Neurocomputing 000 (2017) 1–17

Contents lists available at ScienceDirect

Neurocomputing

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

Spectral salient object detection

Keren Fu[a,*], Irene Yu-Hua Gu[b], Jie Yang[c]

[a] College of Computer Science, Sichuan University, Sichuan 610065, China
[b] Signal Processing Group, Department of Signals and Systems, Chalmers University of Technology, Gothenburg SE-41296, Sweden
[c] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 14 May 2017
Revised 29 August 2017
Accepted 8 September 2017
Available online xxx

Communicated by Prof. Junwei Han

Keywords:
Salient object detection
Object holism
Normalized cut
Graph partition
Regional saliency

ABSTRACT

Many salient object detection methods first apply pre-segmentation on image to obtain over-segmented regions to facilitate subsequent saliency computation. However, these pre-segmentation methods often ignore the holistic issue of objects and could degrade object detection performance. This paper proposes a novel method, *spectral salient object detection*, that aims at maintaining objects holistically during pre-segmentation in order to provide more reliable feature extraction from a complete object region and to facilitate object-level saliency estimation. In the proposed method, a hierarchical spectral partition method based on the normalized graph cut (Ncut) is proposed for image segmentation phase in saliency detection, where a superpixel graph that captures the intrinsic color and edge information of an image is constructed and then hierarchically partitioned. In each hierarchy level, a region constituted by superpixels is evaluated by criteria based on figure-ground principles and statistical prior to obtain a regional saliency score. The coarse salient region is obtained by integrating multiple saliency maps from successive hierarchies. The final saliency map is derived by minimizing the graph-based semi-supervised learning energy function on the synthetic coarse saliency map. Despite the simple intuition of maintaining object holism, experimental results on 5 benchmark datasets including ASD, ECSSD, MSRA, PASCAL-S, DUT-OMRON demonstrate encouraging performance of the proposed method, along with the comparisons to 13 state-of-the-art methods. The proposed method is shown to be effective on emphasizing large/medium-sized salient objects uniformly due to the employment of Ncut. Besides, we conduct thorough analysis and evaluation on parameters and individual modules.

1. Introduction

Studies from neurobiology and cognitive psychology indicate that human brains are capable of selecting a certain visual contents for further processing [1,2]. Modeling human bottom-up visual attention on images, referred to as *bottom-up saliency detection*, is aimed at detecting salient image parts that can easily attract human attention. Bottom-up saliency detection has gained increasing research interest recently. Under this theme there are two sub-types [3,4], namely *eye fixation modeling* [4–7] and *salient object/region detection* [3,8,9]. In this paper, we address the second type. The recent advance in salient object detection is driven by high-level applications such as automatic object segmentation [10,11], content-aware image editing [12–15] and retrieval [16,17].

Many existing methods for salient object detection in still images, e.g., [9,18–21] essentially employ certain pre-segmentation techniques. Resultant regions/superpixels are then considered as basic processing units and fed into saliency computation. This mean not only facilitates computation but also avoid pixel-level noise. Typical techniques include clustering-based segmentation (e.g., Meanshift [22], SLIC superpixels [23]), or merging-based segmentation (e.g., graph-based [24]). Unfortunately, since these methods are based on local image properties, they could result in highly over-segmented regions, where an object breaks up into small regions that ignore the holistic object. Such regions not only easily introduce noise, but also keep one from assessing the object as a whole entity (Fig. 1). As a salient object detection method is aimed at emphasizing the entire object uniformly in the resultant saliency map [3] (the ground truth in Fig. 1), a pre-segmentation that is consistent to the human visual perception and retains holistic object, intuitively, can contribute to more accurate saliency estimation. Retaining holistic object allows more reliable feature extraction and analysis such as colors, shapes and texture from a complete object region.

Segmenting complete regions of arbitrary objects in a pre-segmentation stage is a very challenging task. This is because the task of saliency detection aims at detecting generic arbitrary

* Corresponding author.
E-mail address: fkrsuper@scu.edu.cn (K. Fu).

http://dx.doi.org/10.1016/j.neucom.2017.09.028
0925-2312/

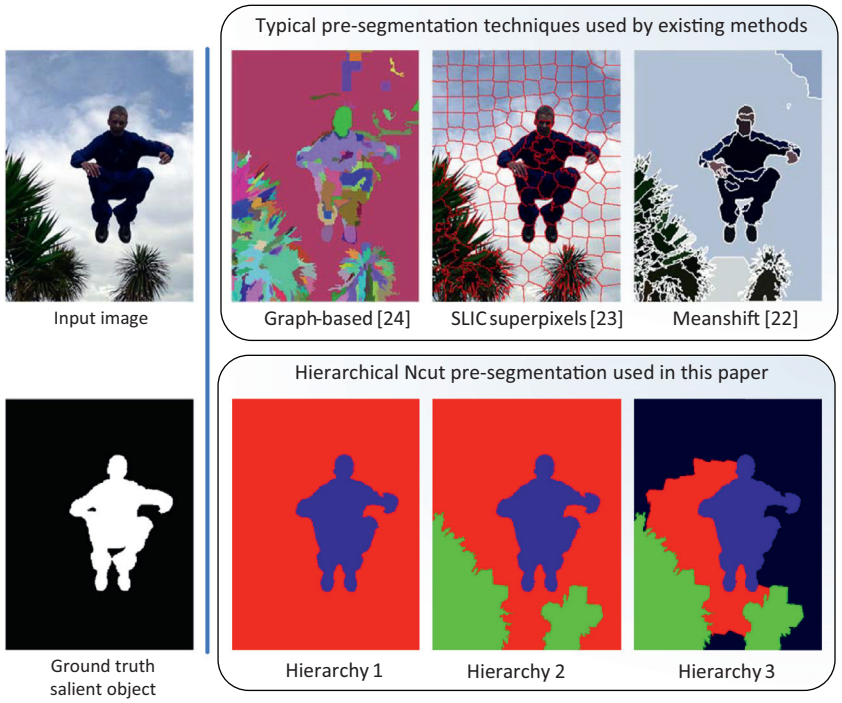

Fig. 1. Comparison between several commonly used segmentation methods and the proposed Ncut-based pre-segmentation. Row-1 (columns 2–4): graph-based segmentation [24], SLIC superpixels [23], and Meanshift segmentation [22]); Row-2 (columns 2–4): results from three hierarchies from the proposed method, where different segments are assigned with different colors. The parameters of the graph-based segmentation [24] were chosen in a similar way as in [9]. About 200 SLIC superpixels [23] were generated similar to [20,21]. The parameter setting of Meanshift [22] is similar to [8]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

objects which human eyes attend. Therefore, specifying category-dependent prior to assist segmentation is not feasible. To remedy this, we propose to utilize a spectral partition technique—the normalized graph cut (Ncut) [25] for salient object detection, because Ncut is unsupervised and shows more agreement to we human perception. As written by Shi and Malik [25], "Rather than focusing on local features and their consistencies in the image data, Ncut aims at extracting the global impression of an image". In this paper, we propose a hierarchical spectral partition method for the segmentation phase of saliency detection. The proposed method uses a superpixel graph to capture the intrinsic image color and edge information, and is based on the following observations: (a) Ncut has a strong discriminative power to separate image contents in object-level because it is a global criterion. It has hence the potential to maintain object holism and return complete boundary of an object (Fig. 1). (b) A salient object often has some unique appearance in terms of color or texture as comparing to its surroundings, implying some visual dissimilarity [5,9]. Therefore, a complete boundary of object is preferred by Ncut. A very low Ncut cost is often achieved if an entire object is separated from the remaining image.

Although Ncut has been widely used for image segmentation [25–27], applying it for inducing saliency computation and hence saliency maps is not well-studied. Furthermore, the aim of image segmentation is different from that of salient object detection, and hence, using Ncut solely is not adequate for rendering a saliency map. To further address this issue, we incorporate Ncut with several regional saliency metrics. Some other works [9,10,20,28,29] append graph cut to saliency maps as a second stage to achieve figure-ground segmentation. Our method differs from theirs since we firstly use Ncut to retain object holism prior to saliency computation, whereas [9,10,20,28,29] do not preserve object holism during saliency detection and might achieve less satisfactory saliency maps. In this paper, we believe conducting saliency detection on holism-retained segmentation could lead to better detection accuracy. Besides, the proposed method differs obviously from the above methods on technical aspects as well as implementations. The main novelties of this paper are summarized below:

1. For maintaining the holism of objects, Ncut is employed for separating salient image contents, where a novel hierarchical spectral partition method is introduced for pre-segmentation. It partitions a superpixel graph that captures the intrinsic color and edge information in an image. A binary segmentation tree is later generated, where an entire object is likely to be retained in emerging hierarchies.
2. For modeling image saliency in different hierarchies, Ncut is incorporated with regional saliency metrics. Three regional saliency metrics are introduced based on figure-ground principles and statistical prior. Salient objects are enhanced by integrating intermediate saliency maps from successive hierarchies.
3. Despite the simple intuition of maintaining object holism, we show that the proposed method achieves state-of-the-art performance on 5 benchmark datasets. Parameters of the proposed method are evaluated both quantitatively and comprehensively.

Although part of our work is published in the conference paper [30], this paper has significantly extended and improved our previous work, where we incorporate additional edge term in graph affinity computation, employ constrained Ncut for the first hierarchy for better cut initialization, and also conduct thorough evaluation on parameters and modules. In addition, more technical details and further extensive test results on objects in complex background are included.

The reminder of the paper is organized as follows. Section 2 describes the related work on salient object detection. Section 3 briefly reviews the fundamental of normalized graph cut, upon which our proposed method is built. Section 4 describes the proposed method in details. Experimental results,

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

performance evaluation and comparisons are then included in Section 5. Finally, conclusions are drawn in Section 6.

2. Related work

Early methods on bottom-up salient object detection are based on the assumption that salient objects are unique in color and present high color contrast to the rest parts of an image. Cheng et al. [9] propose a regional saliency measure as the color histogram contrast to other regions. Perazzi et al. [20] formulates complete contrast and saliency estimation using high dimensional Gaussian filters. Fu et al. [18] detect salient objects by superpixel-based color contrast and color distribution, which are non-linearly integrated to achieve complementary performance. Shen and Wu [19] solve the saliency detection issue as a low rank matrix recovery problem, where salient objects are represented by a noisy sparse matrix while the background is indicated by a low rank matrix. Wang and coworkers [31] compute the pixel-wise image saliency by aggregating complementary appearance contrast measures with spatial priors. Margolin et al. [32] define the distinctness of patches as L1-norm in PCA coordinates and combine it with color distinctness. Li et al. [33] model patch saliency by dense and sparse reconstruction errors, where the dictionaries for reconstruction are obtained from image boundary.

Graph representation is recently used for rendering image saliency. Based on the graph, characteristics of salient objects, such like high color contrast to surroundings, compact color distribution, connectivity to image borders, are modeled. Gopalakrishnan et al. [34] perform random walks on graphs to find salient objects. Wei et al. [35] propose to treat boundary parts of an image as the background. The patch saliency is defined as the shortest geodesic distance on a graph to image boundary. Yang et al. [21] use four image borders as seeds and propagate saliency by graph-based manifold ranking. Fu et al. [36] propagate saliency energy from a coarse energy map based on geodesic distance. Zhu et al. [37] propose a saliency detection method based on robust background estimation. Gong et al. [38] propose a new saliency propagation algorithm employing the teaching-to-learn and learning-to-teach strategies to explicitly improve the propagation quality. More recently, Zhang et al. [39] perform saliency detection based on minimum barrier distance and show its robustness over the extensively used geodesic distance. Fu et al. [40] propose a manifold-preserving graph-based diffusion technique and apply it to saliency diffusion. Li et al. [41] propose to locate coarse saliency regions by fixation prediction and later refine the results by a multi-layer graph-based algorithm.

While the above works are unsupervised saliency computational models, other methods estimate object saliency by supervised learning resorting to human annotations. Liu et al. [42] segment salient objects by learning and inferencing a conditional random field (CRF). Jiang et al. [43] use a random forest regressor to map multiple features to a regional saliency score. Mai et al. [44] propose a data-driven approach for aggregating pixel-level saliency maps. Tong et al. [45] use bootstrap learning to enhance the detection performance. [46] proposes a salient object detection system via proposal subset optimization. The optimization framework is based on maximum a posteriori principle and outputs a compact set of proposal windows containing salient objects. Wang et al. [47] propose correspondence saliency transfer, where initial saliency is obtained by transferring salient region masks of support images from a large reference dataset and later refined via random-walk-with-restart algorithm. Qi et al. [48] incorporate the global restricted Boltzmann machine (RBM) with local conditional random field (CRF) into a unified framework and learn to infer image saliency cues. Recently, deep learning and convolutional neural networks (CNN) are employed [49–51]. Their strong power over traditional supervised/unsupervised saliency models is witnessed. Note that there also exist other saliency works which focus on different fields, such as co-salient objects detection [52–54], video and event saliency discovery [55–57], but these are beyond the scope of this paper.

Most of the above methods focus on developing novel saliency computation methods either on pixel level, superpixel level, or region level, but does not explicitly consider whether these computational units could reflect the global meanings and object holism. Several more closely related studies to our work are [58–60]. The work of Yan's [58] and Cheng's [59] also focus on generating good segmentation/abstraction in multi-scales. Yan et al. [58] propose a hierarchical saliency detection method that merges regions according to user-defined scales to eliminate small-sized distracters. Cheng et al. [59] measure the saliency based on a hierarchical soft abstraction, that includes a pixel layer, a histogram layer, a Gaussian Mixture Model layer and a clustering layer. Liu et al. [60] propose saliency tree as an enhancement on an initial saliency map computed on primitive regions. As comparison to their methods, our method is different from [58–60] on motivations. We use the normalized graph cut to retain objects holistically for enhancing the saliency detection, whereas works in [58–60] are based on the local merging or clustering. Besides, a graph is even not explicitly used in [58–60]. As reported in Section 5.3, we achieve better accuracy for salient object detection than [58–60].

As aforementioned in Section 1, there are some previous literatures [9,10,20,28,29] involving both the graph cut and saliency detection, where the graph cut usually serves as a post-processing (i.e., applied to the outcome of saliency detection to achieve a binary segmentation). The main difference in the proposed method is that we utilize Ncut to enhance saliency detection. In the proposed method, the Ncut is applied prior to the saliency detection, and hence can augment the resulting saliency maps.

3. Reviews of normalized graph cut

This section briefly reviews the theory of normalized graph cut (Ncut). Ncut proposed by Shi and Malik [25] normalizes the cost of graph cut by using the total edge connection towards all nodes in a graph. Let a similarity graph (whose edge weights measure the similarity between vertices) be defined as $G = (V, E)$, $\mathbf{W}$ be the affinity matrix of G (i.e., similarity matrix), $\mathbf{D}$ be the degree matrix of G (a diagonal matrix with diagonal entry $d_i = \sum_j w_{ij}$, where w_{ij} is entry of $\mathbf{W}$). The purpose of Ncut is to find a cut that partitions V into two vertex sets A and B (s.t. $A \bigcup B = V$, $A \bigcap B = \emptyset$) such that the following cost is minimized:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{1}$$

where $cut(A, B) = \sum_{u \in A,\ v \in B} e(u, v)$ is the total weight of edges connecting A and B, the association $assoc(A, V) = \sum_{u \in A,\ t \in V} e(u, t)$ is the total weight of edges between A and all nodes (i.e., V) in the graph, $assoc(B, V)$ is similarly defined. The rationale behind using Ncut is to minimize the similarity between A and B. Shi and Malik [25] show that minimizing $Ncut(A, B)$ is equivalent to minimizing the following energy function:

$$E(\mathbf{y}) = \frac{\mathbf{y}^T(\mathbf{D} - \mathbf{W})\mathbf{y}}{\mathbf{y}^T\mathbf{D}\mathbf{y}} \tag{2}$$

where $\mathbf{y}$ is a discrete indication vector under the constraint $\mathbf{y}^T\mathbf{D}\mathbf{1} = 0$ (where $\mathbf{1}$ is the vector of ones). The exact solution to (2) is NP hard [25]. However a continuous approximation of $\mathbf{y}$ can be obtained as the eigenvector associated with the second smallest eigenvalue of the following generalized eigen-system:

$$(\mathbf{D} - \mathbf{W})\mathbf{y} = \lambda\mathbf{D}\mathbf{y} \tag{3}$$

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

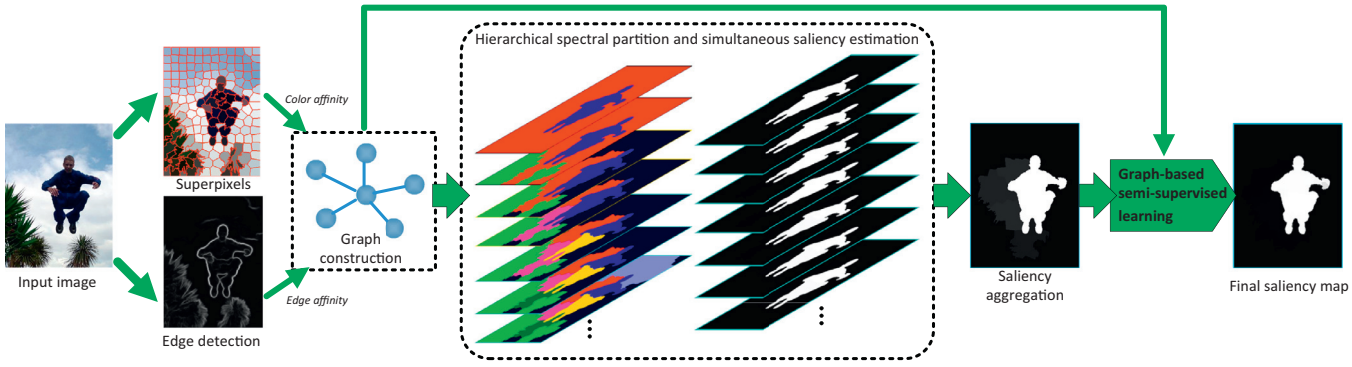

Fig. 2. The block diagram of the proposed method.

where $\mathbf{y}$ and λ are the eigenvector and eigenvalue, respectively. It is worthy noting that since Ncut [25] deploys the smallest eigenvectors (i.e., eigenvectors corresponding to smallest eigenvalues) to perform graph partition, it is known as a *spectral partition technique* and is quite popular (has over 10K Google citations).

As described previously, the Ncut partitions a graph into two parts in a normalized discriminative fashion. We adopt Ncut for saliency detection as it is expected to generate better segmentation hypotheses for salient object detection as comparing with the clustering-based or merging-based segmentation.

4. The proposed method

In this section, we describe the proposed method for salient object detection that integrates hierarchical spectral partition and saliency estimation. The block diagram of the proposed method is shown in Fig. 2. For each input image, we first decompose it into superpixels and simultaneously perform edge detection. We then use superpixels as vertices to construct a graph that captures intrinsic colors and edge information of an image (Section 4.1). Hierarchical spectral partition that is aimed at retaining the entire object is applied to the graph (Section 4.2), meanwhile saliency estimation is carried out on successive hierarchies (Section 4.3). The resultant intermediate saliency maps are integrated/aggragated and then refined by graph-based semi-supervised learning (Section 4.4), yielding to a final saliency map.

4.1. Graph construction

Instead of conducting Ncut on a pixel graph [25], we choose to construct a superpixel graph that increases the efficiency and facilitates the subsequent saliency estimation. An image is first segmented into N SLIC superpixels [23], since SLIC method produces a more regular lattice and therefore is suitable to build a graph. Note that more recent superpixel generation algorithms such as in [61] may also be employed. Next, we construct an initial graph $G = (V, E)$ whose vertices V constitute of all the superpixels, and E is a set of graph edges. Hereafter, notation v_i refers to the ith "superpixel", or "graph vertex", where "node" and "vertex" are used interchangeably in the remaining text. Let $e_{ij} \in E$ denote the graph edge between v_i and v_j. The corresponding affinity weight of the edge is denoted as w_{ij}. In the proposed method, a superpixel is connected to its spatially adjacent superpixels to form graph edges. Additionally, each pair of boundary superpixels (i.e., superpixels that touch image boundary) are connected to each other, similar to the close-loop graph in [21]. Different from existing methods [21,37] that only consider colors of superpixels, the graph in the proposed method captures both color and edge information. Our affinity matrix $\mathbf{W} = [w_{ij}]_{N\times N}$ is defined as:

$$w_{ij} = \begin{cases} \sqrt{S_{ij}^{\text{color}} \cdot S_{ij}^{\text{edge}}} & \text{if } v_i, v_j \text{ are spatially adjacent} \\ S_{ij}^{\text{color}} & \text{else if } v_i, v_j \in B \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where B denotes the set of boundary superpixels. The color affinity S_{ij}^{color} and the edge affinity S_{ij}^{edge} are defined as:

$$\begin{aligned} S_{ij}^{\text{color}} &= \exp(-\lambda_c ||\mathbf{c}_i - \mathbf{c}_j||) \\ S_{ij}^{\text{edge}} &= \exp(-\lambda_e \max_{i' \in \bar{ij}} ||f_{i'}||) \end{aligned} \tag{5}$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ are averaging CIELab colors in v_i and v_j, $\bar{ij}$ is a straight line connecting centers of two superpixels on the image plane, i' is an arbitrary pixel on $\bar{ij}$, $||f_{i'}||$ refers to the edge magnitude at i' that can be derived from an edge detector [62], and λ_c and λ_e are parameters controlling the damping rate of the two terms. Using (4), large color difference and strong intervening image edges between two superpixels would lead to small graph affinity, meaning they are less likely to belong to the same region. It is worthy noting that in practice although most superpixels in an image have their centers inside due to the compactness nature of SLIC superpixels [23], there may be few superpixels whose centroids are located out of them, resulting in less accurate intervening edge cue. Hence for such a case, a pixel location is sampled randomly inside each superpixel, and is used instead of the centroid to compute the edge affinity. Fig. 3 shows an example of superpixels, edge detection and local graph connection.

4.2. Hierarchical spectral partition

In this section, we describe the proposed hierarchical partition method in detail. Once the graph affinity matrix $\mathbf{W}$ is obtained, we form the 1st hierarchy by applying the spectral partition technique, namely Ncut to G to divide the superpixels into two segments. Since the subsequent partitions will be based on the first hierarchy, here we introduce the constrained Ncut [63] which enables grouping priors, to provide us a good partition initialization. Based on the fact that image boundary nodes are likely to be background, a prior is incorporated here which intends to group boundary nodes into one segment. To be specific, the constrained Ncut in the first hierarchy is conducted by solving the below eigen-system:

$$(\mathbf{D} - \mathbf{W} + \mathbf{U}\Gamma\mathbf{U}^T)\mathbf{y} = \lambda\mathbf{D}\mathbf{y} \tag{6}$$

where $\mathbf{D}$, $\mathbf{W}$, $\mathbf{y}$, and λ are of the same meanings as in (3). Given boundary node set B, $\mathbf{U}$ is an $N \times m$ matrix where the m denotes the number of "must-link" conditions [63]. Each column of $\mathbf{U}$ corresponds to a pairwise constraint and has zero entries except for

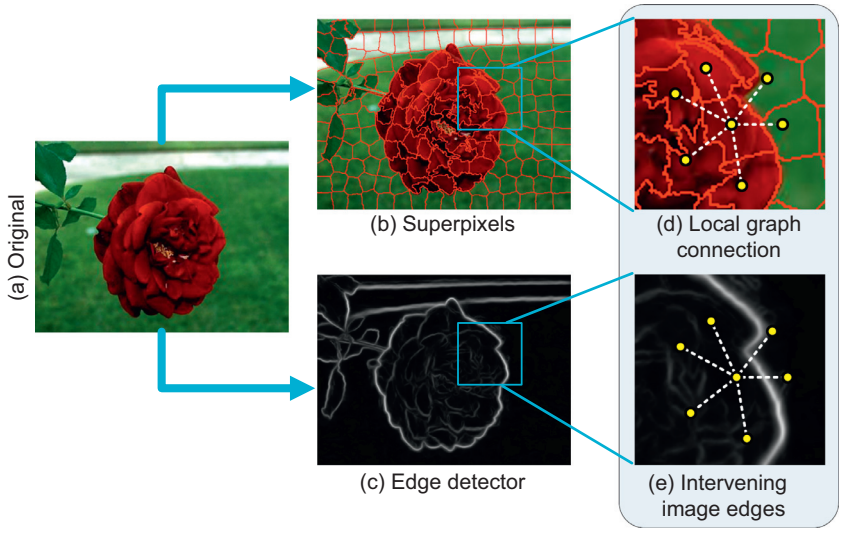

Fig. 3. Resultant superpixels, edge detection and graph connection in the proposed method. (a) The original image; (b) contours of SLIC superpixels [23] (red) marked on the original image; (c) corresponding edge detection from [62]; (d) illustration of local connections of a node, where yellow circles denote vertices and dash lines denote graph edges. A node is connected to its adjacent spatial neighbors; (e) The intervening edge magnitude between nodes revealing how they are likely to be in a same region. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

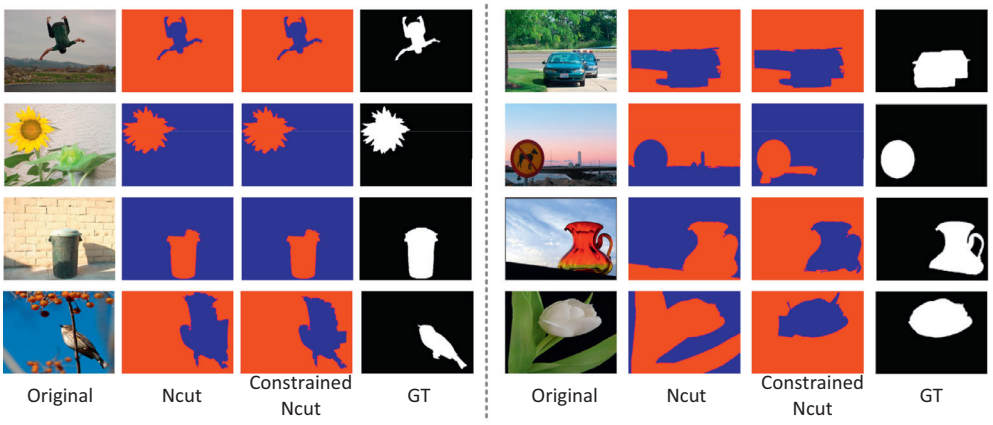

Fig. 4. Example results from ASD [8] after applying Ncut and constrained Ncut to each image. The resulting segments are labeled with either red or blue. GT denotes the ground truth masks of salient objects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

two boundary nodes that are considered in a "must-link". The entries for such two boundary nodes are respectively 1/2 and −1/2. Since we need to have at least $|B|-1$ conditions to glue up all boundary nodes, $m=|B|-1$ is set and the "must-links" among boundary nodes are designed in an arbitrary sequential permutation of boundary node indexes. Meanwhile, $\mathbf{\Gamma}$ is a diagonal matrix containing weights on its main diagonal for each of these conditions and such weights can be set as a relatively large number (e.g., 100 in practice). For more details about constrained Ncut, readers are referred to [63]. It is obvious that constrained Ncut only adds little computational load over Eq. (3). After eigen-solving, the partition is conducted by thresholding the eigenvector corresponding to *the smallest non-zero eigenvalue of* (6). Since vertices of the graph correspond to superpixels, the partition of the graph yields segments constituted of superpixels.

Some results of such partition are shown in Fig. 4. Observing Fig. 4, one can see applying Ncut on the constructed graph sometimes generates two segments that keep the objects holistically, making it easy for identifying salient objects. Note if the segmentation methods [22,24] are employed to achieve equal bi-segmentation, the involved parameters have to be carefully tuned. On the other hand, it can be noticed from Fig. 4 that neither Ncut nor constrained Ncut always generate satisfactory segments, e.g., objects could adjoin other background regions despite the cost of normalized cut (2) is minimized. Therefore, further partition is essential.

Fig. 5 shows our partition strategy. Because splitting two dissimilar vertex sets in one segment yields a low Ncut cost computed by (2), in the next hierarchy we choose the region that renders the lowest Ncut cost (highlighted by the red arrows in Fig. 5) in the previous hierarchy and further implement Ncut Eq. (3), whose affinity matrix can be formed by taking elements in certain rows and columns of the original affinity matrix $\mathbf{W}$. Gradually as Ncut is applied multiple times, a binary structured tree is generated as shown in Fig. 5. Note that in each hierarchy, only two new segments are generated from the previous hierarchy. For notation convenience, we use R_n^m to specify a segment in the tree, indicating that it firstly appears in the mth hierarchy with label n.

Stopping criterion: To retain objects holistically, an ideal time to terminate the partition process is when entire objects are separated from the background. Unfortunately, it is practically difficult to obtain such a criterion due to the lack of prior knowledge as well as the ambiguity in the definition of what an object is. Instead of choosing a single hierarchy level to compute the image saliency,

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

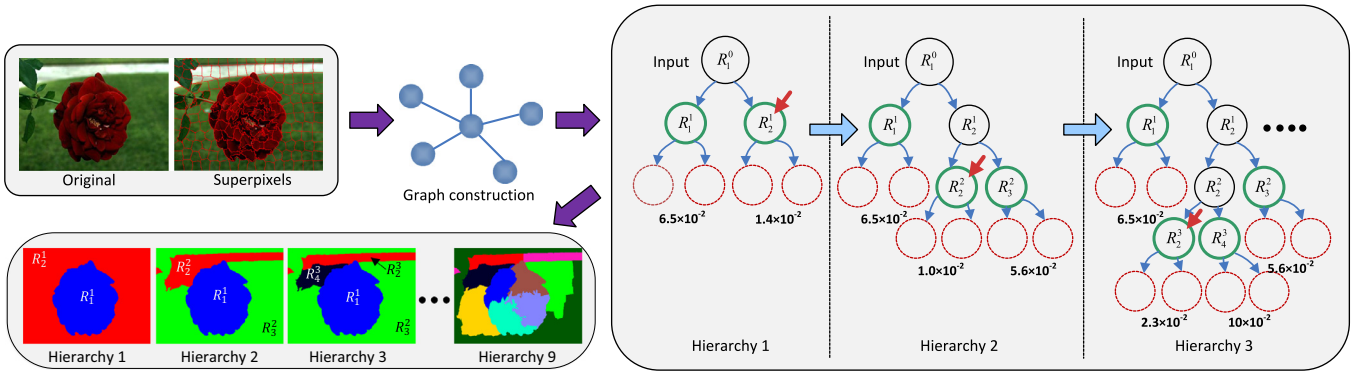

Fig. 5. Hierarchical spectral partition. A segment with the lowest Ncut cost is selected from the current hierarchy (indicated by the red arrow) to continue the next partition. In this figure, the green solid circles indicate tree leaves (segments) generated in the current hierarchy. The red dotted circles stand for two segments generated if Ncut is applied. Meanwhile the cost of the Ncut is shown by the text below the red dotted circles. The entire flower, namely R_1^1 will get split in the 7th hierarchy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a trade-off is to use all emerging hierarchies. Without loss of generality, assuming the probabilistic occurrence of a holistic object is equal in all emerging hierarchies, we choose to generate a fixed number of hierarchies. The successive partition process is stopped once a pre-determined number of "leaves" of the binary tree are generated.

4.3. Saliency assignment

In this section, we describe the proposed saliency assignment approach where segments from Ncut are incorporated with regional saliency measures to model image saliency. The following criteria are considered:

1. *Saliency measure* $\mathcal{M}1$: If treating image boundary as "pseudo background" [21,35], then, segments highly differentiating from boundary superpixels of an image would be considered as more salient. Let $\mathbf{c}_n^m$ denote the mean color of R_n^m. Such color difference can be measured intuitively as the discrepancy between $\mathbf{c}_n^m$ and colors of all boundary superpixels [39,43]. However, we find it difficult to suppress background regions since a background region could also have high contrast to other background regions, e.g., in the 1st image of Fig. 4 the "land" behind the "jumping man" has high discrepancy from the "sky". To solve this, we consider only a small amount of boundary superpixels that have closest colors to R_n^m. This effectively suppresses a segment whose color occurs frequently in boundary superpixel set, but still can highlight a segment whose color is distinctive from all boundary superpixels. Specifically, for a certain segment R_n^m, $\mathcal{M}1$ is definied as:

$$\mathcal{M}1_n^m = \sum_{k=1}^{K} ||\mathbf{c}_n^m - \mathbf{c}_k^{border}||_2 / K \tag{7}$$

where $\{\mathbf{c}_k^{border}\}_{k=1}^K$ denote K boundary superpixel colors that are most similar to $\mathbf{c}_n^m$. For K, we use one quarter of the total number of boundary superpixels.

2. *Saliency measure* $\mathcal{M}2$: Human tend to have center bias when viewing an image [64]. Therefore, segments near the image center have high probability to gain more attention, i.e., being more salient. To take into account of this, $\mathcal{M}2$ is defined as:

$$\mathcal{M}2_n^m = \frac{1}{|R_n^m|} \sum_{v_i \in R_n^m} \exp\left(||\mathbf{p}_i - \mathbf{p}_c||_2^2 / \sigma^2\right) \tag{8}$$

where $\mathbf{p}_i$ is the two-dimensional location vector of superpixel v_i on the image plane, and $\mathbf{p}_c$ is the image center, $|R_n^m|$ is the number of superpixels contained in R_n^m, and σ is the standard deviation which is adaptively set to one third of the longest dimension of the input image.

3. *Saliency measure* $\mathcal{M}3$: From the figure-ground principle [65], a segment surrounded by other segments is likely to be perceived as "figure", thereby should be salient. Such "surroundedness" means a segment should have a closed outer contour and do not touch image borders. Let l_n^m be the number of borders that R_n^m touches, namely $l_n^m \in \{0, 1, 2, 3, 4\}$. Considering the special cases e.g., the half-length portrait in photography that objects could touch one of image borders, we choose to suppress the saliency of R_n^m if $l_n^m > 1$ and simply define $\mathcal{M}3$ as:

$$\mathcal{M}3_n^m = \begin{cases} 0 & \text{if } l_n^m > 1 \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

The combination of the above three hypotheses $\mathcal{M}1_n^m$, $\mathcal{M}2_n^m$, $\mathcal{M}3_n^m$ starts by normalizing $\mathcal{M}1$ into interval [0,1] by considering all segments in the current hierarchy. The combined saliency level of a segment R_n^m is finally defined as:

$$s_n^m = \mathcal{M}1_n^m \times \mathcal{M}2_n^m \times \mathcal{M}3_n^m \tag{10}$$

It is worth noting that besides the above three measures, we have tried to consider other bottom-up saliency hypotheses, e.g., the center-surround contrast by computing the color difference of a segment to its neighbor segments as a fourth measure. Unfortunately, we find they are not as effective as the above three when used in our framework. One explanation may be that, for example, center-surround contrast is very important in other saliency detection task, e.g., eye-fixation prediction, but it has minor contribution to our scheme which aims detecting salient objects (relatively big in sizes). Similar observation was also reported in [43].

By applying the final saliency criterion (10) to all segments, an intermediate saliency map is obtained from each hierarchy. It is worthy noting that because only two segments are generated in a new hierarchy and the saliency estimation ($\mathcal{M}1, \mathcal{M}2, \mathcal{M}3$) of segments does not involve other segments, so each intermediate saliency map can be computed in an incremental fashion after the first hierarchy. Fig. 6 shows an example of different hierarchies, and the corresponding saliency assignment, where all intermediate saliency maps are normalized into [0,1] for visualization.

4.4. The final saliency map

To form the final saliency map, we sum all intermediate saliency maps because all hierarchies are considered as equal probability on retaining object holism. In practical cases, we also find it less reliable to use only one hierarchy. In our system, we construct a tree structure with a moderate number of leaves (see

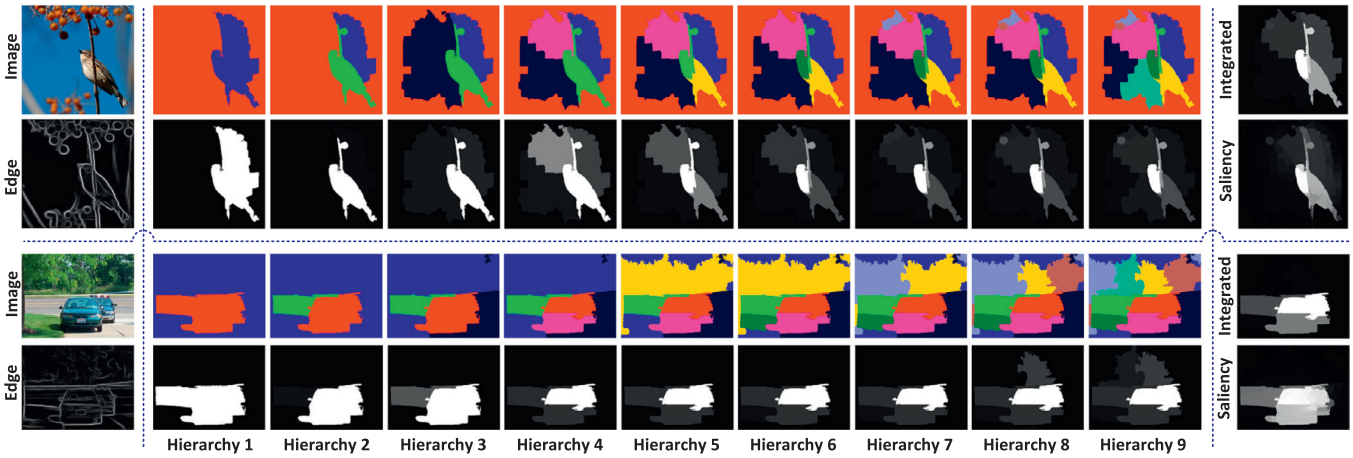

Fig. 6. Two examples of the proposed hierarchical spectral partition and saliency assignment. Column-1: the original image and edge detection. Last column: the integrated saliency map and the final saliency map. Columns 2–10: segments (labeled in different colors) from different hierarchies (hierarchy 1–9) with the corresponding intermediate saliency maps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Section 5.2.3 for details on choosing hierarchies), since an image usually does not contain many semantic contents that need to be separated.

To further refine the final result to maintain the superpixel-level details, a graph-based semi-supervised learning scheme is applied based on G to enforce manifold smoothness. Let us define s_i be the saliency value of v_i after integration, and $\mathbf{s}$ be the vector form $\mathbf{s} = [s_1, s_2, ..., s_N]^T$. The output saliency vector $\tilde{\mathbf{s}} = [\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_N]^T$ is computed by minimizing the learning energy function [66] as follows:

$$E(\tilde{\mathbf{s}}) = \frac{1}{N}\left\{\sum_{i,j=1}^{N} w_{ij}\left(\tilde{s}_i - \tilde{s}_j\right)^2 + \mu \sum_{i=1}^{N} d_i(\tilde{s}_i - s_i)^2\right\} \tag{11}$$

where w_{ij} is the edge weight between v_i and v_j as defined in (4), and $d_i = \sum_j w_{ij}$ is the degree of v_i. $\mu = 0.01$ is set for obtaining sufficient smoothing. In (11), the first summation enforces similar graph neighbors to take similar learning output, while the second summation requires the output fitting to the integrated result. The solution to (11) is computed analytically as:

$$\tilde{\mathbf{s}} = \left(\mathbf{I} - \frac{1}{1+\mu}\mathbf{D}^{-1}\mathbf{W}\right)^{-1}\mathbf{s} \tag{12}$$

where $\mathbf{I}$ is an $N \times N$ identity matrix, $\mathbf{D}$ and $\mathbf{W}$ are the degree matrix and the affinity matrix of G, respectively. The last column of Fig. 6 shows the linearly integrated map and the final saliency map. As summary, the pseudo code of the proposed spectral salient object detection algorithm is given in Algorithm 1.

Algorithm 1 Spectral Salient Object Detection.

Input: Image $\mathbf{I}$, graph affinity $\mathbf{W}$, boundary constraint matrix $\mathbf{U}$, hierarchy number t;
Output: Superpixel-wise saliency $\tilde{\mathbf{s}}$;
Initialization: hierarchy index $\tau = 0$, hierarchy $R = \{R_1^0\}$ (R_1^0 refers to the entire image), the regional saliency scores $\mathcal{S} = \emptyset$ of R;
while $\tau < t$ **do**
  $\tau \leftarrow \tau + 1$;
  if $\tau == 1$ **then**
    Divide R_1^0 into R_1^1 and R_2^1 using constrained Ncut (Eq. (6));
    Update R via: $R \leftarrow R/\{R_1^0\}$, and $R \leftarrow R \bigcup \{R_1^1, R_2^1\}$;
    Compute $\mathcal{S} = \{s_1^1, s_2^1\}$ by Eq. (10), and formulate the intermediate saliency map;
  else
    Denote the segment rendering the lowest Ncut cost in R as R_n^m;
    Divide R_n^m into R_n^τ and $R_{\tau+1}^\tau$ using Ncut (Eq. (3)), where the graph is a sub-graph of G and its affinity matrix is derived from $\mathbf{W}$;
    Update R via: $R \leftarrow R/\{R_n^m\}$, and $R \leftarrow R \bigcup \{R_n^\tau, R_{\tau+1}^\tau\}$;
    Compute $\mathcal{S} \leftarrow \mathcal{S}/\{s_n^m\}$, $\mathcal{S} \leftarrow \mathcal{S} \bigcup \{s_n^\tau, s_{\tau+1}^\tau\}$ by Eq. (10), and formulate the intermediate saliency map;
  end if
end while
Sum hierarchical intermediate saliency maps to get superpixel-wise saliency $\mathbf{s}$.
Apply graph-based semi-supervised learning Eq. (12) to $\mathbf{s}$ and obtain $\tilde{\mathbf{s}}$.

5. Experiments and results

In this section, we first describe the experiments and evaluations conducted to test the effectiveness of the proposed method. We then describe the comparisons made with several state-of-the-art methods on five datasets, with results included.

5.1. Setup

5.1.1. Datasets and a reference model

Five datasets having ground truth salient object masks were used for our tests, including ASD [8], ECSSD [58], MSRA+, PASCAL-S [67] and DUT-OMRON [21]. ASD [8] contains 1000 images selected from the MSRA database [42]. Pixel-level ground truth is provided by [8]. In this dataset, each image usually contains one single object. ECSSD [58] contains 1000 images with diversified patterns in both foreground and background. Ground truth masks are produced by five subjects. MSRA+ formulated by us contains 4000 images that belong to MSRA [42] but are not contained in ASD, to avoid duplicated evaluation. Mostly, each image of MSRA+ has an unambiguous salient object, and MSRA+ is expected more challenging than ASD. DUT-OMRON [21] contains 5168 images manually selected from more than 140,000 images. Images of this dataset have one or more salient objects and relatively complex background. Three types of ground truth annotations are avaliable, i.e., bounding box, eye-tracking, and pixel-wise masks. PASCAL-S [67] is built on the validation set of PASCAL VOC 2010 segmentation challenge and contains 850 natural images. The ground truth

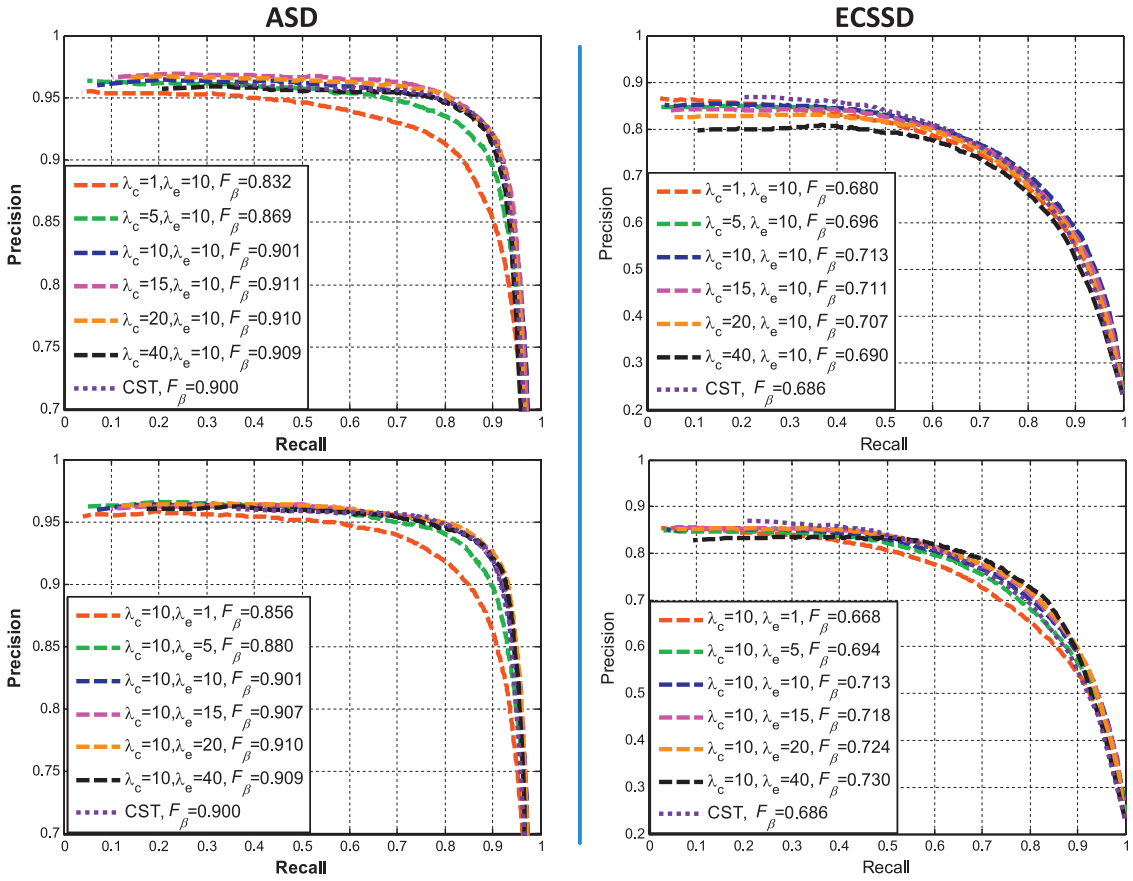

Fig. 7. Sensitivity tests on the performance of the proposed method with respect to the parameters λ_c (top) and λ_e (down) in (5). The performance of CST [47] is also shown.

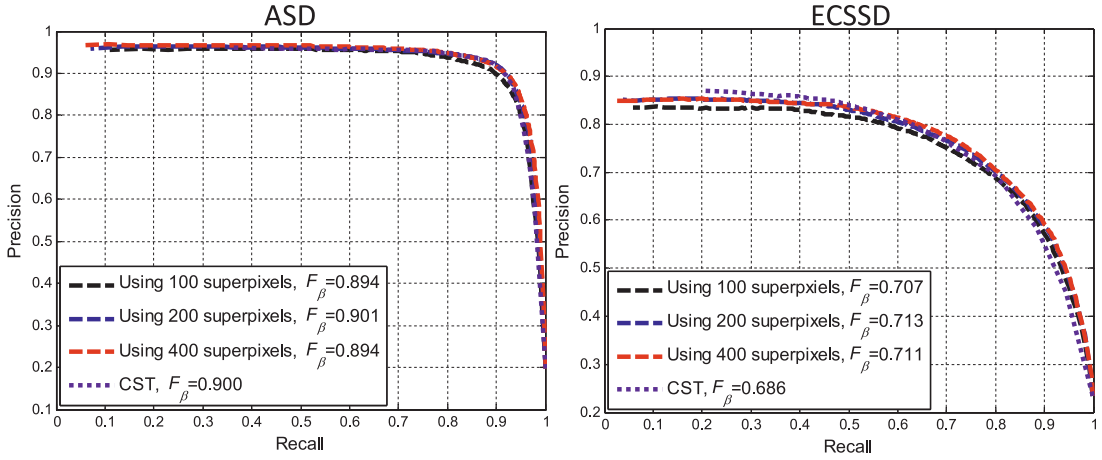

Fig. 8. Robustness of the proposed method against superpixel numbers. The performance of CST [47] is also shown.

annotations are labeled by 12 subjects on full segmentation of all objects in the images. PASCAL-S is designed to avoid dataset bias.

To show the robustness of our method, we include a recent method, namely CST (Correspondence Saliency Transfer) [47], as a reference in the following evaluation. Since CST requires a large reference dataset from which saliency information is transferred, we choose DUT-OMRON as suggested in [47]. As a result, CST was not evaluated on DUT-OMRON.

5.1.2. Parameter settings

To test the proposed methods, several parameters need to be chosen empirically in our tests. For the graph affinity computation, $\lambda_c = 10$ and $\lambda_e = 10$ were used. All edge maps for testing the proposed method were obtained from a state-of-the-art edge detector [62] that is based on the structured random forest. The structured learning used enables the edge detector to consider both color and texture information of image data. Other edge detectors can also be used to replace the above edge detector as long as generates reasonably good edge maps. The number of hierarchies was fixed as $t = 9$ in all our tests, leading to 10 segments in the last hierarchy. We use $N = 200$ superpixels on each image. Evaluations by varying various parameters are included in the next subsection. In the following text, the proposed method is abbreviated as "**SS**" (short for **S**pectral **S**alient object detection).

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

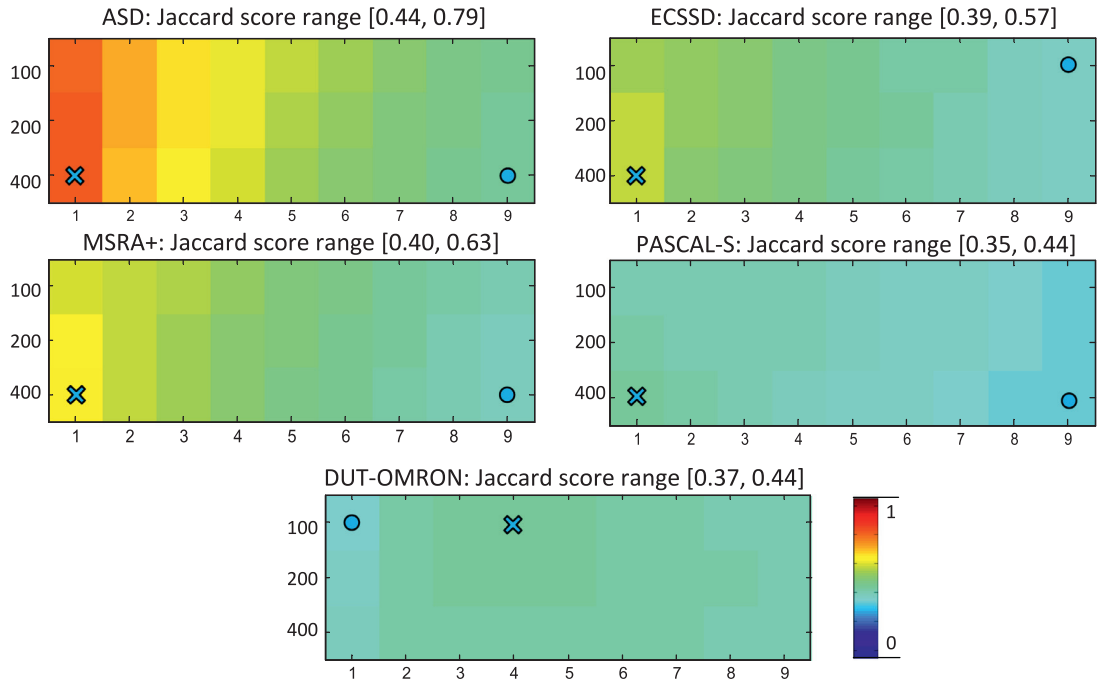

Fig. 9. Validation on Ncut segmentation by varying hierarchy (τ) and superpixel number (N). Vertical axis: the number of superpixels N (from 100 to 400); horizontal axis: the index of hierarchy level ($\tau \in [1, 9]$). A block in the matrices with warm color implies a high Jaccard coefficient $C_{Jaccard}$. The maximums in the matrices are indicated by blue "x" whereas the minimums in the matrices are indicated by blue "o". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

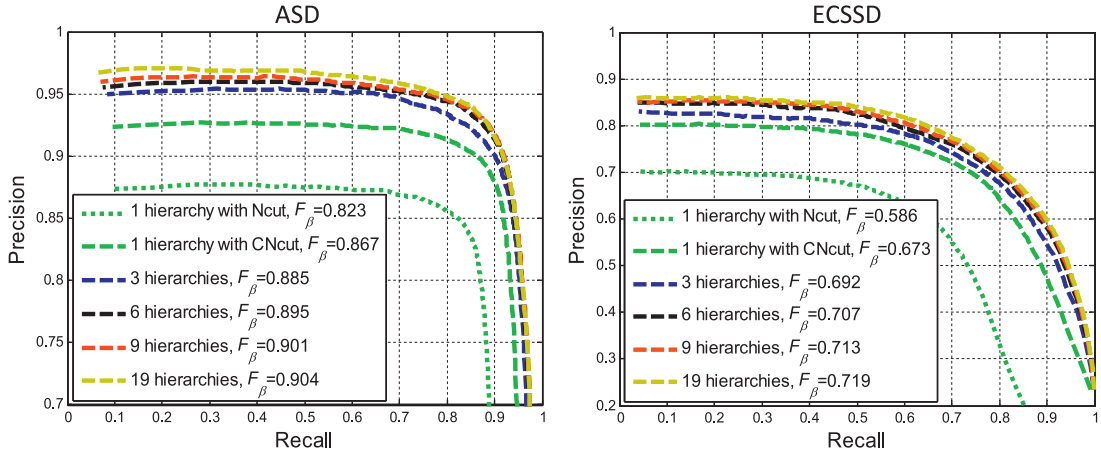

Fig. 10. The performance versus the number of hierarchies t. Left: results from ASD dataset; Right: results from ECSSD dataset.

Table 1
Average CPU seconds on ASD dataset, all based on using Matlab code.

Methods	PCA	MR	DRFI	DSR	wCtrO	ST	TLLT	CST	SS
Time (s)	4.8	0.5	5.7	4.4	0.2	75.6	2.2	41.8	**1.4**

5.1.3. Metrics for performance evaluation

Given a saliency map S_{map} and the ground truth map Gt, three widely adopted metrics are used for the evaluation of the proposed method. They are briefly introduced as follows:

1. *Precision–Recall (PR)* [8,9] is defined as:

$$\text{Precision}(T) = \frac{|M(T) \cap Gt|}{|M(T)|}, \quad \text{Recall}(T) = \frac{|M(T) \cap Gt|}{|Gt|} \tag{13}$$

where $M(T)$ is the binary mask obtained by directly thresholding the saliency map S_{map} with the threshold T, and $|\cdot|$ is the total area of the mask(s) inside the map. By varying T from 0 to 255, a precision–recall curve can be obtained.

2. *F-measure* (F_β) [8,9] is defined as:

$$F_\beta = \frac{(1+\beta^2)\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \tag{14}$$

where β is the weight between the precision and the recall. $\beta^2 = 0.3$ is usually set since the precision is often weighted more than the recall [8]. In order to get a single-valued score, existing works usually first binarize a saliency map into a foreground mask map, leading to a single precision and recall

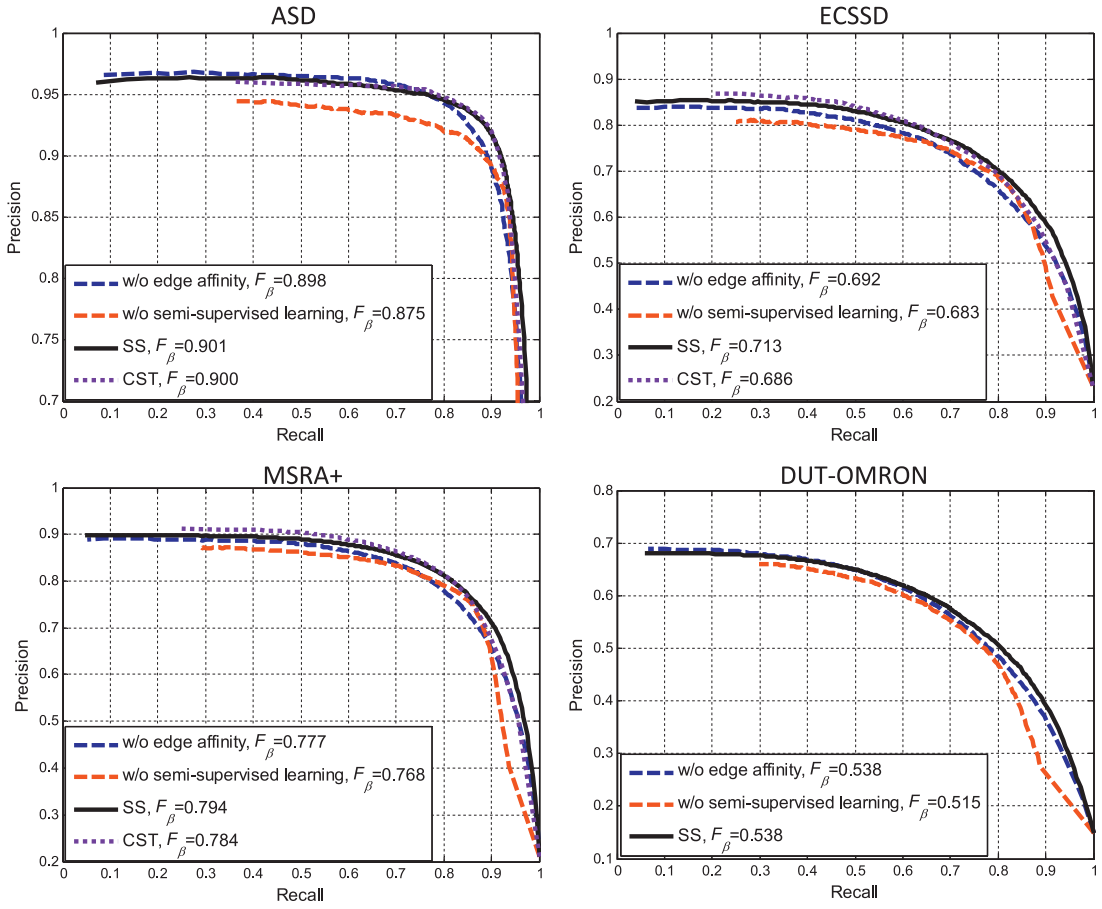

Fig. 11. Validation of edge affinity and graph-based semi-supervised learning. Results are achieved by removing individual components from our complete system. In this figure, SS stands for our complete method. The performance of CST [47] is also shown.

values. The most common way to do this is by adaptive thresholding, where the adaptive threshold is defined as *two times of the mean value* of the saliency map [8,9].

3. *Mean absolute error (MAE)* [20,59] is defined as:

$$MAE = \frac{1}{W \cdot H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S_{map}(x, y) - Gt(x, y)| \quad (15)$$

where $S_{map}(x, y)$ and $Gt(x, y)$ correspond to the saliency value and ground truth value at pixel location (x, y). W and H are the width and height of S_{map}.

Among the above three metrics, high precision–recall curves, high F_β, and low MAE indicate good saliency models.

5.2. Tests and evaluation on the proposed method

5.2.1. Sensitivity of the performance with respect to parameters

The two damping factors λ_c and λ_e in (5) play an important role in controlling the graph affinity, and therefore the Ncut results. The suitable ranges of these two parameters were tested. We conducted the tests on ASD and ECSSD. Precision–recall curves were obtained by varying λ_c while fixing $\lambda_e = 10$, and then varying λ_e while fixing $\lambda_c = 10$. The resulting precision–recall curves with F-measure F_β are shown in Fig. 7. From Fig. 7, one can observe that too small λ_c or λ_e (e.g., 1 or 5) leads to degenerated performance. The reason is that these exponential damping factors in (4) are not adequate to render a small affinity when superpixels have relatively large color differences or intervening edge magnitudes. When λ_c and λ_e values increase up to approximately 10, the performance improves. However, a further increase of these values provides very little improvement on both datasets, and sometimes even leads to a bit degeneration (see Fig. 7, when $\lambda_c = 40$). In all, we find the performance is stable when parameters λ_c and λ_e are within [10, 20]. We set $\lambda_c = 10$ and $\lambda_e = 10$ in the experiments.

Besides, we have found our method is generally robust to superpixel numbers. Precision–recall curves with F-measure F_β on ASD and ECSSD by tuning the superpixel number N from 100 to 400 are documented in Fig. 8. The overall perform changes slightly even if one uses $\times 4$ superpixels to perform detection.

5.2.2. Validation on Ncut segmentation

Ncut plays an essential role to retain an object holistically through hierarchies. To see how well the hierarchies generated by Ncut could retain object holism, we adopt the well-known Jaccard coefficient $C_{Jaccard}$ as the measure. Given two binary masks $\mathcal{R}$ and $\mathcal{R}'$, $C_{Jaccard}$ is defined as:

$$C_{Jaccard} = \frac{|\mathcal{R} \bigcap \mathcal{R}'|}{|R \bigcup R'|}$$

The evaluation was conducted by varying both the hierarchy and the number of superpixels. Giving a specific number $N \in \{100, 200, 400\}$ and a hierarchy index $\tau \in [1, 9]$, the segment in the current hierarchy having the highest $C_{Jaccard}$ score with the ground truth mask was found. For certain N and t, $C_{Jaccard}$ scores were averaged over an entire dataset.

Fig. 9 shows the evaluation results (in the matrix form) on all five datasets. Observing Fig. 9, one can see that the Ncut in the first hierarchy ($\tau = 1$) has, in general, the strongest

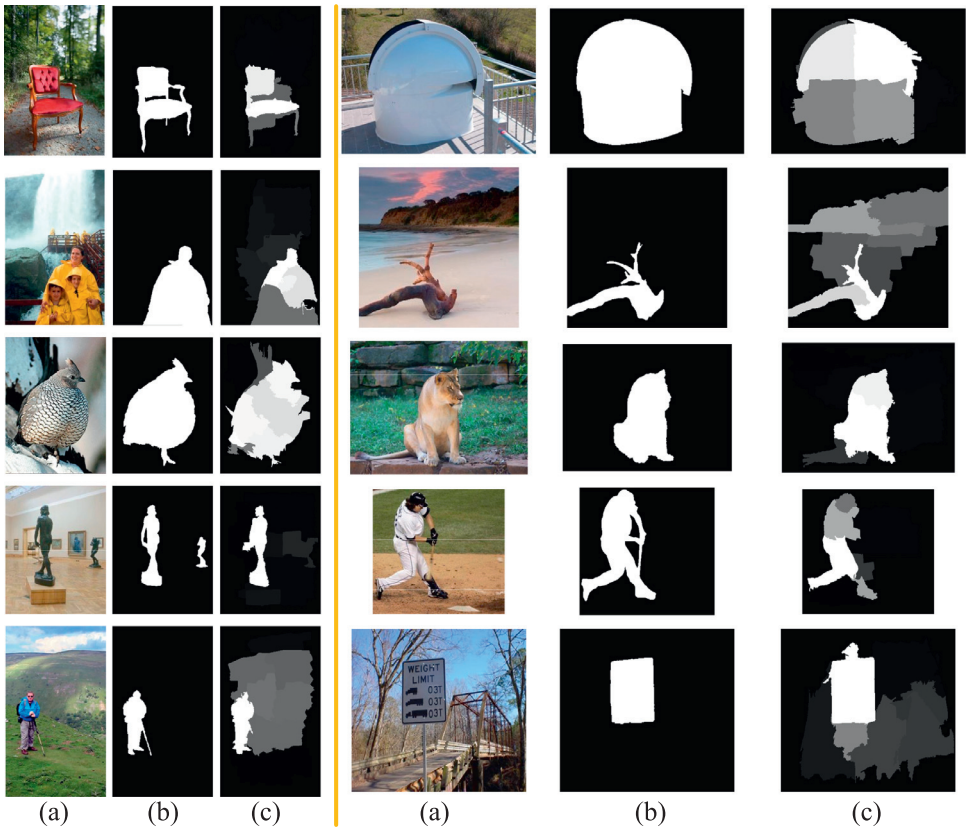

Fig. 12. Visual results of the proposed method without using semi-supervised learning. (a) Original images. (b) Ground truth masks. (c) Results without semi-supervised learning refinement, where object holism is kept by integrating Ncut and saliency assignment from multiple hierarchies.

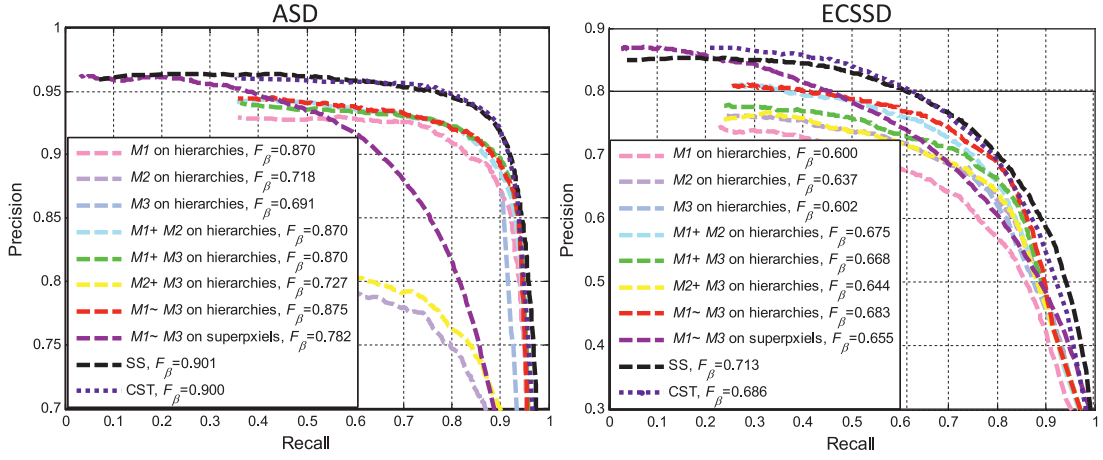

Fig. 13. Behavior of individual saliency measures (without using semi-supervised learning) on ASD (left) and ECSSD (right) datasets. In this figure, SS stands for our complete method. The performance of CST [47] is also shown. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

capability of segmenting entire salient objects from the background. When τ increases, the Jaccard coefficient $C_{Jaccard}$ decreases as the holistic object is less likely to be retained. This indicates that one should use a moderate number of top hierarchies. Furthermore, the results on ASD has the highest $C_{Jaccard} = 0.79$, implying this dataset is the easiest among the five datasets for the proposed method to retain object holism. The following $C_{Jaccard}$ scores are from MSRA+ ($C_{Jaccard} = 0.63$) and ECSSD ($C_{Jaccard} = 0.57$). Two most difficult datasets are PASCAL-S and DUT-OMRON, where $C_{Jaccard}$ reaches only 0.44 on both datasets. This indicates that much room still exists for further improving holism-retained segmentation for saliency detection, especially on difficult datasets like PASCAL-S and DUT-OMRON. Another interesting observation is that on DUT-OMRON, the first hierarchy ($\tau = 1$) seems not as effective as usual to achieve the highest $C_{Jaccard}$ scores. This is caused by complex contents in DUT-OMRON, which make it very difficult to keep holistic objects by using Ncut once. Therefore further partitions are required and the maximum $C_{Jaccard}$ appears when $\tau = 4$.

5.2.3. Performance versus the number of hierarchies

Fig. 10 shows the experimental results from using different hierarchies on ASD and ECSSD datasets. From Fig. 10, one can observe that using the first hierarchy (i.e., partition an image

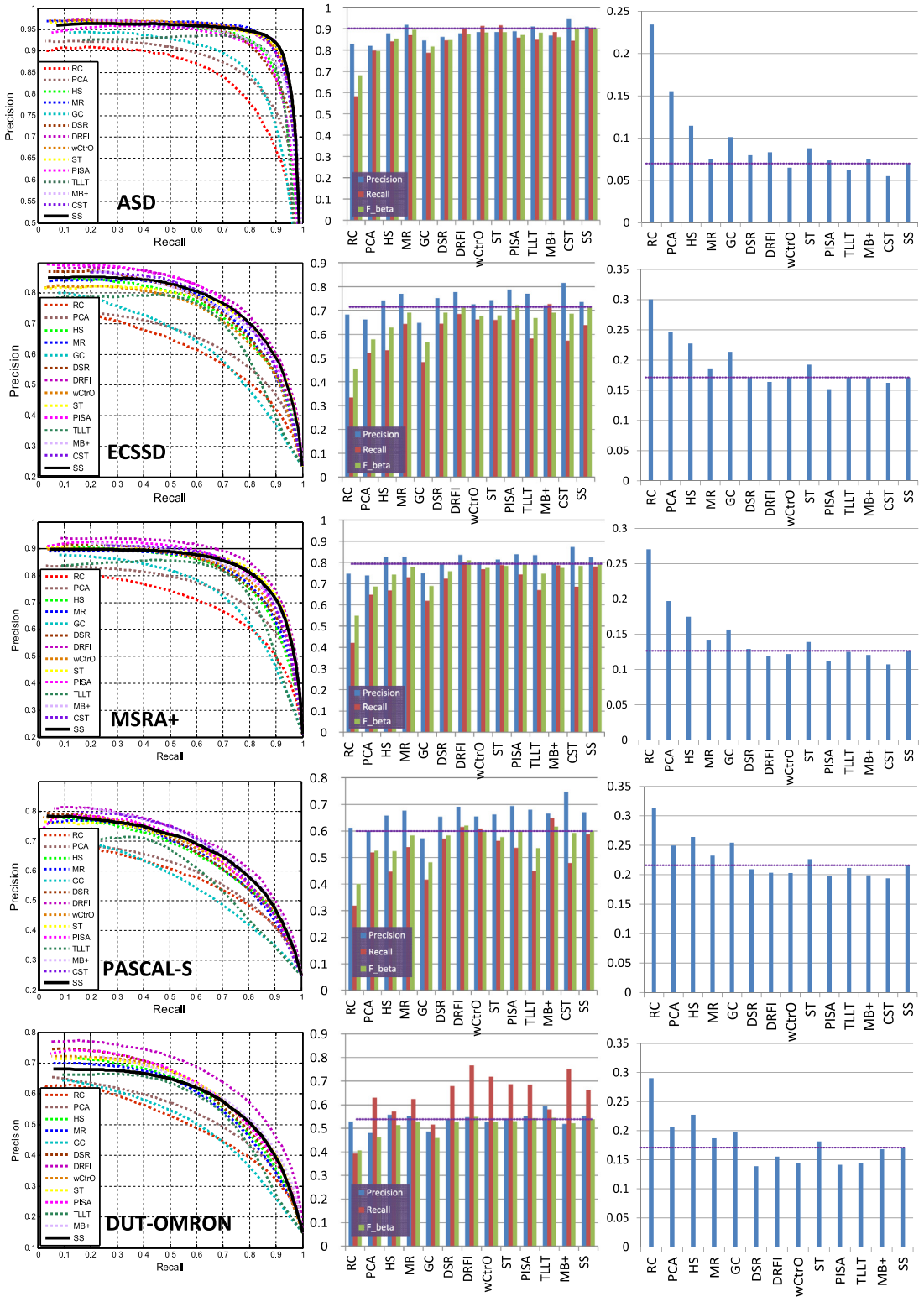

Fig. 14. Performance evaluations from the proposed method SS and the 13 existing methods on five benchmark datasets. Column-1: precision–recall curves; Column-2: precision, recall and F-measure by adaptive threshold; Column-3: mean absolute error (MAE). Rows from top to down: the results on ASD, ECSSD, MSRA+, PASCAL-S, and DUT-OMRON datasets.

only into two segments) is shown to be less effective as it is insufficient for separating a whole object from the background. This is clearly validated in Fig. 4. On the other hand, the performance of using constrained Ncut (CNcut in Fig. 10) is much higher than using traditional Ncut. Such quantitative results are also consistent with the qualitative observations in Fig. 4. Increasing the number of hierarchies from $t = 1$ to $t = 3$ improves the performance with a large margin, indicating the effectiveness of integrating multiple hierarchies. When increasing t from 6 to 9 and further to 19, the performance gain is very limited. Therefore, we use $t = 9$ hierarchies in the experiments.

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

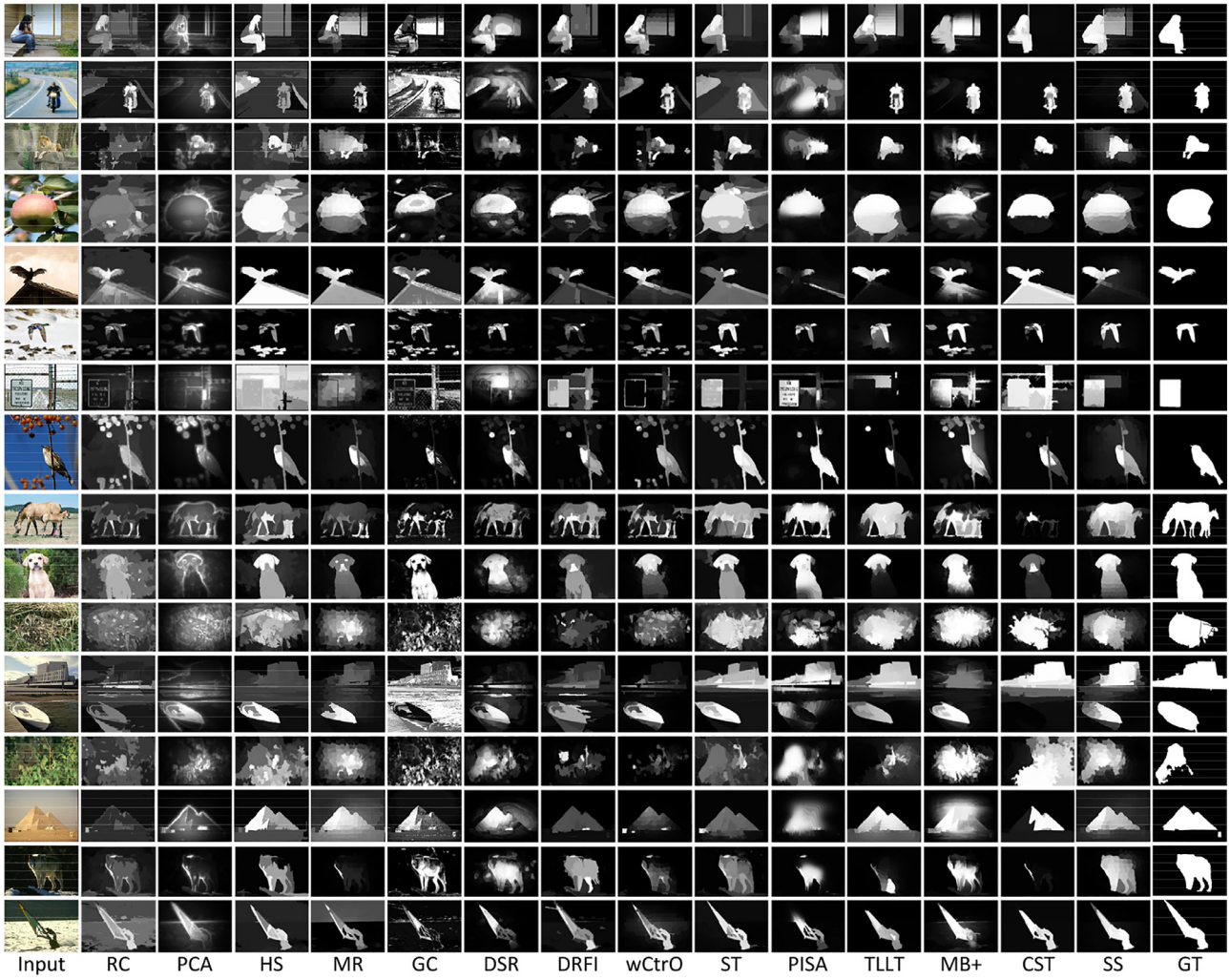

Fig. 15. Visual comparisons of SS to 13 existing methods.

5.2.4. Validation of individual modules

We have validated the individual impact of edge detection and semi-supervised learning. Experiments were conducted by removing these two modules from our complete method. To remove edge detection, we use only color affinity (i.e., S_{ij}^{w}) instead of the square root of both color and edge affinity. To remove semi-supervised learning, we just use integrated saliency maps from hierarchies. Fig. 11 shows the resulting precision–recall curves with F-measure F_β on four databases: ASD, ECSSD, MSRA+, and DUT-OMRON. From Fig. 11 one can see that introducing edge affinity and semi-supervised has moderate improvement over the entire system while removing them leads to certain performance drop. Fig. 12 shows some visual results without semi-supervised learning, where holistic objects are highlighted from the background. This provides further support to using Ncut for retaining object holism.

The impact of the three regional saliency measures $\mathcal{M}1, \mathcal{M}2, \mathcal{M}3$ were also tested. There exist 7 possible combination of them. In addition, we show the performance of applying these measures directly to superpixels without any hierarchies. Fig. 13 shows the quantitative evaluation. One can observe that combining the three measures (the red dash curves) consistently outperforms using only one or combining two of them. Besides, without the enhancement from hierarchies, jointly computing the three measures on superpixels (the purple dash curves) can hardly achieve high precision under high recall. This reveals that in such a case holistic salient objects are not detected.

5.3. Comparisons with 13 state-of-the-art methods

On the five aforementioned datasets, we test and compare the proposed method SS with 13 state-of-the-art methods including: RC (Region Contrast) [9], PCA [32], HS (Hierarchical Saliency) [58], MR (Manifold Ranking) [21], GC (Global Cue) [59], DSR (Dense and Sparse Reconstruction) [33], DRFI (Discriminative Regional Feature Integration) [43], wCtrO (background weighted Contrast with Optimization) [37], Saliency Tree [60], PISA (Pixelwise Image Saliency by Aggregation) [31], TLLT (Teaching-to-Learn and Learning-to-Teach saliency) [38], MB+ (enhanced Minimum Barrier saliency) [39], CST (Correspondence Saliency Transfer) [47]. All of the contenders are unsupervised computational saliency models except DRFI[1] and CST[2]. Note that DRFI, wCtro, ST, DSR are among the best scoring models in a recent benchmark [68].

Quantitative comparisons are shown in Fig. 14. The precision–recall (PR) curves obtained by using a fixed threshold T, $T \in [0, 255]$

[1] DRFI is the best model in a recent benchmark [68] thanks to DRFI's supervised learning strategy.

[2] CST is implemented using the code from http://github.com/shenjianbing/saliencytransfer.

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

are shown in Fig. 14 (1st column). Observing the results, the proposed method achieves state-of-the-art performance and outperforms many contenders on the five benchmark databases. Interestingly, SS achieves top precision values when recall is greater than 0.6, may be due to the use of Ncut and entire objects are uniformly highlighted. This is very useful since for some applications such as Saliency Cut [9], the initial masks are usually generated by selecting a fixed threshold under high recall [9]. While DRFI which is based on supervised learning outperforms SS on MSRA+, PASCAL-S, DUT-OMRON, ECSSD, SS shows superior performance to DRFI on ASD although it works without supervised learning.

Another evaluation is carried out by using an adaptive threshold to binarize saliency maps [8,9]. The resultant single precision, recall and F-measure scores of each method are shown in Fig. 14 (2nd column). Notably, on F-measure, the proposed method achieves *top three* on ASD, ECSSD, MSRA+ and PASCAL-S. On DUT-OMRON the proposed method performs less satisfactory (rank 4th) as comparing to the other four datasets, however it still outperforms DSR, ST and wCtrO. Comparing the performance under the MAE metric as shown in Fig. 14 (3rd column), SS does not achieve the lowest MAE on any database, but the performance is still comparable. Fig. 15 further shows some visual comparisons, where SS performs well on highlighting large salient objects, and the saliency maps from SS are closer to the ground truth with more visual agreement. This should be attributed to the leverage of Ncut and infers potential advantages of SS over existing models.

5.4. Computational efficiency

The average running time of SS on ASD (default settings mentioned in Section 5.1) is 1.4 s, where 16% of the time is taken by superpixel segmentation, and 27% by multi-scale random forest edge detection. The hierarchical spectral partition, incremental saliency computation, and final semi-supervised learning in total take the remaining 57% of the time. Table 1 shows the average running time of SS compared to several state-of-the-art models[3] on a laptop equipped with an Intel i7-4720HQ 2.6 GHz CPU and 8 GB memory by un-optimized Matlab code.

Regarding to the eigen-solving problems in SS, the brutal-force solution has time complexity $O(N^3)$, where N denotes the number of graph/sub-graph nodes. Fortunately, our superpixel graph could be deemed approximately as a regular graph and its affinity matrix $\mathbf{W}$ is very sparse. Using Lanczos algorithm makes finding the second smallest eigenvector $O(\xi N)$ complexity, where ξ is the maximum number of power iterations.

5.5. Limitation

The proposed spectral salient object detection algorithm sometimes fails to highlight the most focused objects in the scenes as shown in Fig. 16. The reasons are two folds. First, Ncut is a biased cut on fairly large sets of vertices. When both large and small objects exist in the scene, our algorithm tends to pop-up the large one and meanwhile suppresses the small one. Second, in such scenes, the small objects that attract attention are related to high-level and semantic attributes. For example, “persons” are likely to grasp more attention. Since the proposed method is fully bottom-up and leverages only low-level cues, it lacks ability to find out semantic objects that are less prominent on low-level features, e.g., colors, contrast, boundary intensities.

[3] We only have compared the running time on Matlab, so those methods whose released code are in Matlab form are considered here.

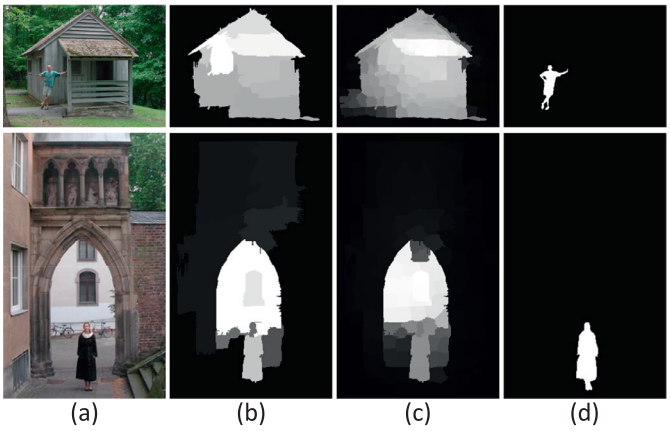

Fig. 16. Two failure cases. (a) Input images from DUT-OMRON; (b) results without semi-supervised learning refinement; (c) finaly saliency maps and (d) ground truth masks.

6. Conclusion

The proposed spectral salient object detection method, has been tested and evaluated on five benchmark datasets. By applying Ncut to a superpixel graph that captures color and edge information in image, our method is shown to be effective on detecting salient objects holistically, especially salient objects of large/moderated sizes. By using the binary segmentation tree from the hierarchical spectral partition, segments in each tree hierarchy are effectively incorporated with regional saliency metrics to estimate object saliency. Experimental results show the proposed method is capable of detecting and emphasizing salient objects uniformly by integrating intermediate saliency maps from successive hierarchies. Comparisons to 13 existing models on five benchmark datasets demonstrate the state-of-the-art performance of the proposed method under widely used criteria. Future works include exploiting some enhanced graph cut approaches to assist saliency computation, and incorporating more sophisticated saliency estimation and integration methods with our current work.

Acknowledgment

This research is partly supported by the National Science Foundation, China, under No. 61703077, 61773270, U1733111 and National Key Research and Development Program of China (2017YFB0802300, 2016YFC0801100) and the National Key Scientific Instrument and Equipment Development Projects of China (2013YQ49087904).

References

[1] A. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[2] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurbiol. 4 (1985) 219–227.
[3] A. Borji, D. Sihite, L. Itti, Salient object detection: a benchmark, in: Proceedings of European Conference on Computer Vision (ECCV), 2012.
[4] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 185–207.
[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. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2007.
[7] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Proceedings of Advances in Neural Information Processing Systems Conference (NIPS), 2005.
[8] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[9] M. Cheng, G. Zhang, N. Mitra, X. Huang, S. Hu, Global contrast based salient region detection, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[10] E. Rahtu, J. Kannala, M. Salo, J. Heikkila, Segmenting salient objects from images and videos, in: Proceedings of European Conference on Computer Vision (ECCV), 2010.
[11] L. Wang, J. Xue, N. Zheng, G. Hua, Automatic salient object extraction with contextual cue, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2011.
[12] F. Stentiford, Attention based auto image cropping, in: Proceedings of Workshop on Computational Attention and Applications (ICVS), 2007.
[13] L. Marchesotti, C. Cifarelli, G. Csurka, A framework for visual saliency detection with applications to image thumbnailing, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2009.
[14] Y. Ding, X. Jing, J. Yu, Importance filtering for image retargeting, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011.
[15] W. Wang, J. Shen, Y. Yu, K.-L. Ma, Stereoscopic thumbnail creation via efficient stereo saliency detection, IEEE Trans. Vis. Comput. Graph. 23 (8) (2017) 2014–2027.
[16] T. Chen, M. Cheng, P. Tan, A. Shamir, S. Hu, Sketch2photo: internet image montage, ACM Trans. Graph. 28 (5) (2006) 1–10.
[17] Y. Gao, M. Shi, D. Tao, C. Xu, Database saliency for fast image retrieval, IEEE Trans. Multimed. 17 (3) (2015) 359–369.
[18] K. Fu, C. Gong, J. Yang, Y. Zhou., Salient object detection via color contrast and color distribution, in: Proceedings of Asian Conference on Computer Vision (ACCV), 2012.
[19] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[20] F. Perazzi, P. Krahenbul, et al., Saliency filters: contrast based filtering for salient region detection, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[21] C. Yang, L. Zhang, et al., Saliency detection via graph-based manifold ranking, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[22] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.
[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[24] P. Felzenszwalb, D. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.
[25] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[26] Y. Weiss, Segmentation using eigenvectors: a unifying view, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 1999.
[27] T. Kim, K. Lee, S. Lee, Learning full pairwise affinities for spectral segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1690–1703.
[28] P. Mehrani, O. Veksler, Saliency segmentation based on learning and graph cut refinement, in: Proceedings of British Machine Vision Conference (BMVC), 2010.
[29] Y. Fu, J. Cheng, Z. Li, H. Lu, Saliency cuts: an automatic approach to object segmentation, in: Proceedings of IEEE International Conferenceon Pattern Recognition (ICPR), 2008.
[30] K. Fu, C. Gong, I. Gu, J. Yang, X. He, Spectral salient object detection, in: Proceedings of IEEE International Conference on Multimedia and Expo, 2014.
[31] K. Wang, L. Lin, J. Lu, C. Li, K. Shi, Pisa: pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence, IEEE Trans. Image Process. 24 (10) (2015) 3019–3033.
[32] R. Margolin, A. Tal, L. Zelnik-Manor, What makes a patch distinct, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[33] X. Li, H. Lu, L. Zhang, X. Ruan, M. Yang, Saliency detection via dense and sparse reconstruction, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013.
[34] V. Gopalakrishnan, Y. Hu, D. Rajan, Random walks on graphs for salient object detection in images, IEEE Trans. Image Process. 19 (12) (2010) 3232–3242.
[35] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, in: Proceedings of European Conference on Computer Vision (ECCV), 2012.
[36] K. Fu, C. Gong, I. Gu, J. Yang, Geodesic saliency propagation for image salient region detection, in: Proceedings of IEEE International Conference on Image Processing (ICIP), 2013.
[37] W. Zhu, S. Liang, Y. Wei, Saliency optimization from robust background detection, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[38] C. Gong, D. Tao, W. Liu, S. Maybank, M. Fang, K. Fu, J. Yang, Saliency propagation from simple to difficult, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[39] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, R. Mech, Minimum barrier salient object detection at 80 fps, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015.
[40] K. Fu, I.Y. Gu, C. Gong, J. Yang, Robust manifold-preserving diffusion-based saliency detection by adaptive weight construction, Neurocomputing 175 (2016) 336–347.
[41] S. Li, C. Zeng, S. Liu, Y. Fu, Merging fixation for saliency detection in a multilayer graph, Neurocomputing 230 (2017) 173–183.
[42] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[43] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: a discriminative regional feature integration approach, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[44] L. Mai, Y. Niu, F. Liu, Saliency aggregation: a data-driven approach, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[45] N. Tong, H. Lu, X. Ruan, M. Yang, Salient object detection via bootstrap learning, in: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[46] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, R. Mech, Unconstrained salient object detection via proposal subset optimization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5733–5742.
[47] W. Wang, J. Shen, L. Shao, F. Porikli, Correspondence driven saliency transfer, IEEE Trans. Image Process. 25 (11) (2016) 5025–5034.
[48] J. Qi, S. Dong, F. Huang, H. Lu, Saliency detection via joint modeling global shape and local consistency, Neurocomputing 222 (2017) 81–90.
[49] X. Wang, L. Zhang, L. Lin, Z. Liang, W. Zuo, Deep joint task learning for generic object extraction, in: Proceedings of Neural Information Processing Systems (NIPS), 2014.
[50] L. Wang, H. Lu, M. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[51] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[52] D. Zhang, J. Han, C. Li, J. Wang, X. Li, Detection of co-salient objects by looking deep and wide, Int. J. Comput. Vis. 120 (2) (2016) 215–232.
[53] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2017) 865–878.
[54] D. Zhang, J. Han, J. Han, L. Shao, Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining, IEEE Trans. Neural Netw. Learn. Syst. 27 (6) (2016) 1163–1176.
[55] J. Yao, X. Liu, C. Qi, Foreground detection using low rank and structured sparsity, in: Proceedings of 2014 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2014, pp. 1–6.
[56] D. Zhang, J. Han, L. Jiang, S. Ye, X. Chang, Revealing event saliency in unconstrained video collection, IEEE Trans. Image Process. 26 (4) (2017) 1746–1758.
[57] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Trans. Image Process. 24 (11) (2015) 4185–4196.
[58] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[59] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, N. Crook, Efficient salient region detection with soft image abstraction, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013.
[60] Z. Liu, W. Zou, O.L. Meur, Saliency tree: a novel saliency detection framework, IEEE Trans. Image Process. 23 (5) (2014) 1937–1952.
[61] J. Shen, Y. Du, W. Wang, X. Li, Lazy random walks for superpixel segmentation, IEEE Trans. Image Process. 23 (4) (2014) 1451–1462.
[62] P. Dollár, C. Zitnick, Structured forests for fast edge detection, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013.
[63] S. Chew, N. Cahill, Semi-supervised normalized cuts for image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1716–1723.
[64] B. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions, JoV 14 (7) (2007).
[65] S. Palmer, Vision Science: Photons to Phenomenology, The MIT Press (1999).
[66] D. Zhou, et al., Learning with local and global consistency, in: Proceedings of Neural Information Processing Systems (NIPS), 2003.
[67] Y. Li, X. Hou, C. Koch, J. Rehg, A. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[68] A. Borji, M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.

Keren Fu received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2011, and dual Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, and Chalmers University of Technology, Gothenburg, Sweden, in 2016, under the joint supervision of Prof. Jie Yang and Prof. Irene Yu-Hua Gu. He is currently a research associate professor with College of Computer Science, Sichuan University, Chengdu, China. His current research interests include visual computing, saliency analysis, and machine learning.

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028

Irene Yu-Hua Gu received the Ph.D. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1992. From 1992 to 1996, she was Research Fellow at Philips Research Institute IPO, Eindhoven, The Netherlands, and post dr. at Staffordshire University, Staffordshire, U.K., and Lecturer at the University of Birmingham, Birmingham, U.K. Since 1996, she has been with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, where she has been a full Professor since 2004. Her research interests include statistical image and video processing, object tracking and video surveillance, machine learning and deep learning, and signal processing with applications to electric power systems. Dr. Gu was an Associate Editor for the IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, and Part B: Cybernetics from 2000 to 2005, and an Associate Editor for the EURASIP Journal on Advances in Signal Processing from 2005 to 2016. She was the Chair of the IEEE Swedish Signal Processing Chapter from 2001 to 2004. She has been with the Editorial board of the Journal of Ambient Intelligence and Smart Environments since 2011.

Jie Yang received his Ph.D. from the Department of Computer Science, Hamburg University, Germany, in 1994. Currently, he is a professor at the Institute of Image Processing and Pattern recognition, Shanghai Jiao Tong University, China. He has led many research projects (e.g., National Science Foundation, 863 National High Tech. Plan), had one book published in Germany, and authored more than 200 journal papers. His major research interests are object detection and recognition, data fusion and data mining, and medical image processing.

Please cite this article as: K. Fu et al., Spectral salient object detection, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.09.028