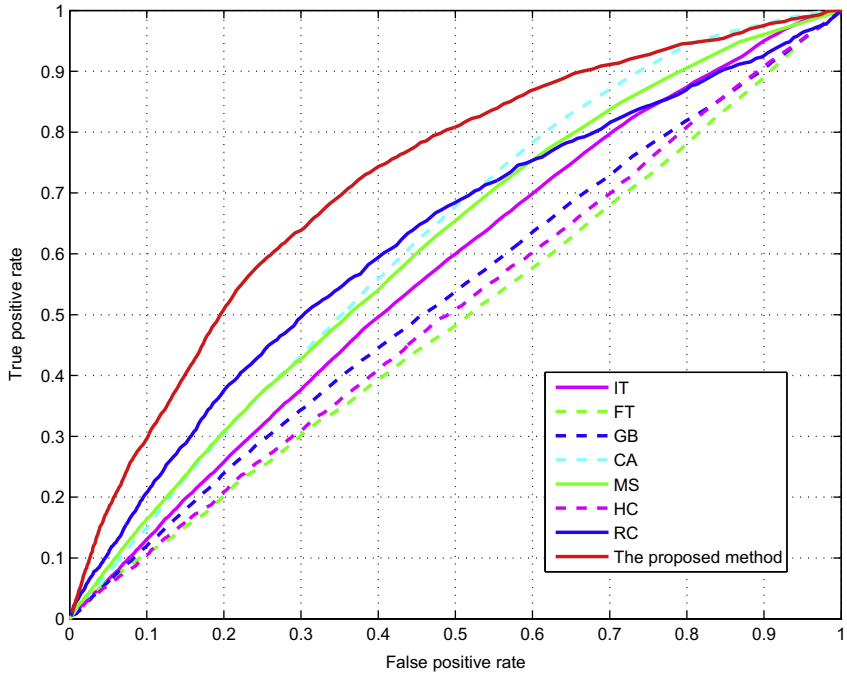


Fig. 9. The ROC curves of the existing methods and the proposed method on the eye tracking dataset provided by Bruce and Tsotsos.

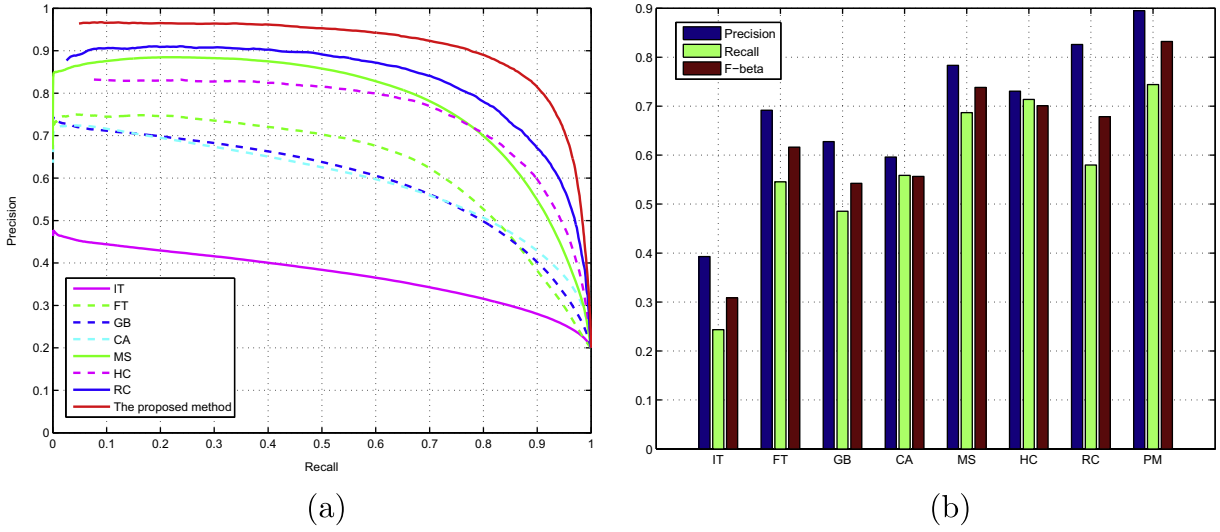


Fig. 10. Comparison results for 1000 MSRA benchmark images in [27,35]: (a) curves of precision versus recall for the fixed thresholding and (b) precision-recall bars for the adaptive thresholding.

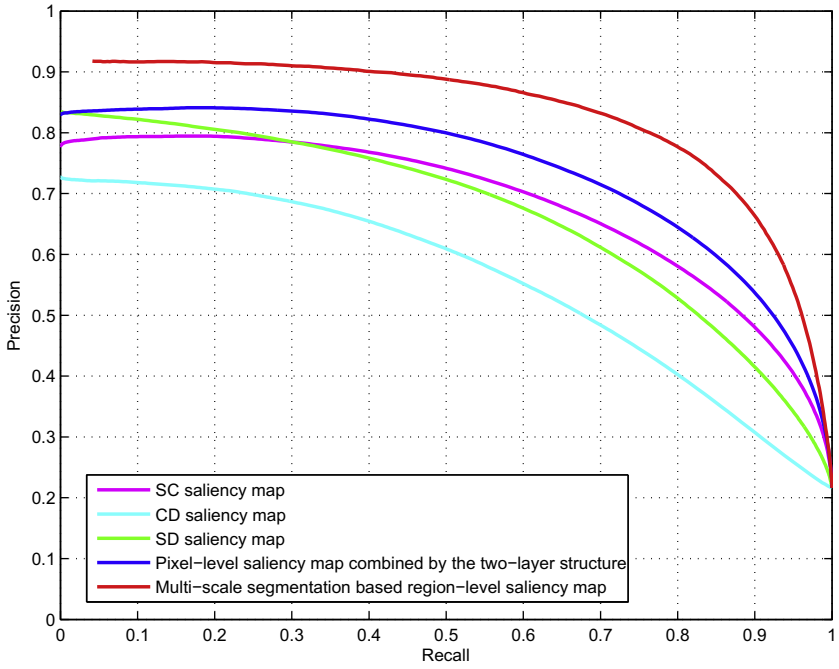


Fig. 11. Curves of precision versus recall over 5000 MSRA benchmark images for different techniques proposed in this paper.

multi-scale segmentation technique is adopted to generate the region-level saliency, a great leap in performance is performed.

The proposed method is based on the global contrast which integrates the entire information features all over the visual field to extract the salient object. So, the method is able to achieve reasonable results, even in the conditions that the salient object is not near the center or there are multiple salient objects in the image as shown in Fig. 12. Although the proposed method performs well in the experiments described in Section 4.2 and 4.3, it may fail to locate the salient object if the appearance of the foreground and background is similar as shown in Fig. 13. Generally, this issue is challenging for the bottom-up saliency detection models which focus on regions possessing distinct low-level features (intensity, color etc.) without object prior and task information. One way to solve this problem is to use the top-down or task-orientated cues.

The proposed method has several parameters to set, which are from the empirical study. All the parameters are not changed during the evaluation on the two datasets. The superior results of the proposed method show the effectiveness of the parameter setting. Also, we can automatically set the spatial distance related parameters according to the resolution, e.g., the parameter σ_{s1}^2 in (3) can be set to $0.75\max(w,h)$, where w and h represent the width and height of the image, respectively. For future research, we plan to use the learning based method to optimize these parameters.

In terms of the computational time of the proposed method, the average running time on the 5000 MSRA images to produce the

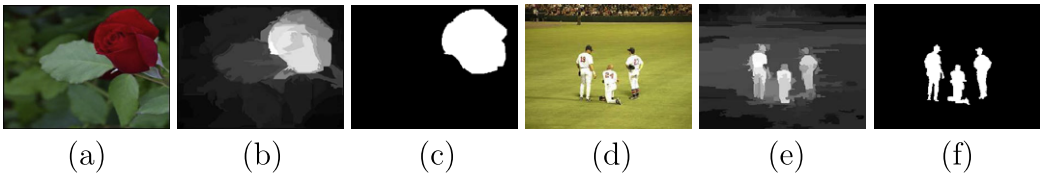


Fig. 12. Two cases: the object is at the corner and there are multiple salient objects, respectively: (a, d) original images, (b, e) saliency maps achieved by the proposed method, (c, f) ground truth masks.

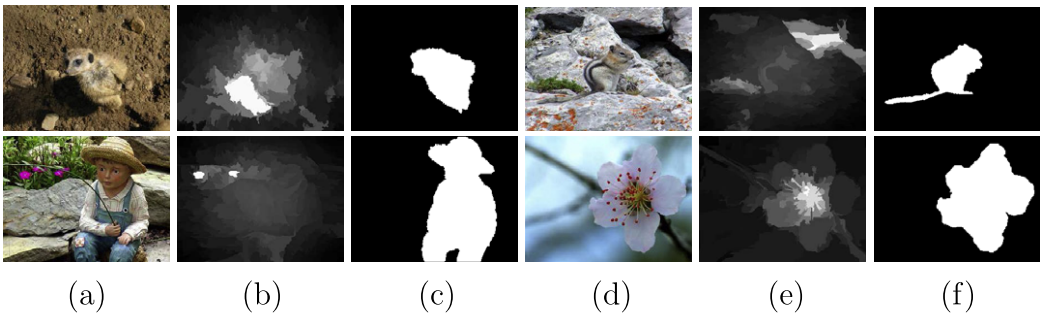


Fig. 13. Failure case: the appearance of the foreground and background is similar: (a, d) original images, (b, e) saliency maps achieved by the proposed method, (c, f) ground truth masks.

pixel-level saliency map is 1.47 s when measured on an Intel 3.20 GHz CPU with 3 GB RAM in Matlab implementation. As for the running time for generating a region-level saliency map, it mainly depends on the efficiency of the adopted segmentation algorithm when the multi-scale segmentation is performed.

5. Conclusion

In this paper, we present a novel method to detect salient regions in images. To overcome the area-depending deficiencies of some state-of-the-art saliency detection methods, the proposed method generates pixel-level saliency maps using joint spatial-color constraint. Firstly, we introduce a spatial constraint based saliency to produce the global contrast in CIE L*a*b* color space, which uses the Gaussian weighting function from the center of the image to distinguish the difference between "center and surround". Secondly, based on the physiological discovery, the color double-opponent saliency is obtained by computing the global contrast of red-green and blue-yellow double opponency. Thirdly, we use the similarity distribution based saliency to detect the salient object and its background. Finally, a two-layer structure is adopted to merge the three components into a pixel-level saliency map. To produce the region-level saliency map, we apply a multi-scale segmentation based method, which can overcome the weakness of inconsistency caused by single-scale over-segmentation. The proposed method is evaluated on the MSRA image dataset and eye tracking dataset. Experimental results show that the proposed method outperforms the state-of-the-art methods on both salient region detection and human fixation prediction with respect to precision, recall and ROC area.

Acknowledgments

This work was partially supported by NSFC (No. 61271289 and 61179060), National High Technology Research and Development Program of China (863 Program, no. 2012AA011503), and The Ph.D. Programs Foundation of Ministry of Education of China (No. 20110185110002).

References

[1] W. James, The Principles of Psychology, Dover Publications Inc., New York, 1950.
[2] R. Desimone, J. Duncan, Neural mechanisms of selective visual attention, Ann. Rev. Neurosci. 18 (1) (1995) 193–222.
[3] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cognit. Psychol. 12 (1) (1980) 97–136.
[4] E. Niebur, C. Koch, Computational architectures for attention, in: R. Parasuraman (Ed.), The Attentive Brain, MIT Press, Cambridge, MA, 1998, pp. 163–186.
[5] 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.
[6] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong, S. Yao, Rate control for videophone using local perceptual cues, IEEE Trans. Circuits Syst. Video Technol. 15 (4) (2005) 496–507.
[7] Z. Chen, J. Han, K.N. Ngan, Dynamic bit allocation for multiple video object coding, IEEE Trans. Multimedia 8 (6) (2006) 1117–1124.
[8] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image Vision Comput. 29 (1) (2011) 1–14.
[9] W. Lin, C.-C.J. Kuo, Perceptual visual quality metrics: A survey, J. Vis. Commun. Image Represent. 22 (4) (2011) 297–312.
[10] J. Han, K.N. Ngan, M. Li, H.J. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Trans. Circ. Syst. Video Technol. 16 (1) (2006) 141–145.
[11] H. Li, K.N. Ngan, Saliency model-based face segmentation and tracking in head-and-shoulder video sequences, J. Vis. Commun. Image Represent. 19 (5) (2008) 320–333.
[12] Q. Zhang, K.N. Ngan, Multi-view video based multiple objects segmentation using graph cut and spatiotemporal projections, J. Vis. Commun. Image Represent. 21 (5–6) (2010) 453–461.
[13] H. Li, K.N. Ngan, Learning to extract focused objects from low dof images, IEEE Trans. Circuits Syst. Video Technol. 21 (11) (2011) 1571–1580.
[14] A. Shokoufandeh, I. Marsic, S.J. Dickinson, View-based object recognition using saliency maps, Image Vision Comput. 17 (5–6) (1999) 445–460.
[15] 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.
[16] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10), Miami Beach, FL, USA, 2010, pp. 1975–1981.
[17] V. Setlur, T. Lechner, M. Nienhaus, B. Gooch, Retargeting images and video for preserving information saliency, IEEE Comput. Graph. Appl. 27 (5) (2007) 80–88.
[18] M. Guttmann, L. Wolf, D. Cohen-Or, Content aware video manipulation, Comput. Vision Image Understand. 115 (12) (2011) 1662–1678.
[19] M. Mirmehdi, R. Perissamy, Perceptual image indexing and retrieval, J. Vis. Commun. Image Represent. 13 (4) (2002) 460–475.

[20] J. Lai, Y. Yi, Key frame extraction based on visual attention model, J. Vis. Commun. Image Represent. 23 (1) (2012) 114–125.
[21] L. Itti, Visual salience, Scholarpedia 2 (9) (2007) 3327.
[22] C. Koch, S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, Human Neurobiol. 4 (1985) 219–227.
[23] N. Bruce, J.K. Tsotsos, Saliency based on information maximization, Adv. Neural Informat. Process. Syst. 18 (2006) 155–162.
[24] L. Itti, P. Baldi, Bayesian surprise attracts human attention, Adv. Neural Inform. Process. Syst. 19 (2006) 547–554.
[25] H. Li, K.N. Ngan, A co-saliency model of image pairs, IEEE Trans. Image Process. 20 (12) (2011) 3365–3375.
[26] D. Gao, N. Vasconcelos, Bottom-up saliency is a discriminant process, in: Proceedings of 2007 IEEE 11th International Conference on Computer Vision (ICCV'07), Rio de Janeiro, Brazil, 2007.
[27] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami Beach, FL, USA, 2009, pp. 1597–1604.
[28] Y.-F. Ma, H.-J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: Proceedings of the Eleventh ACM International Conference on Multimedia (MULTIMEDIA'03), Berkeley, CA, USA, 2003, pp. 374–381.
[29] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395–1407.
[30] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, Adv. Neural Inform. Process. Syst. 19 (2006) 545–552.
[31] T. Liu, J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, in: Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, Minnesota, USA, 2007.
[32] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: Proceedings of the 14th ACM international conference on Multimedia (MULTIMEDIA'06), Santa Barbara, CA, USA, 2006, pp. 815–824.
[33] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, Minnesota, USA, 2007.
[34] S. Geferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, in: Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10), Miami Beach, FL, USA, 2010, pp. 2376–2383.
[35] M.-M. Cheng, G.-X. Zhang, N.J. Mitra, X. Huang, S.-M. Hu, Global contrast based salient region detection, in: Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11), Colorado Springs, CO, USA, 2011, pp. 409–416.
[36] J. Feng, Y. Wei, L. Tao, C. Zhang, J. Sun, Salient object detection by composition, in: Proceedings of 2011 IEEE 13th International Conference on Computer Vision (ICCV'11), Barcelona, Spain, 2011, pp. 1028–1035.
[37] C. Rother, L. Bordeaux, Y. Hamadi, A. Blake, Autocollage, ACM Trans. Graph. 25 (3) (2006) 847–852.
[38] S. Engel, X. Zhang, B. Wandell, Colour tuning in human visual cortex measured with functional magnetic resonance imaging, Nature 388 (6637) (1997) 68–71.
[39] H. Li, L. Xu, G. Liu, Two-layer average-to-peak ratio based saliency detection, Signal Processing: Image Communication 28 (1) (2013) 55–68.
[40] O. Muratov, P. Zontone, G. Boato, F.G.B.D. Natale, A segment-based image saliency detection, in: Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'11), Prague, Czech Republic, 2011, pp. 1217–1220.
[41] K. He, J. Sun, X. Tang, Guided image filtering, in: Proceedings of 11th European Conference on Computer Vision (ECCV'10), Part I, Heraklion, Crete, Greece, 2010, pp. 1–14.
[42] B.W. 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) 1–17.
[43] P. Felzenszwalb, D. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vision 59 (2) (2004) 167–181.
[44] R. Achanta, S. Süsstrunk, Saliency detection using maximum symmetric surround, in: Proceedings of 2010 17th IEEE International Conference on Image Processing (ICIP'10), Hong Kong, 2010, pp. 2653–2656.