Salient Object Detection via Adaptive Region Merging

Jingbo Zhou[1,3], Jiyou Zhai[2], Yongfeng Ren[1,3]
[1]Nanjing University of Information Science & Technology (NUIST),
Nanjing, 210044, P. R. China
[e-mail:jbzhou2013@aliyun.com]
[2]College of Computer and Information, Hohai University,
Nanjing, 211100, P. R. China
[e-mail:jyzhai@163.com]
[3]Faculty of Computer Engineering, Huaiyin Institute of Technology,
Huai'an, 223003, P. R. China
*Corresponding author: Jingbo Zhou

Received December 2, 2015; accepted March 30, 2015;
published September 30, 2016

Abstract

Most existing salient object detection algorithms commonly employed segmentation techniques to eliminate background noise and reduce computation by treating each segment as a processing unit. However, individual small segments provide little information about global contents. Such schemes have limited capability on modeling global perceptual phenomena. In this paper, a novel salient object detection algorithm is proposed based on region merging. An adaptive-based merging scheme is developed to reassemble regions based on their color dissimilarities. The merging strategy can be described as that a region R is merged with its adjacent region Q if Q has the lowest dissimilarity with Q among all Q's adjacent regions. To guide the merging process, superpixels that located at the boundary of the image are treated as the seeds. However, it is possible for a boundary in the input image to be occupied by the foreground object. To avoid this case, we optimize the boundary influences by locating and eliminating erroneous boundaries before the region merging. We show that even though three simple region saliency measurements are adopted for each region, encouraging performance can be obtained. Experiments on four benchmark datasets including MSRA-B, SOD, SED and iCoSeg show the proposed method results in uniform object enhancement and achieve state-of-the-art performance by comparing with nine existing methods.

Keywords: Salient object detection, Image segmentation, Adaptive region merging

This work is sponsored by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 14KJB520006), the CICAEET fund and the PAPD fund.

http://dx.doi.org/10.3837/tiis.2016.09.020 **ISSN : 1976-7277**

1. Introduction

Saliency detection has attracted much attention recently due to its ability to model the human visual attention mechanism, which has its roots in psychology but has been a topic of research in diverse areas such as neuroscience, robotics and computer vision [1,2]. Early efforts of saliency detection aimed to predict the locations of human eye fixations and introduced the fundamental principles of saliency detection. Different from eye fixations prediction, a new sub-field in saliency detection called salient object detection has recently emerged and drawn a lot of research attentions. It aims at compensating the drawback of previous eye fixation prediction models [3-6] on enhancing entire objects. There are two main approaches to salient region detection – top-down and bottom up, where the former is task dependent while the latter seeks to identify pop-out features that enable the extraction of distinct regions in an image. Bottom up saliency models have been developed as a pre-processing step to prioritize the search space for object detection tasks reducing the computational overhead [7]. Top-down approaches include [8] for scene recognition and [9] for tracking. Saliency detection has also been used as a pre-processing step for active segmentation of the objects in point clouds for manipulative tasks in robotics.

Previous salient region detection methods [10-15], which commonly employed segmentation techniques include superpixels [16] and mean shift [17] or graph-based segmentation [18], exploit contrast and rarity properties on local superpixels or regions. These techniques are known to be useful for eliminating background noise and reducing computation by treating each segment as a processing unit. However, individual small segments provide little information about global contents. Such models have limited capability on modeling global perceptual phenomena [19, 20]. **Fig. 1** shows a typical example. The entire flower tends to be perceived as a single entity by human visual system. However, local-segment based algorithms (e.g. [16]) partition the flower into many parts (**Fig. 1** (b)), each of which alone does not reflect the meaning of "flower". In contrast, a coarse segmentation (derived from **Fig. 1** (b)) that attempts to keep semantic holism (**Fig. 1**(c)) which better models such gist. It is easily imagined that saliency computation with the help of such coarse segmentation is conducive to highlighting entire objects while suppressing background.

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

Fig. 1. Different segmentation for salient object detection, (a) Input image, (b) Over-segmentation, (c) Coarse segmentation, (d) Object mask

Since it is important to control segmentation to reflect proper image content, some recent approaches benefit from multiscale strategies to compute saliency on both coarse and fine scales with fusion. Yan et al. [10] define three levels of sizes for regions and merge a region to its neighbor region if it is smaller than defined sizes (Hierarchical Saliency Detection, HS for short). Despite good performance of HS, the underlying problem may be that scale parameters

in HS are crucial to performance. A salient region might not be in the proper level if it is smaller than the defined size. In addition, large background regions with close colors may not be merged together if they are larger than the defined size. Since appropriate merging may facilitate global perceptual phenomena analysis (**Fig. 1**), to find coincidence of salient object in multiscales, in this paper we propose an alternative solution, which generate varied levels by merging similar regions. Compared to HS, we use color dissimilarity and an adaptive technique during merging, while HS merges according to region size. Main advantages that lead to robust performance of the proposed method against HS include: (1) use color dissimilarity and their spatial location (rather than region size), reflecting object saliency that is often indicated by enclosed strong similarities and neighboring location; (2) use an adaptive merging strategy to obtain background and object, which help to better assist highlight objects and suppress background; (3) the number of levels in the proposed method is much larger than HS where only three scales are considered. It leads to robustness in more generic cases. In addition, our method is adaptive, i.e. no specification/manually determination of scale parameters is needed like HS.

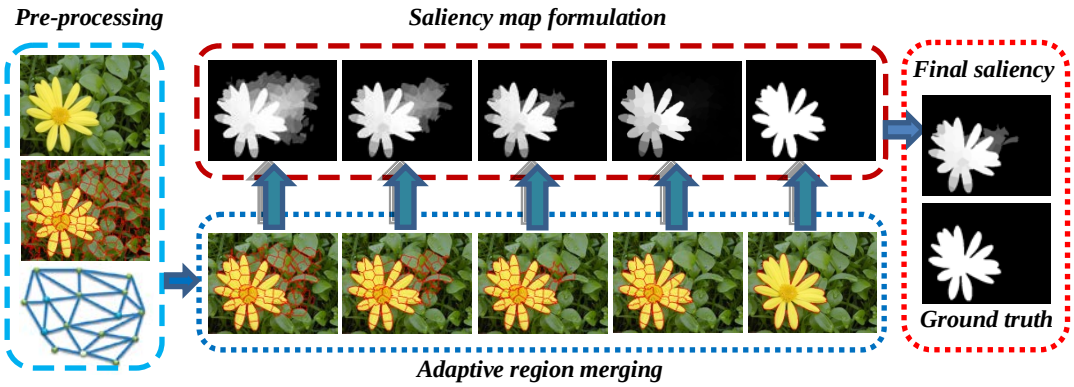


Fig. 2. The block diagram of proposed algorithm

The proposed algorithm can be divided into three stages, such as pre-processing step, adaptive region merging step and saliency measurement step, which are plotted in **Fig. 2**. In pre-processing step, an initial segmentation is required to partition the image into homogeneous regions for measuring saliency. In this paper, we use SLIC algorithm [16] for initial segmentation because it can well preserve the object boundaries. To measure the similarity between superpixels, we consider each superpixel as the node of a graph and construct the weights by color dissimilarity. Since the weights measure the similarity or dissimilarity between a node and its neighbors, we look upon the weight matrix as the similarity matrix. In adaptive region merging step, the superpixels which located at the boundary of the image are treated as the seeds to start the merging process. However, it is possible for a boundary in the input image to be occupied by the foreground object. To avoid this case, we optimize the boundary influences by locating and eliminating erroneous boundaries before the region merging, which similar to [20]. Then, an adaptive region merging, which is adaptive to the image content and it does not need to set the similarity threshold in advance, is designed in the proposed framework. A region R is merged with its adjacent region Q if Q has the lowest dissimilarity with Q among all Q's adjacent regions. In last step, each merged region is just evaluated using three simple region saliency measurements, though more complex features and measurements as in [4] can be adopted. Even though like this, we show

the proposed method already can achieve competitive results against the best methods among the state-of-the-art.

The contributions of this paper are summarized as follows:

(1) We propose a novel salient object detection algorithm which is based on adaptive region merging. Comparing to other state-of-the-art methods, no threshold is needed in region merging process. Under region merging framework, coarse segmentation is conducive to highlighting entire objects while suppressing background in salient object detection.

(2) Performance obtained is similar to other state-of-the-art methods even though simple region saliency measurements are adopted for each region.

The remainder of the paper is organized as follows: Section 2 surveys conventional salient object detection algorithms which are related to our approach. We demonstrate framework of our saliency detection method in detail in Section 3. Then, we demonstrate our experimental results based on four public image datasets and compare the results with other state-of-art saliency detection methods in Section 4. The final section concludes the paper by summarizing our findings.

2. Related Work

The following gives a review of salient object detection algorithms that are related to our approach. A comprehensive survey of salient object detection can be found from [1]. The review on visual attention modeling [2] also includes some analysis on salient object detection.

Saliency models map natural images into saliency maps, in which each image element (pixel, superpixel and region) is assigned a saliency strength or probability. A representative work by Itti et al. is presented in [4]. They proposed a biologically inspired visual attention model and built a system called neuromorphic vision C++ toolkit. Specifically, they proposed the using of a set of feature maps from three complementary channels as intensity, color, and orientation. The normalized feature maps from each channel were then linearly combined to generate the overall saliency map. Bruce and Tsotsos [21] modeled bottom-up saliency as the maximum information sampled from an image. More specifically, their saliency was computed as Shannon's self-information. Oliva and Torralba [22] proposed a Bayesian framework for the task of visual search (i.e., whether a target is presented or not.). Zhang et al. [23] also proposed saliency detection using natural statistics (SUN) based on a similar Bayesian framework to estimate the probability of a target at every location. They also claimed that their saliency measure emerges from the use of Shannon's self-information under certain assumptions. Most of the methods [4, 22-24] based on Gabor or DoG filter responses required many design parameters such as the number of filters, type of filters, choice of the nonlinearities, and a proper normalization scheme. Different from the traditional image statistical models, spectral saliency approaches that operate on the Fourier and cosine (frequency) spectrum have attracted a huge interest [3]. Guo et al. [25] claimed that what plays an important role for saliency detection is the image's phase spectrum. Hou et al. [26] proposed discrete cosine transform (DCT) image signature approach, which defines the saliency using the inverse DCT of the signs in the cosine spectrum.

Early efforts aimed to predict the locations of human eye fixations and introduced the fundamental principles of saliency detection. While these methods were purely bottom-up (stimulus driven), some researchers proposed models of top-down saliency. These use high-level information to guide the saliency computation, equating saliency to the detection of stimuli from certain object classes [11]. While bottom-up saliency predicts eye fixations, these

methods are useful for high-level vision tasks, such as object recognition and localization. However, they require the definition of object classes of interest and are, thus, less generic. A compromise is to either consider the responses of object detectors as features for bottom-up saliency [27], or to formulate saliency as a mid-level vision task. In this area, substantial attention has been devoted to the problem of object saliency [2]. The goal is no longer to predict eye fixations or the locations of specific object classes, but to identify the locations of salient objects, independent of what these may be [28].

Methods in this class are generally based on the earlier principles of bottom-up saliency, e.g. local measures of center-surround contrast [11, 28-29]. Center-surround prior is realized as a Gaussian fall-off map. It is either directly combined with other cues as weights [28], or used as a feature in learning-based methods [11]. This makes strict assumptions about the object size and location in the image. Achanta et al. [30] exploited the central-surround principle by defining the saliency of a pixel as its distance to the average image. Cheng et al. [12] considered a regional contrast based saliency extraction algorithm with mid-level cues which simultaneously evaluates global contrast differences and spatial coherence. From an opposite perspective, recent works [20, 31] introduce boundary prior and treat image boundary regions as background. In [31], the contrast against the image boundary is used as a feature in learning. In [15], saliency estimation is formulated as a ranking and retrieval problem and the boundary patches are used as background queries. These approaches work better for off-center objects but are still fragile and may fail even when an object only slightly touches the boundary. In [32], Goferman et al. proposed a context-aware saliency algorithm to detect the image regions that represent the scene based on four principles of human visual attention.

Most methods implement and combine low level cues heuristically. Recently, a few approaches have adopted more principled global optimization. The work in [33] treats salient objects as sparse noises and solves a low rank matrix recovery problem instead. The work in [34] models salient region selection as the facility location problem and maximizes the sub-modular objective function. These methods adapt viewpoints and optimization techniques from other problems for saliency estimation. In [35], multiple saliency maps from different methods are aggregated into a better one.

It is also worth noting that there are some previous works involving both graph cut and saliency detection [12-13, 36]. Those methods differ from ours as they treat the two steps separately. Saliency detection is conducted first and resulting saliency maps are then used to generate “seeds” or “initial regions” to guide graph cut. The outcome of graph cut is a binary segmentation map. For example in [36], seed regions are generated by saliency detection in [3] and then “MaxFlow” is applied to solve the min-cut problem. In contrast, our saliency detection is induced by the superpixels. According to the regions generated by superpixel algorithm, we merge the similar superpixels into a new region. Thereby in our method the saliency detection takes place prior to the merging regions. In [12], [13], and [36], the results of graph cut highly depend on saliency maps that provide “seeds”, cut performance could suffer from a less accurate saliency map that is derived from less good grouping.

3. Salient Object Detection via Region Merging

This section details the proposed method for salient object detection. Firstly, the graph construction is discussed. Then, an adaptive merging method is described that is used to generate coarse segmentation by merging the similar regions according to their color dissimilarity. Last, regional saliency measures are introduced which describes the formulation of saliency map.

3.1 Graph Construction

Before graph construction, our framework first performs over-segmentation on an input image by using SLIC superpixels. The result is a set of compact superpixels that are homogenous in color and maintain image boundaries. N=200 superpixels are selected for each input image since such number of superpixels suffices for detecting salient objects. Let a graph $G =< V, E >$ be defined where vertices V are, and E are graph edges. Let $G =< V, E >$ be an undirected graph, where $v_i \in V$ is a set of nodes corresponding to superpixels. E is a set of edges connecting the pairs of neighboring nodes. Each edge $(v_i, v_j) \in E$ has a corresponding weight $w((v_i, v_j))$ to measure the dissimilarity of the two nodes connected by that edge. In the proposed algorithm, we consider the color dissimilarity between regions since color statistics is an important attributes of image region. Specially, considering image pixels I_i and I_j, the dissimilarity is defined as

$$d(I_i, I_j) = D(I_i, I_j) \tag{1}$$

where $D(I_i, I_j)$ is the color distance metric between pixels I_i and I_j in the $CIE\ L*a*b*$ space. Suppose that $I_j \in R_B$ where R_B is a region, the average dissimilarity between pixel I_i and region R_B is defined as follows

$$d(I_i, R_B) = \frac{1}{|R_B|} \sum_{I_j \in R_B} D(I_i, I_j) \tag{2}$$

where $|R_B|$ is the number of pixel in region R_B. It is easy to see that pixels with the same color value have the same dissimilarity value under this definition, since the measure is oblivious to spatial relations. Hence, rewriting equation (2) such that the terms with the same color value c_j are grouped together, we get dissimilarity value for each color as

$$d(I_i, R_B) = d(c_i) = \sum_{j=1}^{n_B} p_j \times D(c_i, c_j) \tag{3}$$

where c_i is the color value of pixel I_i, n_B is the number of distinct pixel colors in region R_B, and p_j is the probability of pixel color c_j in region R_B. If $I_i \in R_A$, the dissimilarity between regions R_A and R_B can be written as

$$\begin{aligned} d(R_A, R_B) &= \sum_{I_i \in R_A} \sum_{I_j \in R_B} D(I_i, I_j) \\ &= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} p(c_{A,i})\, p(c_{B,j}) D(c_{A,i}, c_{B,j}) \end{aligned} \tag{4}$$

where $p(c_{s,i})$ is the probability of the i-th color $c_{s,i}$ among all n_s colors in the s-th region R_s, $s = \{A, B\}$. According to equation 4, it is easy to use the color histogram to represent the regions R_A and R_B since color histogram is more robust than the other feature descriptors. The dissimilarity of regions R_A and R_B are measured by computing the difference of the corresponding color histogram. Note that we use the probability of a color in the probability density function (i.e. normalized color histogram) of the region as the weight for this color to emphasize more the color differences between dominant colors.

Since color histogram is an effective method to represent the regions, it is computationally too expensive even for medium sized images. To make our algorithm efficiency, similar to [12], we compute color dissimilarity by building a compact color histogram using color quantization and choosing more frequent colors. For detailed information about histogram based speed up and color space smoothing, please refer to [12].

3.2 Region merging strategy

Based on the graph constructed aforementioned, the next step is to merge the regions with similar characteristics. Regions with similar color are merged on the assumption that they belong to the same object or the background. The conventional region merging methods merge two adjacent regions whose similarity is above a preset threshold [37]. These methods have difficulties in adaptive threshold selection. A big threshold will lead to incomplete merging of the regions belonging to the object, while a small threshold can easily cause over-merging, i.e., some object regions are merged into the background. Moreover, it is difficult to judge when the region merging process should stop. In this paper, we present an adaptive minimal dissimilarity based mechanism to merge the superpixels.

Suppose that Q is an adjacent region of R and denote by $S_Q = \{S_i^Q\}_{i=1,2,\cdots,q}$ the set of Q's adjacent regions. The color dissimilarity between Q and all its adjacent regions, i.e. $d(Q, S_i^Q)$, $i = 1,2,\cdots,q$ are calculated. Obviously, R is a member of S_Q. If the dissimilarity between R and Q is the minimal one among all the color dissimilarities $d(Q, S_i^Q)$, we will merge R and Q. The following merging rule is defined as:

$$Merge\ \ R\ and\ \ Q \quad if\ d(R,Q) = \min_{i=1,2,\cdots,q} d(R, S_i^Q) \tag{5}$$

The merging rule (5) is very simple but it establishes the basis of the proposed region merging process. One important advantage of (5) is that it avoids the presetting of similarity threshold for merging control. However, to merge the regions by region growing, we must choose some regions as the seeds to start. Since different seed make finally result different, it is hard that there is no prior information in the input images. Inspired by [15], we define a two-stage merging mechanism in which the superpixels located in the boundary of the image as the priors of the background at the first stage to guide the merging process.

In the first stage, we try to merge the regions, which located in the boundary of the given image, with their adjacent regions. Specially, we denote the regions that near the boundary of the image as M_B. For each region $B \in M_B$, we form the set of its adjacent regions $S_B = \{A_i\}_{i=1,2,\cdots,r}$. Then for each A_i and $A_i \notin M_B$, we form it's set if adjacent regions

$S_{A_i} = \{S_j^{A_i}\}_{j=1,2,\cdots,k}$ .it is obvious that $B \in S_{A_i}$. The color dissimilarity between A_i and each element S_{A_i} is calculated. If B and A_i satisfy the equation (5), then B and A_i are merged into one region:

$$B = B \cup A_i \tag{6}$$

The above procedure is iteratively implemented. Note that in each iteration, the sets of background regions will be updated. Specifically, M_B expands. The iteration stops when the entire background regions M_B will not find new merging regions.

As stated in [11], it is possible for a boundary in the input image to be occupied by the foreground object. Using such a problematic boundary as the seeds in adaptive region merging may lead to undesirable results, and a typical example is illustrated in the line of **Fig. 3**. In such case, the object is easy to merging into the background, which makes the subsequent results inaccurate in salient object detection. Similar to [20], we therefore optimize the boundary influences by locating and eliminating erroneous boundaries before the background merging (as shown in **Fig. 3** (second line)). The major advantage of erroneous boundary removal is that it helps to relieve the inaccuracy of using all boundaries in cases that one or more of the boundaries happen to be adjacent to the foreground object. Removal of the most irrelevant boundary leads to more accurate outputs.

Fig. 3. Intermediate results of the region merging process by boundary prior (first line) and optimized boundary prior (second line), (a) (b) Intermediate results of first merging round, (c) the final results of first merging round, (d) the results of second merging round

After the first stage merging, some background regions will be merged with the corresponding background seeds. However, there are some regions survived which cannot be merged because they have higher similarity with each other than with the background regions (i.e., regions of the object). For each survived region P , we form the set of its adjacent regions $S_P = \{H_i\}_{i=1,2,\cdots,k}$. Then for each H_i that $H_i \notin M_B$, we form its set of adjacent regions $S_{H_i} = \{S_j^{H_i}\}_{j=1,2,\cdots,k}$. The color dissimilarity between H_i and P is calculated. If H_i and P satisfy the rule (5), i.e.,

$$d(P, H_i) = \min_{i=1,2,\cdots,k} d(H_i, S_j^{H_i}) \tag{7}$$

then P and H_i are merged into one region. Otherwise, P and H_i will not merge.

This process is terminated when there is no region survived. After the second-stage of region merging process is complete, there are some trivial regions, which are inconsistency with the background and the objects, cannot be merged into any large regions. We check these regions and find that either they are the noise in the background (object) or they are the boundary of the object. In former case, the region is surrounded by the background (object) and can be merged into background (object); in latter case, the region is located in the middle of the background and object, and can be merged into neither background nor object.

3.3 Region Saliency Measurements

To show the effectiveness of the proposed region merging and integration scheme, each merged region is just evaluated using three simple region saliency measurements, i.e., figure-ground contrast, center bias and boundary cropping. Even though like this, we show the proposed method already can achieve competitive results against the best methods among the state-of-the-art.

Let R_i be a region at the region merging process. We propose the following regional saliency measures for R_i.

Figure-Ground Contrast: We compute the figure-ground contrast by comparing a region's color distance to all boundary superpixels. As a merged region constitutes of a set of superpixels, the problem boils down to the comparison between two superpixel sets, and is defined as:

$$S_i^{fg} = \frac{\sum_{R_j \in \mathrm{M_B}} \left\| d(R_i, R_j) \right\|}{\left| R_i \right| \cdot \left| \mathrm{M_B} \right|} \tag{8}$$

where M_B represents the set containing all boundary superpixels. Notation $|\cdot|$ indicates the number of elements in the set, i.e., the number of superpixels. Different from the previous regional contrast hypothesis [12], here we only compare a region with a potential background, i.e. boundary superpixels according to the verified boundary hypothesis [15, 31]. This is more efficient to compute for regions in different levels as boundary set M_B is always fixed.

Center Bias: Statistical results in [2] shows that human attention is center biased, indicating that distinctive regions close to image center is likely to be salient [10, 33]. Therefore, the mask with a Gaussian distribution $G(p)$ is applied at the image center, and the average probability value lying in each region is computed:

$$S_i^{cb} = \frac{\sum_{j \mid R_j \in R_i} G(p_j)}{\left| R_i \right|} \tag{9}$$

where $G(p_j)$ corresponds to Gaussian value of location p_j. Although it has been argued in [31] that boundary hypothesis is more generic than the center prior, we still find the latter useful when there are multiple regions disconnected from image boundary but scattered in the whole image.

Boundary Cropping: Boundary hypotheses [15, 31] imply that regions touching image borders are likely to be background. This phenomenon can be explained by the "surroundedness" in Gestalt laws [38]: a region with a complete/closed contour is likely to be perceived as figure. We simply incorporate this cue by cropping saliency of regions according to numbers of image borders they touch (suppose an image has four borders), defined as:

$$S_i^{bc} = \begin{cases} 1 & if\ l_i \le 1 \\ 0 & otherwise \end{cases} \tag{10}$$

where l_i is the number of image borders that R_i touches. Equation (10) implies that a region cropped by more than one image borders will be suppressed in the computed intermediate saliency map. This measurement can maintain objects that touch none or one border such as the half-length portrait in photography.

Combination of Regional Saliency Measures: Since salient regions are assumed to achieve high scores under all three metrics above, linear combination or multiplication can be considered. Similar to [15], we chose multiplication as good background suppression is observed. Furthermore, S_i^{bc} can effectively suppress image boundary-touching regions if the multiplication is used. Hence, the final saliency score for the region R_i is defined as:

$$S_i = S_i^{fg} \cdot S_i^{bc} \cdot S_i^{cb} \tag{11}$$

where S_i^{fg} S_i^{cb} S_i^{bc} respectively denotes the "figure-ground" contrast saliency, "center bias" saliency, and "boundary cropping" saliency. This regional saliency score is further assigned to the corresponding superpixels and pixels in the image to formulate intermediate saliency maps.

4. Experimental Results and Analysis

In this section, we evaluate the performance of our proposed algorithm over several datasets that are widely used in previous works, e.g. [2, 11, 12]. Next, we describe the datasets shortly and report both quantitative and qualitative comparisons of our approach with state-of-the-art approaches in detail. To save space, we compare our method with several prior ones, including SVO [39], PCAS [40], RC [12] and DRFI [11], which are the top four models or their improvements in survey [2]. In addition, we also consider well-known methods, such as CA [32], FT [30], HS [10], LRMR [33] and MR [15], that are not covered in [2]. For all methods aforementioned, we run the codes from author's website accordingly and use the results for fair comparison. We have not compared with eye fixation models such as Itti's [4] and Hou's [3] due to different purposes of the methods.

4.1 Datasets and evaluation measures

The proposed method is evaluated on four publicly-available datasets with a ground truth in the forms of accurate human-marked labels for the salient regions. (1) MSRA-B [28] includes 5000 images, originally containing labeled rectangles from nine users drawing a bounding box around what they consider the most salient object. There is a large variation among images including natural scenes, animals, indoor, out-door, etc. We use the salient object (contour) in

[11] as binary masks. The ASD dataset [30] is a subset (binary masks are provided) of the MSRA-B, and thus we no longer make the evaluation on it. (2) SED [41] contains two subsets, the first of which is a single-object database (SED1) with 100 color images and only one salient object in each image. The second is a two-object database (SED2), which also has 100 color images with two salient objects in each image. Pixel-wise ground truth annotation for the salient objects in both SED1 and SED2 are provided. (3) SOD [42] is a collection of salient object boundaries based on the Berkeley segmentation dataset. Seven subjects are asked to choose the salient object(s) in 300 images. This dataset contains many images which contains multiple objects making it challenging. (4) iCoSeg is a publicly available co-segmentation data set [43], including 38 groups of totally 643 images. Each image is along with pixel-wise ground truth annotation, which may contain one or multiple salient objects. In this paper, we use it to evaluate the performance of salient object detection.

We exploit the measures used in [30], i.e., the PR (precision-recall) curve, to evaluate the performance of our proposed algorithm and other state-of-the-art methods. Precision is the fraction of detected salient pixels belonging to the salient object in the ground truth, and recall corresponds to the percentage of salient pixels correctly assigned.

$$precision(T) = \frac{|S(T) \cap G|}{|S(T)|} \tag{12}$$

$$recall(T) = \frac{|S(T) \cap G|}{|G|} \tag{13}$$

where G is the ground truth map, $|\cdot|$ denotes the sum area of masks. $S(T)$ is the binary mask obtained by directly thresholding a saliency map using threshold T. The PR curve is created by varying the saliency threshold T from 0 to 255 that determines whether a pixel is on the salient object.

To obtain F-Measure, we follow [30] to segment a saliency map by the threshold τ defined as follows:

$$\tau = \frac{2}{H * W} \sum_{x=1}^{H} \sum_{y=1}^{W} S(x, y) \tag{14}$$

where W and H are the width and height of the saliency map in pixels, respectively, and $S(x, y)$ is the saliency value of the pixel at position (x, y). If the saliency value of a superpixel is larger than threshold, it is considered as the part of salient object. In many applications, high precision and high recall are both required. We thus estimate the F-Measure [30] as:

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

where β is set to 0.3 as suggested in [12] to emphasize the precision. We also exploit mean absolute error (MAE) [13] to evaluate all algorithms aforementioned, i.e.,

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \left\| S(x,y) - G(x,y) \right\| \tag{16}$$

where W and H are the width and height of the saliency map in pixels, $S(x,y)$ and $G(x,y)$ are the saliency value and ground-truth of the pixel at position (x,y) respectively. The reason of using MAE as a compensation criterion is that precision-recall curves are insensitive to the uniformness of a saliency map. For example, by pixel-wisely multiplying a ground truth map with a 2D Gaussian centered inside the mask with arbitrary variance, one can still obtain a good precision-recall curve with such heterogeneous map. On the other hand, MAE can be affected by small error accumulation since it sums all pixel-wise errors. With these characteristics, we use it as the measurement of the saliency map in our experiment.

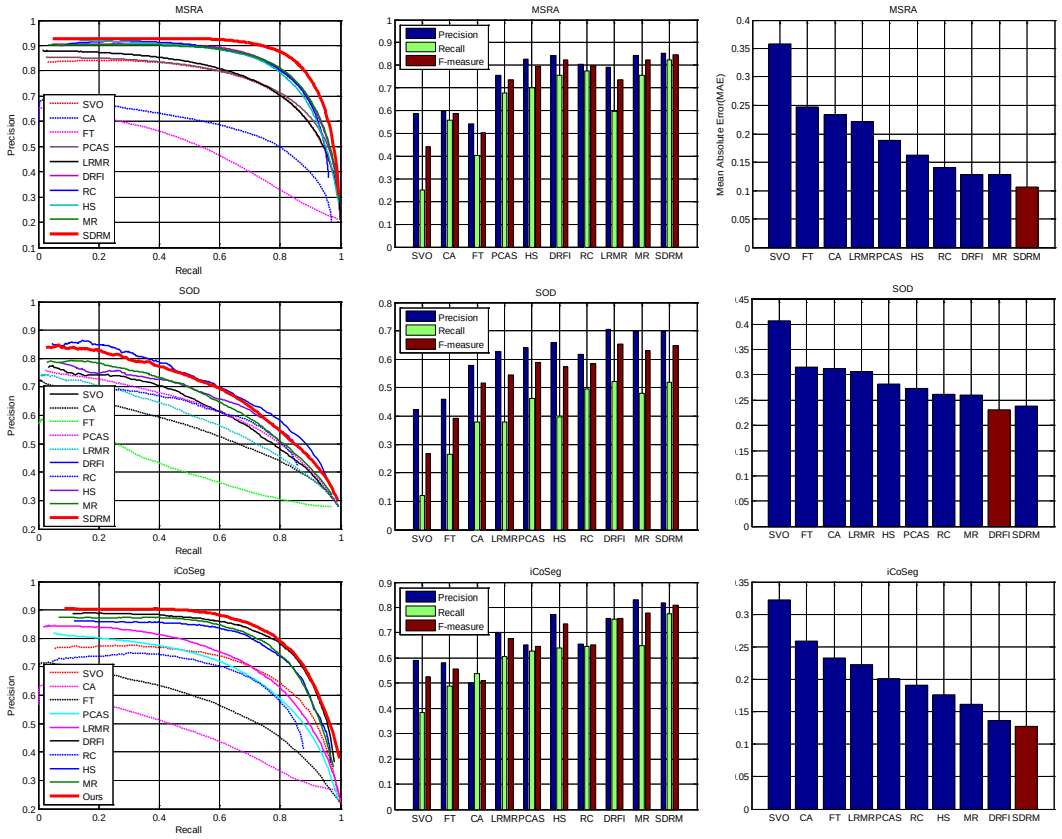


Fig. 4. Quantitative comparisons of saliency maps produced by different approaches on MSRA-B, SOD and iCoSeg dataset

4.2 Quantitative comparison

The quantitative comparison is shown in **Fig. 4** - **Fig. 5**. Generally speaking, the precision indicates the performance of the saliency detection algorithms compared with saliency map of ground-truth. To compare the proposed model with others, we always see the precision value for different algorithms, for the precision value is the ratio of the correctly detected region over the whole detected region. The performance of our method on precision-recall curves is

comparable to the most recent techniques including HS, DRFI and MR. Our method significantly outperforms DRFI on MSRA-B and iCoSeg. Marginal improvement is observed on SED1, SED2 and SOD. Besides, observing PR curves, our method is comparable to HS [10] and MR [15] on all the four datasets. Intuitively, our approach has limited ability when discovering the right boundary of salient objects in the image that with complex background (higher recall). It can be seen from SOD and SED1 dataset that the recall is not so well when it compare with DRFI. The reason might be that clutter background affects the region merging process (detailed in section 4.4), which leads to false border of the object. However, the improvements over state-of-the-arts are slightly better when considering their performance and especially the adaptability of our model to different datasets.

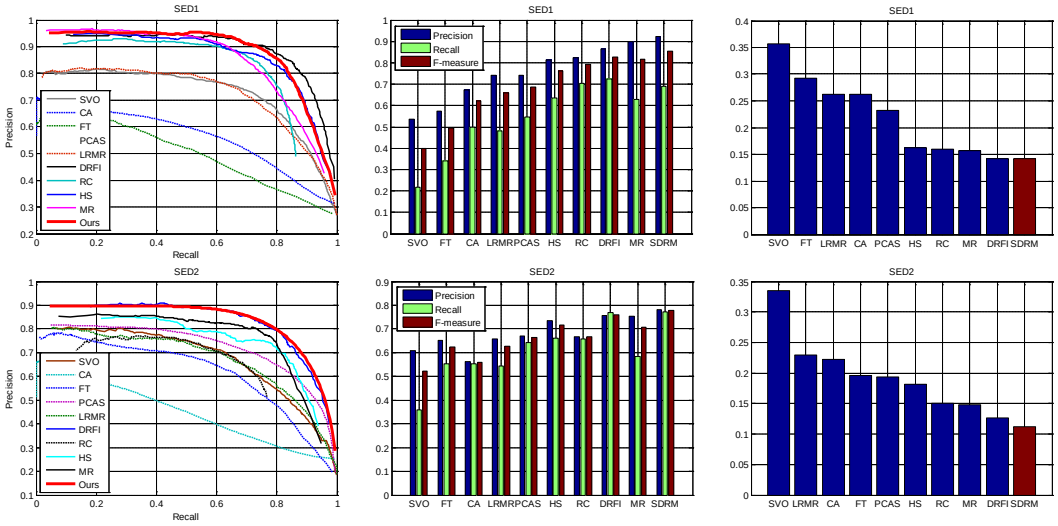


Fig. 5. Quantitative comparisons of saliency maps produced by different approaches on SED dataset

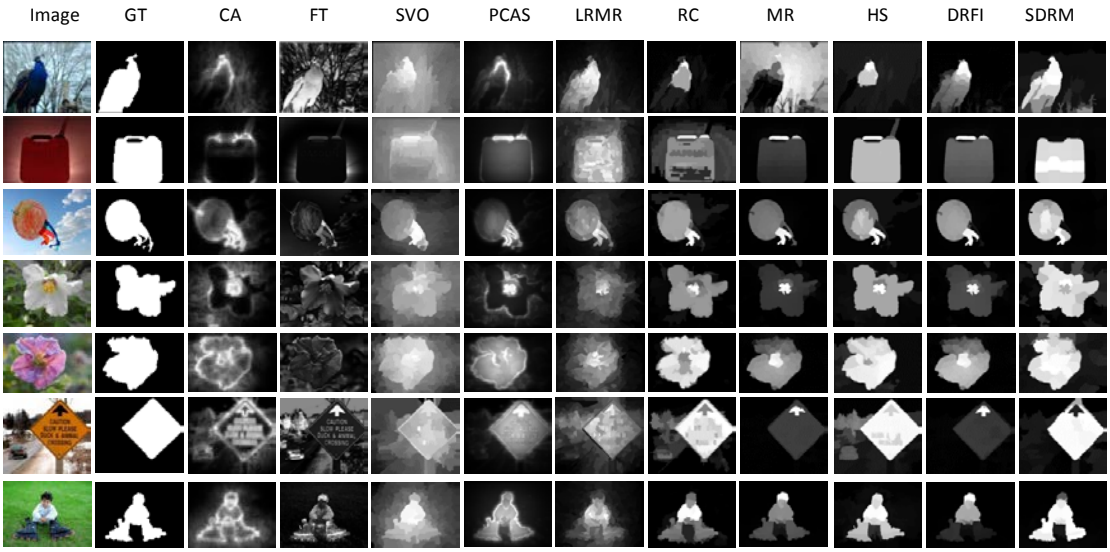


Fig. 6. Visual comparisons of saliency maps on MSRA-B dataset

Adaptive threshold experiments were carried out, where the adaptive threshold is defined as two times the mean value of a saliency map [30]. Results are shown in the middle column in

Fig. 4 and **Fig. 5**. Our method achieves both the highest precision and F-measure on MSRA-B, SED1 and SED2 datasets, providing further support to the effectiveness of the proposed method. Second best precision and F-measure for our method are observed on SOD and iCoSeg. For SED1 whose images contain single objects in more complex scenarios, our method performs close to DRFI [11]. An observation on SED2 is since this dataset has many labeled objects which violate the boundary prior, MR performs less well than other datasets.

To further evaluate the methods, we compute the MAE values [13]. As shown in the right column in **Fig. 4** and **Fig. 5**, our method produces consistently the lowest error on MSRA-B, SED and iCoSeg datasets, indicating more robustness against different datasets. Despite good performance in precision-recall curves and F-measure, LRMR [33], CA [32], FT [30] and SVO [39] have the higher MAE due to the weak background suppression.

4.3 Visual comparison

We also provide the visual comparison of different methods in **Fig. 6 - Fig. 9**. As can be seen, our method effectively suppresses background clutter and uniformly emphasizes the foreground objects. In most visual comparisons, much clearer object boundaries are obtained compared to other methods, e.g. last row in **Fig. 6**, 1st, 2nd, 4th rows in **Fig. 8**, and 1st row in **Fig. 9**. In addition, the proposed method is able to deal with images containing clutter background (e.g. 1st row in **Fig. 6**, 6th and 7th rows in **Fig. 9**). Our region merging scheme effectively combines them into background, preserving perceptual homogeneity.

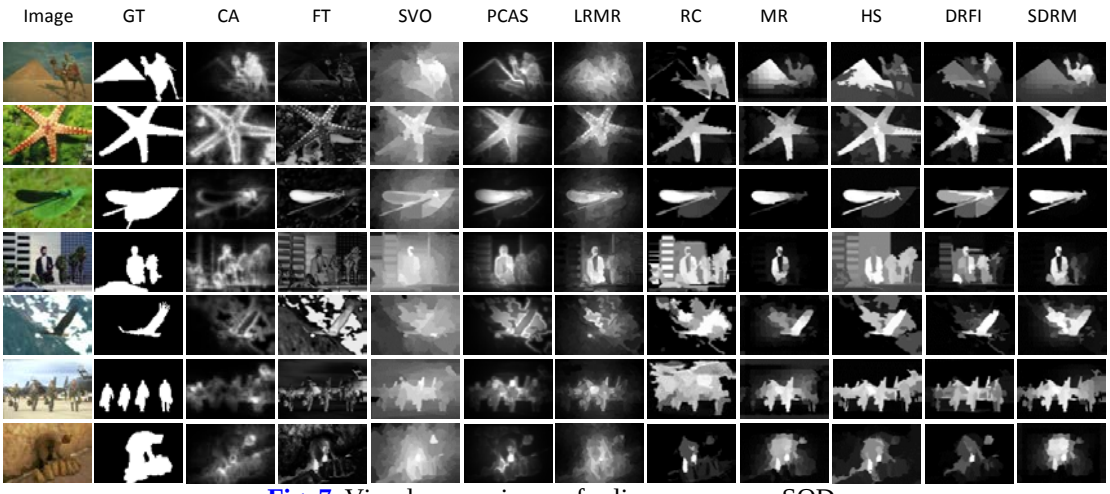


Fig. 7. Visual comparisons of saliency maps on SOD

It is also worth pointing out that our approach performs well when the object touches the image border, e.g. the first two rows in MSRA-B dataset in **Fig. 6**, the second row in SOD in **Fig. 7** and first row in iCoSeg in **Fig. 9**, even though it violates the pseudo-background assumption used in MR [15]. MR, in which the first stage is based on the pseudo-background assumption, can not label the saliency seeds correctly when the object slightly touches the image border (e.g. 1st row in **Fig. 6**). Another side-effect of this operation is the risk of missing useful object parts. This is consistently observed on SED dataset. As two objects in one image may be of different saliency levels, one of the two objects in an image can be “lost” after thresholding, leading to a performance drop (e.g. 6th row in **Fig. 8**). In contrast, such risk is avoided in our method as no threshold is used to binarize the saliency map for performance boosting.

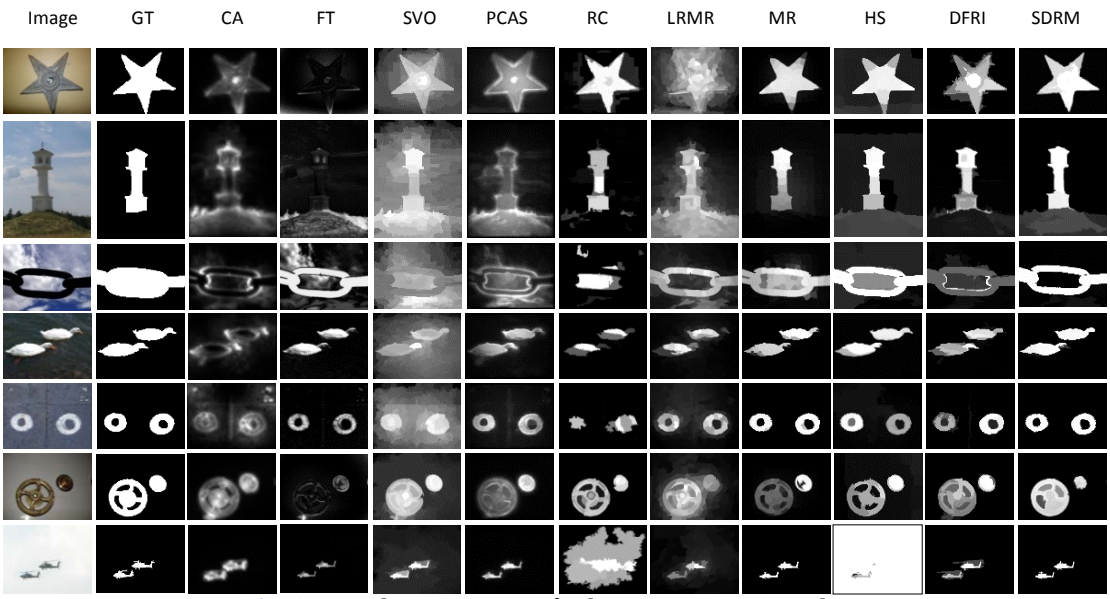


Fig. 8. Visual comparisons of saliency maps on SED dataset

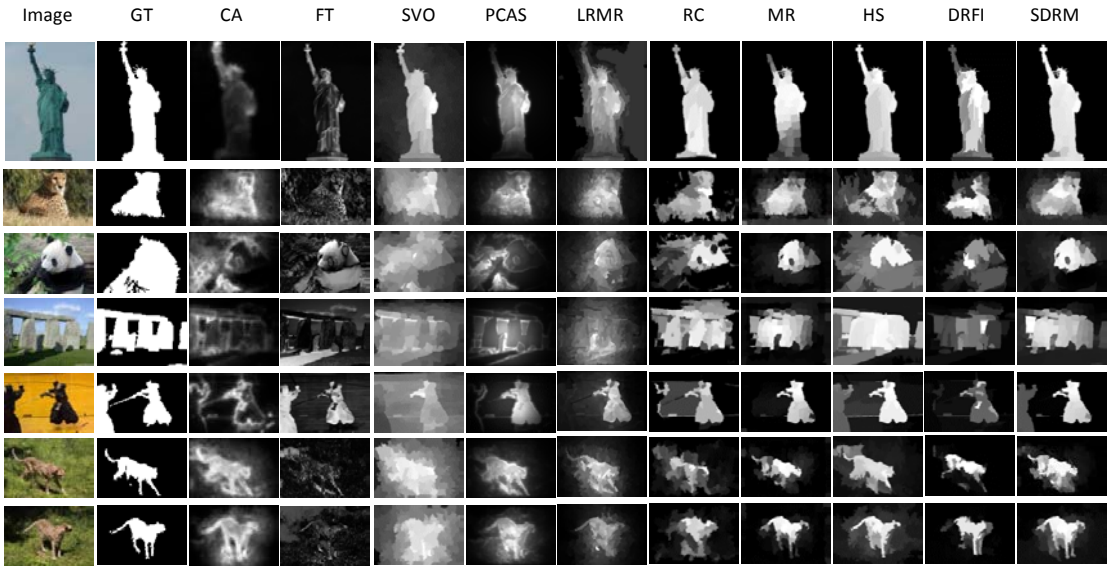


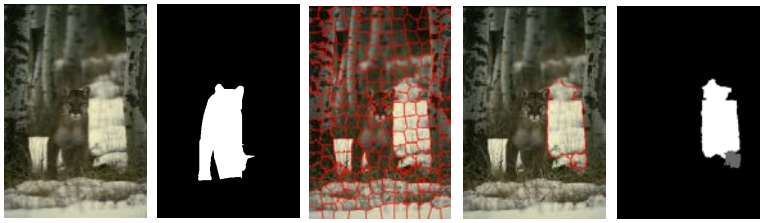
Fig. 9. Visual comparisons of saliency maps on iCoSeg dataset

For other state-of-the-art approaches, it can be seen that while SVO [39] detects the salient regions, parts of the background are erroneously detected as salient. FT [30], which exploited the central-surround principle by defining the saliency of a pixel as its distance to the average image, generally detect the foreground from input images. However, it is easy to influence by the background that the salient area contains not only salient object but also clutter background. As relies mostly on patterns, CA [32] detects the outlines of the saliency objects, while missing their interior. By relying solely on color, RC [12] can mistakenly focus on distinct background colors, e.g., the sky is captured instead of the iron chain in the third row on SOD database. PCAS [40] relies mostly on patterns; hence, it detects the outlines of the saliency objects, while missing their interior. LRMR [33], which integrates the high-level priors, focus

on the center and the warm color of image. It is worth mentioning that the salient objects with warm colors such as red and yellow are more pronounced. HS [10] considers the situation that locally smooth regions could be inside a salient object and globally salient color, contrarily, could be from the background. However, it is easy to lose the small target, such as the last row in SED in **Fig. 9**. DRFI [11], which is based on multi-level image segmentation, uses the supervised learning approach to map the regional feature vector to a saliency score, and finally fuses the saliency scores across multiple levels, yielding the saliency map. When most of the images contain only one object in training set, it has limited ability to discover all the salient objects within one image.

4.4 Sensitivity to region merging

As stated previously, the proposed framework detects salient regions depending on the process of adaptive region merging. The results of salient object detection are affected by that of the region merging more or less. Our method obtains accurate saliency maps mainly based on the aspects that the framework can segment the object effectively in region merging process. However, this framework may fail for those images in which salient object(s) is hided in more complex background. In such case, adaptive region merging algorithm merges the object into background and the noise in background is considered as the target falsely, which makes the detection results of subsequent salient object detection inaccuracy.



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

Fig. 10. Failure cases of the proposed algorithm, (a) Input images, (b) ground truth, (c) SLIC segmentation, (d) region merging result, and (e) the saliency maps of the proposed algorithm

Fig. 10 shows an example. The lion in **Fig. 10** is not “pop-out” as a whole due to the small color dissimilarity between the lion and the trees in the background, but it is still a salient object due to the object recognition capability of the trained human brain. In adaptive region merging framework, the superpixels that located at left, top and right boundary of the image are regarded as the seeds to start the merging process. As the color of lion is similar to the seeds, it is merged into background in the first stage. Since the snow in the background is dissimilar to the trees and the lion, it is segmented as the object in the following process. In the process of region saliency measurement, it is detected as the salient object, which was only to be expected. Our future work will focus on high-level knowledge, which could be beneficial to handle more challenging cases and other kinds of saliency cues or priors to be embedded into our framework.

5. Conclusion

We propose a bottom-up method to detect salient regions in images based on adaptive region merging. Different from most existing salient object detection algorithms that commonly employed segmentation techniques to eliminate background noise and reduce computation by

treating each segment as a processing unit, the proposed framework use adaptive region merging to combine the similar regions. Based on the merging results, we use three region saliency measurements to generate the saliency maps which have capability on modeling global perceptual phenomena. We evaluate the proposed algorithm on large datasets and demonstrate promising or comparable results with comparisons to state-of-the-art methods. For handle more challenging cases, our future work will focus on high-level knowledge and other kinds of saliency cues or priors to be embedded into our framework.

References

[1] A. Toet, "Computational versus Psychophysical Bottom-Up Image Saliency: A Comparative Evaluation Study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2131-2146, 2011. Article (CrossRef Link)

[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185-207, 2013. Article (CrossRef Link)

[3] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1-8, June, 2007. Article (CrossRef Link)

[4] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998. Article (CrossRef Link)

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, pp. 545-552, 2006.

[6] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 470-477, June 16-21, 2012. Article (CrossRef Link)

[7] J. Zhou, S. Gao, Y. Yan, et al., "Saliency detection framework via linear neighborhood propagation," *IET Image Processing*, vol. 8, no. 12, pp. 804-814, 2014. Article (CrossRef Link)

[8] X. Shi, N. Bruce, J. Tsotsos, "Fast, recurrent, attentional modulation improves saliency representation and scene recognition," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1-8, 2011. Article (CrossRef Link)

[9] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1007-1013, 2009. Article (CrossRef Link)

[10] Q. Yan, L. Xu, J. Shi, J. Jia, "Hierarchical Saliency Detection," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1155-1162, June 23-28, 2013. Article (CrossRef Link)

[11] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2083-2090, June 23-28, 2013. Article (CrossRef Link)

[12] M. Cheng, G. Zhang, N. Mitra, X. Huang, S.Hu, "Global contrast based salient region detection," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 409-416, June 16-21, 2011. Article (CrossRef Link)

[13] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 733-740, June, 2012. Article (CrossRef Link)

[14] Zhou J, Ren Y, Yan Y, et al., "Salient object detection: manifold-based similarity adaptation approach," *Journal of Electronic Imaging*, vol. 23, no. 6, pp. 063004-063004, 2014. Article (CrossRef Link)

[15] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, "Saliency Detection via Graph-Based Manifold Ranking," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3166-3173, June 23-28, 2013. Article (CrossRef Link)

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274-2282, Nov., 2012. Article (CrossRef Link)
[17] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, May, 2002. Article (CrossRef Link)
[18] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167-181, Sep., 2004. Article (CrossRef Link)
[19] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898-916, May, 2010. Article (CrossRef Link)
[20] C. Li, Y. Yuan, W. Cai, et al., "Robust saliency detection via regularized random walks ranking," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2710-2717, June 7-12, 2015. Article (CrossRef Link)
[21] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in Neural Information Processing Systems*, pp. 155-162, 2006.
[22] A.Oliva, A. Torralba,M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *Proc. of International Conference on Image Processing*, pp. 253-256, 2003. Article (CrossRef Link)
[23] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, 8(7):32, pp. 1-20, 2008. Article (CrossRef Link)
[24] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, 8(7):13, pp. 1-18, 2008. Article (CrossRef Link)
[25] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1-8, 2008. Article (CrossRef Link)
[26] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194-201, 2012. Article (CrossRef Link)
[27] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. of 6th Int. Conf. Computer Vision Systems*, pp. 66-75, 2008. Article (CrossRef Link)
[28] T. Liu, Z. Yuan, J. Sun, et al, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no.2, pp. 353-367, 2011. Article (CrossRef Link)
[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. of ICCV*, pp. 2106-2113, 2009. Article (CrossRef Link)
[30] R. Achanta, S. Hemami, F. Estrada, & S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1597-1604, June, 2009. Article (CrossRef Link)
[31] Y. Wei, F. Wen, W. Zhu, & J. Sun, "Geodesic saliency using background priors," in *Proc. of ECCV*, pp. 29-42, 2012. Article (CrossRef Link)
[32] S. Goferman, L. Zelnik-Manor, & A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no.10, pp.1915-1926, 2012. Article (CrossRef Link)
[33] X. Shen, Y. Wu, "A Unified Approach to Salient Object Detection via Low Rank Matrix Recovery," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 853-860, 2012. Article (CrossRef Link)
[34] Z. Jiang, L. Davis, "Submodular Salient Region Detection," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2043 - 2050, 2014. Article (CrossRef Link)
[35] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1131-1138, 2013. Article (CrossRef Link)
[36] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Proc. of Int. Conf. Pattern Recognit. (ICPR)*, pp. 1-4, Dec., 2008. Article (CrossRef Link)

[37] M. Sonka, V. Hlavac, R. Boyle, *Image Processing, Analysis and Computer Vision*, Thomson, 2007.
[38] S. E. Palmer, V*ision Science: Photons to Phenomenology*. Cambridge, MA, USA: MIT Press, 1999.
[39] K. Chang, T. Liu, H. Chen, and S. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. of 13th IEEE International Conference on Computer Vision (ICCV)*, pp. 914-921, 2011. Article (CrossRef Link)
[40] R. Margolin, A. Tal and L. Zelnik-Manor, "What Makes a Patch Distinct?" in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1139-1146, 2013. Article (CrossRef Link)
[41] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 73-80, 2010. Article (CrossRef Link)
[42] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1-8, 2007. Article (CrossRef Link)
[43] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3169-3176, 2010. Article (CrossRef Link)

Jingbo Zhou received his PhD degree in control science and engineering from Nanjing University of Science and Technology (NUST) in 2013. His research interests include pattern recognition, image processing, data clustering, etc.

Jiyou Zhai is a doctoral student majoring in College of Computer and Information at Hohai University from September 2012 to present. He is a lecturer at Nanjing Institute of Technology. His research interests include pattern recognition, image processing, etc.

Yongfeng Ren received his PhD degree in computer science and engineering from College of Computer and Information at Hohai University in 2016. His research interests include pattern recognition, image processing, etc.