Normalized Cut-Based Saliency Detection by Adaptive Multi-Level Region Merging

Keren Fu, *Student Member, IEEE*, Chen Gong, Irene Yu-Hua Gu, *Senior Member, IEEE*, and Jie Yang

***Abstract*— Existing salient object detection models favor over-segmented regions upon which saliency is computed. Such local regions are less effective on representing object holistically and degrade emphasis of entire salient objects. As a result, the existing methods often fail to highlight an entire object in complex background. Toward better grouping of objects and background, in this paper, we consider graph cut, more specifically, the normalized graph cut (Ncut) for saliency detection. Since the Ncut partitions a graph in a normalized energy minimization fashion, resulting eigenvectors of the Ncut contain good cluster information that may group visual contents. Motivated by this, we directly induce saliency maps via eigenvectors of the Ncut, contributing to accurate saliency estimation of visual clusters. We implement the Ncut on a graph derived from a moderate number of superpixels. This graph captures both intrinsic color and edge information of image data. Starting from the superpixels, an adaptive multi-level region merging scheme is employed to seek such cluster information from Ncut eigenvectors. With developed saliency measures for each merged region, encouraging performance is obtained after across-level integration. Experiments by comparing with 13 existing methods on four benchmark datasets, including MSRA-1000, SOD, SED, and CSSD show the proposed method, Ncut saliency, results in uniform object enhancement and achieves comparable/better performance to the state-of-the-art methods.**

***Index Terms*— Salient object detection, normalized cut, clustering, region merging, saliency map.**

I. Introduction

SALIENCY detection is a long-standing problem in computer vision and plays a critical role in understanding the mechanism of human visual attention. Applications to vision and graphics are numerous, especially in solving problems that require object-level prior such as "proto object" detection [1] and segmentation [2], [3], content based image editing [4]–[7], and image retrieval [8].

To benefit complex computer vision tasks, a new sub-field in saliency detection called *salient object/region detection* has recently emerged and drawn a lot of research attentions. It aims at compensating the drawback of previous eye fixation prediction models [1], [9]–[11] on enhancing entire objects. To achieve this, many previous salient region detection methods [12]–[16] exploit contrast and rarity properties on local *superpixels* or *regions*. Commonly employed segmentation techniques include superpixels [17], mean shift [18] or graph-based segmentation [19]. These techniques are known to be useful for eliminating background noise and reducing computation by treating each segment as a processing unit.

Such segmentation methods, which are usually employed by previous models to generate over-segmentation, take merely local color similarity into account and could lead to regions less effective on representing object holism. As a result, these regions prevent the enhancement of holistic salient objects. An example is shown in Fig. 1 where state-of-the-art methods fail to capture the entire human body as they are based on local segments, and separate the body into fragment regions. Only the head region of the girl is emphasized by most of these methods as it is somewhat more salient than other body parts. This example indicates that a better grouping to cluster an object as a whole can be useful for saliency detection as the self-organization capability of human vision system captures the entire object as a whole, as indicated by the ground truth image (GT) in Fig. 1.

Towards better grouping of objects and background, in this paper we study unsupervised graph cut for the purpose of decomposing image into visual clusters. More specifically, we focus on the Normalized graph cut (Ncut) that partitions a graph in a normalized energy minimization fashion. The Ncut can be solved in an efficient way by solving a generalized eigen-system. The resulting eigenvectors contain good cluster information. In this paper, we propose to directly induce saliency maps via eigenvectors of the Ncut, contributing to accurate saliency estimation of visual clusters. Thereby, the proposed method, Ncut saliency (NCS), highlights the entire human body (Fig. 1). To leverage Ncut eigenvectors for inducing saliency maps, we develop an adaptive multi-level region merging method to turn Ncut cluster information into regions. Saliency measures can then be easily applied to these regions. To the best of our knowledge, the Normalized graph cut (Ncut) has not been used to directly induce saliency maps in previous works.

Manuscript received January 8, 2015; revised July 23, 2015; accepted September 24, 2015. Date of publication October 1, 2015; date of current version October 23, 2015. This work was supported in part by the National Natural Science Foundation, China, under Grant 61273258 and in part by 863 Plan, China, under Grant 2015AA042308. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jean-Philippe Thiran. *(Corresponding author: Jie Yang.)*

K. Fu is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg SE-412 96, Sweden (e-mail: fkrsuper@sjtu.edu.cn).

C. Gong is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, 235 Jones Street, Ultimo, NSW 2007, Australia (e-mail: goodgongchen@sjtu.edu.cn).

I. Y.-H. Gu is with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg SE-412 96, Sweden (e-mail: irenegu@chalmers.se).

J. Yang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jieyang@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2485782

Fig. 1. An example case where state-of-the-art methods [12]–[15], [20], [21] fail to detect the entire object. About processing units, SF [15], LR [20], MR [21] work on superpixel level. RC [14] uses single level graph-based segmentation [19]. DRFI [13] uses multi-scale graph-based segmentation [19]. HS [12] uses a hierarchical segmentation method. The proposed method (NCS) uses the eigenvectors of Ncut and succeeds in highlighting the entire object.

One of related work that is worth mentioning is from [12], as it is a method shares some common motivation with ours. To alleviate impact of small-scale patterns with high contrast, Yan *et al.* [12] define three levels of sizes for regions, and merge a region to its neighbors if it is smaller than pre-defined sizes. Our method differs from theirs as we discover holistic information of objects by using the Ncut. It is a non-parametric method whereas the scale parameters in [12] are manually determined. Another difference lies in the way of generating multiple segmentation levels.

It is also worth noting that there are some previous works involving both graph cut and saliency detection [2], [14], [15], [22], [23]. 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 [23], seed regions are generated by saliency detection in [1] and then "MaxFlow" is applied to solve the min-cut problem. In contrast, our saliency detection is induced by the Ncut. Thereby in our method the graph cut takes place prior to saliency detection. In [2], [14], [15], [22], and [23], 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. It is also worth noting that there are many applications besides figure-ground segmentation where saliency maps can be useful. Another concurrent work [24] uses Ncut as a post-enhancement to refine a preliminary saliency map. Their idea is somewhat similar to [23], where saliency detection is performed first. In this paper, we adopt a very different computational scheme for saliency detection as compared to [24].

The main contributions of the paper are twofold: (a) Apply the Ncut to salient region detection, and induce a saliency map by Ncut eigenvectors for better visual clustering; (b) Embed saliency detection in an adaptive multi-level merging scheme to discover cluster information conveyed by Ncut eigenvectors.

A preliminary version of our system was described in [25]. This paper differs from [25] in several ways. First, we start the problem from Ncut viewpoint whereas [25] starts from multi-level viewpoint. For technical details, the framework in this paper is the same as [25], but modifications have been made to improve the performance. For example, 2-ring graph topology and boundary connection are adopted in this paper whereas [25] only uses local neighbor graph. Edge detection is incorporated for graph affinity in this paper whereas [25] only considers superpixel color difference. Besides, more technical details and extensive results are provided, showing the proposed method outperforms the previous method in [25].

The reminder of the paper is organized as follows. Section II describes the related work on salient object detection. Section III briefly reviews the fundamental of Normalized graph cut, based on which we build our method. Section IV gives detailed description of the proposed method by adaptive multi-level region merging. Experimental results and performance evaluation are included in Section V. Finally, conclusions are drawn in Section VI.

II. Related Work

The literature of salient object detection is huge and we refer readers to comprehensive surveys [11], [26]. There are a number of ways to classify existing methods. We classify previous works in terms of processing units upon which saliency is computed. This is based on the starting point of this paper.

A. Pixel-Based

Zhai and Shah [27] introduce image histograms that only model the luminance channel to calculate pixel-level saliency. Pixel-level spatial saliency is measured as the luminance contrast between image pixels. Achanta *et al.* [28] provide a saliency approximation by subtracting the average color from low-pass filtered result of image. This operation is equivalent to combining center-surround differences of all bandwidth to detect objects of different sizes. Shi *et al.* [29] compute pixel-wise image saliency by aggregating complementary appearance contrast measures with spatial priors. Liu *et al.* [30] segment salient objects by aggregating pixel saliency cues in a conditional random field. Their saliency cues include center-surround histogram contrast, saliency maps from the spectral residual method [1], and color spatial distributions. The linear weights for those cues are learned under the Maximized Likelihood (ML) criteria by tree-weighted belief propagation. Cheng *et al.* [31] measure saliency by hierarchical soft abstraction. They form a 4-layer hierarchical structure (respectively are pixel layer, histogram layer, GMM layer and clustering layer) with an index table to associate cross-layer relations efficiently. Saliency estimation using color contrast and distribution are conducted on the coarse layers and then propagated to the pixel layer. The drawback of using pixels as the basic units is that simple computation of color contrast [27], [28] is less satisfactory for complex scenes whereas incorporating holistic pixel-wise information like [30] requires heavy computation. Additionally, it is easily affected by small-scale noise in an image.

B. Patch/Region/Superpixel-Based

Patches have also been considered for computing object saliency in early time. Gopalakrishnan *et al.* [32] perform random walks on graphs constructed from patches to find salient objects. The global pop-out and compactness properties of salient objects are modeled in random walks by the equilibrium access time performed on a complete and k-regular graph. Goferman *et al.* [7] combine local and global features to estimate the patch saliency in multi-scales. To consider both local and global factors, they compute saliency of a certain patch as its contrast to the nearest patches in the image. Under this framework, inner parts of an object are often attenuated due to the edge preference. Margolin *et al.* [33] define patch distinctness as L1 norm in PCA coordinates and combine it with color distinctness. Unfortunately, using local patch contrast [7], [33] could cause edges highlighted. Besides, patches are less good on edge-preserving rendering since they could contain edges or large color variation inside.

To overcome the disadvantage of patches, much effort has focused on pre-segmentation techniques to obtain edge-aware superpixels/regions and shown success in eliminating unnecessary details and producing high quality saliency detection. Examples in this category include: Cheng *et al.* [14] extend the method in [27] and incorporate color histograms. A regional contrast saliency measure is proposed in [14] as the color contrast to other regions. Perazzi *et al.* [15] propose a saliency filter that formulates complete contrast and saliency estimation using high dimensional Gaussian filters. A Bayesian framework is adopted in [34]. First, saliency points are applied to obtain a coarse location of the saliency region. Based on the rough region, a prior map is computed for the Bayesian model. Wei *et al.* [35] treat boundary parts of an image as the background. The superpixel saliency is defined as the shortest geodesic distance to image boundary. Since a salient object often does not adjoin image boundary, the geodesic distance between image boundary and the object should be large. Shen and Wu [20] solve saliency detection as a low rank matrix recovery problem, where salient objects are represented by a sparse matrix (noise) while the background is indicated by a low rank matrix. This sparse and low rank assumption however cannot be satisfied in complex scenes, leading to unsatisfactory results. Yang *et al.* [21] utilize similar boundary priors as [35] however propagate saliency via graph-based manifold ranking from four image borders separately. Four saliency maps generated are then multiplied to achieve the final one. Despite these efforts, as mentioned in section I, small local segments alone hardly reflect object holism and global meanings.

C. Multi-Scale Based

Since over-partitioned segments have limited capabilities in modeling holistic properties as shown in Fig. 1, a number of recent approaches employ multi-scale segmentation schemes to extract non-local contrast information. Yan *et al.* [12] merge regions according to user-defined scales (e.g., 3 size scales in their case) to eliminate small-size distracters. Jiang *et al.* [13] learn several optimal scales from a series of manually defined scales using a least-square estimator. However, segmentation methods used above are still based on local clustering and do not reflect holistic information of objects. Although multi-level segmentation is used in our method, we exploit the Ncut for saliency detection and embed saliency estimation in an adaptive multi-level region merging scheme.

III. The Normalized Cut: Review

The Normalized graph cut (Ncut) proposed by Shi and Malik [36] normalizes the cost of graph cut by using the total edge connections towards all nodes in a graph. Given a similarity graph $G = (V, E)$ (a graph whose edges measure the similarity between vertices), let $\mathbf{W}$ be its adjacency matrix, $\mathbf{D}$ be its degree matrix (a diagonal matrix with diagonal entry $d_i = \sum_j w_{ij}$, where w_{ij} is entry of $\mathbf{W}$), and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ be its Laplacian matrix. The cost of a cut between two subsets A and B is defined as $cut(A, B) := \sum_{v_i \in A, v_j \in B} w_{ij}$. Let $\bar{A}$ be the complement of A. For a given number k of subsets, the Ncut aims to choose a partition $A_1, \ldots, A_k$ that minimizes:

$$Ncut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{assoc(A_i, V)} \tag{1}$$

where $assoc(A_i, V) := \sum_{v_i \in A, v_j \in V} w_{ij}$ is a measure of set size, i.e. the larger $|A_i|$ is, the higher $assoc(A_i, V)$ will be. The exact solution for (1) is NP hard. However, by defining a discrete indicating vector for each A_i and relaxing the discrete constraints [37], the continuous indicating vectors for the multi-cluster Ncut in (1) can be derived from the first k smallest eigenvectors of $\mathbf{D}^{-1}\mathbf{L}$, or the first k smallest eigenvectors of the following system:

$$(\mathbf{D} - \mathbf{W})\mathbf{v} = \lambda \mathbf{D}\mathbf{v} \tag{2}$$

where $\mathbf{v}$ and λ denote the eigenvector and eigenvalue. The solution of the 2-way Ncut ($k = 2$) is given by its second smallest eigenvector. For detailed mathematical derivation, see [36], [37]. The Ncut is tightly related to the spectral clustering. As the continuous indicating vectors for the multi-cluster Ncut contain cluster information, k-means clustering can be applied to those eigenvectors to obtain cluster labels, known as spectral clustering [37].

It is worth noting that comparing to the min-cut that minimizes the summation of the numerators in (1), the Ncut normalizes each cut cost as a fraction of the total edge connections to all the nodes in the graph. Due to this normalization, the Ncut is a biased cut on fairly large set of vertices. Note our goal for salient object detection is to find good grouping of visual contents, usually large objects, meanwhile prevent grouping of small clusters that are usually noise. The Ncut rightly satisfies this demand. In addition, the Ncut in (1) is a global criterion that partitions the graph in a non-parametric way. It is also efficient to compute. An example of Ncut eigenvectors generated by our method is shown in Fig. 2. It implies that Ncut eigenvectors contain good cluster information that well groups visual contents together, i.e. objects and background. Contents likely to be in the same cluster (i.e. similar pseudo colors in each eigenvector in Fig. 2) should

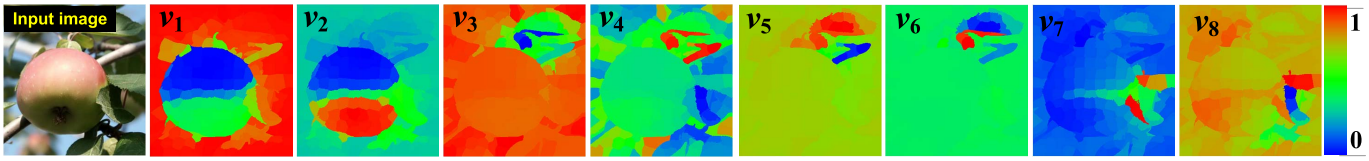


Fig. 2. An example of top eight smallest eigenvectors ($\mathbf{v}_1$ to $\mathbf{v}_8$) from Ncut. Eigenvectors are visualized in pseudo colors.

be treated as a whole and rendered the same saliency during the saliency detection.

IV. Proposed Methodologies

This section details the proposed method for salient object detection. Firstly, the graph construction is discussed in IV-A. Details on applying the Ncut to obtain cluster information are given in Section IV-B. An adaptive graph-based merging method is described in IV-C that is used to generate multi-level segmentation by discovering the cluster information from the Ncut. Regional saliency measures are introduced in IV-D. Finally, IV-E describes the formulation of the final saliency map.

A. Graph Construction for the Ncut

1) Pre-Processing: We first over-segment an input image into superpixels using the SLIC algorithm [17]. The result is a set of compact superpixels that are homogenous in color and maintain image boundaries. $N \approx 200$ superpixels are selected for each input image since such number of superpixels suffices for detecting salient objects [21]. Let the ith superpixels be R_i^0 and the corresponding average CIELab colors and spatial locations be $\mathbf{c}_i^0$ and $\mathbf{p}_i^0$, $i = 1, 2, \ldots, N$. The superscript "0" indicates the initial superpixel level. Define a graph $G = (V, E)$ whose vertices V are superpixels and E are edges (to be detailed later). Thereby, only a small set of graph nodes needs to be considered. This drastically increases the efficiency compared to a pixel-level graph [36].

2) Construct Edge Connections of Graph: To improve the Ncut performance, two extensions are made upon [25]. First, we extend the local range of graph connections by constructing connection between superpixel R_i and R_j that satisfies either $\{R_j \in N_i\}$ or $\{R_j \in N_k, R_k \in N_i\}$ (2-ring graph, illustrated in Fig. 3), and N_i denotes the neighborhood (adjacency) of superpixel R_i. This extension improves the Ncut performance (discussion for this is in Section V-D). Furthermore, we associate boundary superpixels with each other (Fig. 3). The rationale behind is that boundary superpixels have high probability to belong to a same background (e.g. in Fig. 3, in the sea behind the man).

3) Construct Edge Weights of Graph: The edge weights of graph for the Ncut encode the similarity between nodes. Given two connected superpixels R_i and R_j in the graph, we derive the entries of matrix $\mathbf{W}$ from the combination of color and intervening image edge cues between superpixels. Firstly, we define a joint metric that measures the distinction between the superpixels as below:

$$d_{ij}^{app+edge} = (1-\alpha) d_{ij}^{app} + \alpha d_{ij}^{edge} \tag{3}$$

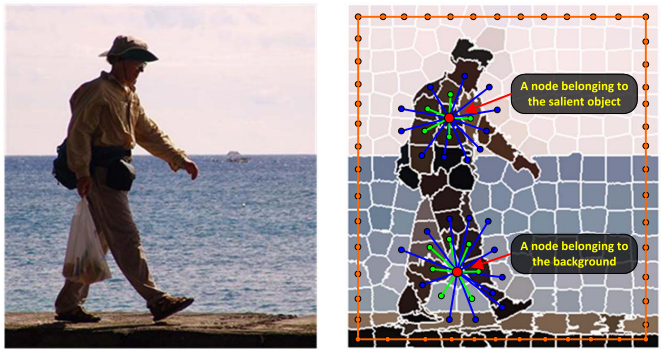


Fig. 3. Connection of graph edge E. Left: an input image. Right: edge connection between superpixels. A vertex (specified by a red dot) connects to both its adjacent superpixels (green connection) and superpixels sharing common boundaries with its adjacent superpixels (blue connection), resulting in a 2-ring graph topology. Connection between arbitrary boundary superpixels (brown dots) is constructed at the same time.

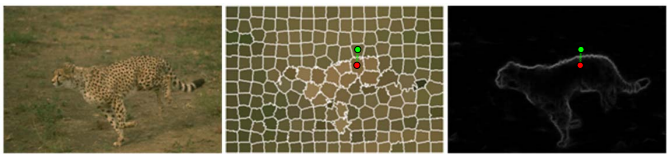


Fig. 4. From left to right: (a) an input image, (b) superpixel representation, and (c) edge detection by [39]. In this case the edge magnitude between two specified superpixels are more significant than superpixel color difference.

where d_{ij}^{app} is the appearance difference and d_{ij}^{edge} is the intervening contour magnitude. The appearance difference d_{ij}^{app} is defined as:

$$d_{ij}^{app} = ||\mathbf{c}_i^0 - \mathbf{c}_j^0|| \tag{4}$$

CIELab color difference is widely used for graph construction in graph-based salient region detection [21], [35] and shown to be effective. The intervening contour magnitude d_{ij}^{edge} is defined as:

$$d_{ij}^{edge} = \max_{\mathbf{p} \in l(\mathbf{p}_i^0, \mathbf{p}_j^0)} \mathbb{E}(\mathbf{p}) \tag{5}$$

where $l(\mathbf{p}_i^0, \mathbf{p}_j^0)$ is the straight line connecting the locations of superpixels R_i^0 and R_j^0, $\mathbf{p}$ runs over every pixel location on the line, and $\mathbb{E}(\mathbf{p})$ is the corresponding edge probability on an edge map $\mathbb{E}$. (5) is known as *intervening contour cue* [38]. In certain cases, using an edge map provides better delineation between objects and background than (4). For example, despite very weak difference in superpixel colors in Fig. 4, the edge detection well highlights the entire plot

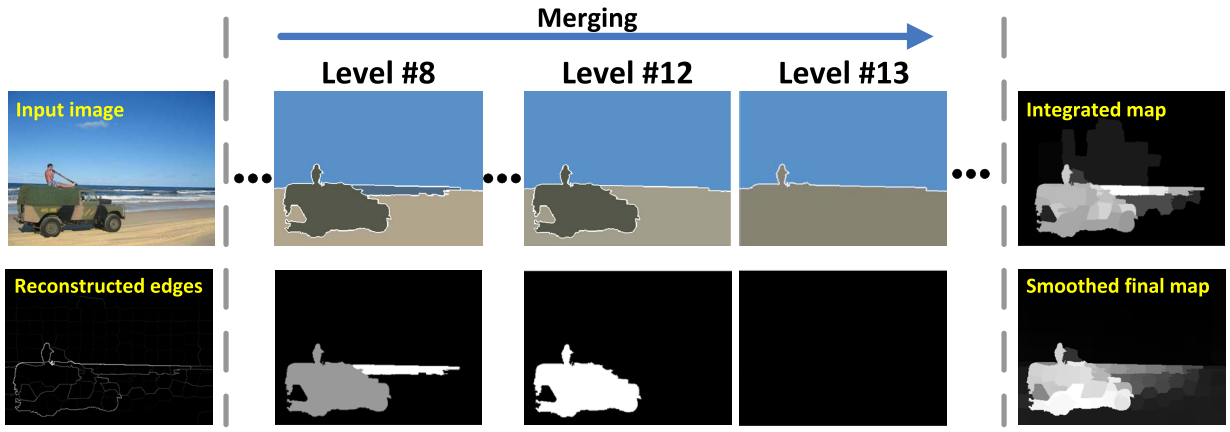


Fig. 5. An illustrative example for region merging. The first column shows the original image and the reconstructed graph edges, based on which the merging proceeds. A high intensity indicates a high edge weight between two superpixels. Only weights among adjacent superpixels are shown. Three columns in the middle show merged regions at level #8, #12 and #13 together with the intermediate saliency maps. A total of 16 levels is obtained in this example.

of the cheetah due to texture distinction. It is worth noting that any edge detector that outputs boundary probability map can be employed. We use the structured random forest edge detector [39] that works on multi-scales and has state-of-the-art performance with reasonably fast processing speed.

As mentioned before, boundary superpixels are all connected with each other. One can observe some disadvantages when computing (5) between boundary superpixels. Since the spatial distance between two boundary superpixels could be large, e.g. one superpixel is on the left border while the other is on the right border of an image, d_{ij}^{edge} can be extremely large as there often exists strong edges on the straight line connecting them. This can prevent clustering between potential background. Therefore, for d_{ij}^{edge} among all boundary superpixels, $d_{ij}^{edge} = d_{ij}^{app}$ is set to alleviate such problem. Noting before integration in (3), d_{ij}^{edge} and d_{ij}^{app} are normalized to [0, 1] by dividing their global maximum respectively.

In (3), α specifies the relative importance of the edge detector. By incorporating edge detection term, better object-background delineation is observed, leading to better Ncut results. This incorporation is optional and can be easily disabled by letting $\alpha = 0$.

Finally, the entry of matrix $\mathbf{W}$ is defined by the joint metric in (3) as:

$$w_{ij} = \begin{cases} \exp(-\dfrac{d_{ij}^{app+edge}}{\sigma^2}) & \text{if } R_i^0 \text{ and } R_j^0 \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where σ^2 is empirically set as 0.1 for all experiments. Diagonal entries of $\mathbf{W}$ are set to zeros to prevent self-loops in the graph.

B. Apply the Ncut to Obtain Cluster Information

We then solve (2) for the generalized eigenvectors $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{nvec}$ (correspond to $nvec + 1$ smallest eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{nvec}$). The resulting eigenvectors of Ncut are the soft indicator vectors of different clusters [37]. Eigenvectors themselves do not contain any saliency information but only cluster information. Also eigenvectors themselves cannot be directly added because they may not correspond to each other.

Since saliency detection is often conducted by applying saliency measures to regions, e.g. regional contrast, we further consider turning such cluster information into region so that those measures can be applied easily. The essence is that eigenvectors are soft cluster labels, and each individual eigenvector implies the extent of superpixels belonging to different clusters. Hence, the difference between values of vertices on the eigenvectors can be integrated, indicating "inter-class distance" [40]. We reconstruct the graph edge e_{ij} between the two connected R_i^0 and R_j^0 by integrating the difference between values of vertices on $nvec$ smallest eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{nvec}$:

$$e_{ij} = \sum_{k=1}^{nvec} \frac{1}{\sqrt{\lambda_k}} |\mathbf{v}_k(R_i^0) - \mathbf{v}_k(R_j^0)| \quad (7)$$

where $\mathbf{v}_k(R_i^0)$ indicates the value in eigenvector $\mathbf{v}_k$ corresponding to the superpixel R_i^0. The weighting by $1/\sqrt{\lambda_k}$ is motivated by the physical interpretation of the generalized eigenvalue problem as a mass-spring system [40]. In practice, $nvec = 8$ suffices while further increasing it introduces extra noise due to large approximation errors of eigenvectors [36]. We have also observed in our experiments that the performance of Ncut for clustering visual contents is affected by the topology of the graph. Since long range graph connections facilitate the propagation of local grouping cues across larger image regions, larger graph radius may make clustering better. Similar phenomenon has been observed in [41], where graph affinity is constructed on pixel level. Hence in this paper the 2-ring graph connection is a trade-off between clustering performance and region connectivity.

C. Graph-Based Adaptive Merging of Vertices

We apply a graph-based region merging scheme [19] to adaptively discover the cluster information in (7). Let $R^l = \{R_1^l, R_2^l, \ldots\}$ be a partition of V in the lth level

and $R_k^l \in R^l$ corresponds to its kth part. Since vertices correspond to superpixels, partition of the graph will result in regions. A criterion D is defined to measure the pairwise difference between R_i^l, R_j^l:

$$D_{ij}^l = D(R_i^l, R_j^l) = \text{mean}_{v_k \in R_i^l, v_m \in R_j^l, e_{km} \in E}\{e_{km}\} \quad (8)$$

where "mean" is an averaging operator over graph edges connecting R_i^l and R_j^l. To discover the cluster information, an adaptive threshold Th is defined to control the bandwidth of D_{ij}^l: at level l, we fuse R_i^l, R_j^l into one cluster (i.e. region) if their difference $D_{ij}^l \leq Th$. Indexes of these sets are found by searching the minimum repetitively as below:

$$i^*, j^* = \arg\min_{i,j} D_{ij}^l, \quad \text{s.t.} \quad D_{ij}^l \leq Th \quad (9)$$

Merging in a level stops when (9) results in no solution, i.e., pairwise difference between two arbitrary regions is larger than Th. At the next level $l+1$, Th is increased as:

$$Th \leftarrow Th + T_s \quad (10)$$

where T_s is a step length and the merging continues as above. T_s is automatically computed by $T_s = (e_{max} - e_{min})/n$ in all experiments, where e_{max}, e_{min} are the maximum and minimum of reconstructed graph edges, respectively, and n is defined as the "quantifying number". $n = 30$ is determined empirically (see V-C). The proposed "merging and adapting" procedure continues until all regions in R^l are merged together, i.e., $|R^l| = 1$. We start the merging from initial superpixels $\{R_1^0, R_2^0, \ldots, R_N^0\}$, and Th is initialized to be T_s in the 1st level (level #1). As Th increases and merging proceeds, multi-levels are obtained.

D. Regional Saliency Measures During Merging

Let R_i^l be a region at merging level l. We propose the following regional saliency measures for R_i^l:

1) 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:

$$S_{i,l}^{fg} = \frac{\sum_{j,k|R_j^0 \in R_i^l, R_k^0 \in B} ||\mathbf{c}_j^0 - \mathbf{c}_k^0||}{|R_i^l| \cdot |B|} \quad (11)$$

where B represents a 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 [14], here we only compare a region with a potential background, i.e. boundary superpixels according to the verified boundary hypothesis [21], [35]. This is more efficient to compute for regions in different levels as boundary set B is always fixed. In practice, as $|B|$ remains a constant for all regions, it is omitted in our implementation.

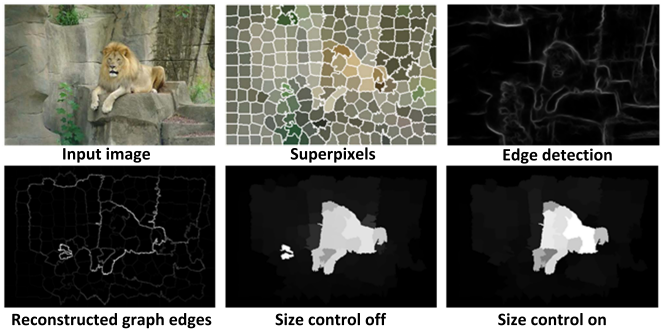


Fig. 6. Illustration for size control. Though a small area of "leaves" is grouped by Ncut, it is suppressed in the integrated saliency map ($N_{low} = 4$). No smoothing is performed at this point.

2) Center Bias: Statistical results in [26] and [42] show that human attention is center biased, indicating that distinctive regions close to image center are likely to be salient [12], [16], [20]. Therefore, the mask with a Gaussian distribution $\mathbb{G}(\mathbf{p})$ is applied at the image center, and the average probability value lying in each region is computed:

$$S_{i,l}^{cb} = \frac{\sum_{j|R_j^0 \in R_i^l} \mathbb{G}(\mathbf{p}_j^0)}{|R_i^l|} \quad (12)$$

where $\mathbb{G}(\mathbf{p}_j^0)$ corresponds to Gaussian value of location $\mathbf{p}_j^0$. Although it has been argued in [35] 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.

3) Boundary Cropping: Boundary hypotheses [21], [35] imply that regions touching image borders are likely to be background. This phenomenon can be explained by the "surroundedness" in Gestalt laws [43], [44]: 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:

$$S_{i,l}^{bc} = \begin{cases} 1 & \text{if } \ell_i^l \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where ℓ_i^l is the number of image borders that R_i^l touches. (13) 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.

4) 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] and [21], we chose multiplication as good background suppression is observed. Furthermore, $S_{i,l}^{bc}$ can effectively suppress image boundary-touching regions if the multiplication is used. Hence, the final saliency score for the region R_i^l is defined as:

$$S_{i,l}^{final} = S_{i,l}^{fg} \cdot S_{i,l}^{cb} \cdot S_{i,l}^{bc} \quad (14)$$

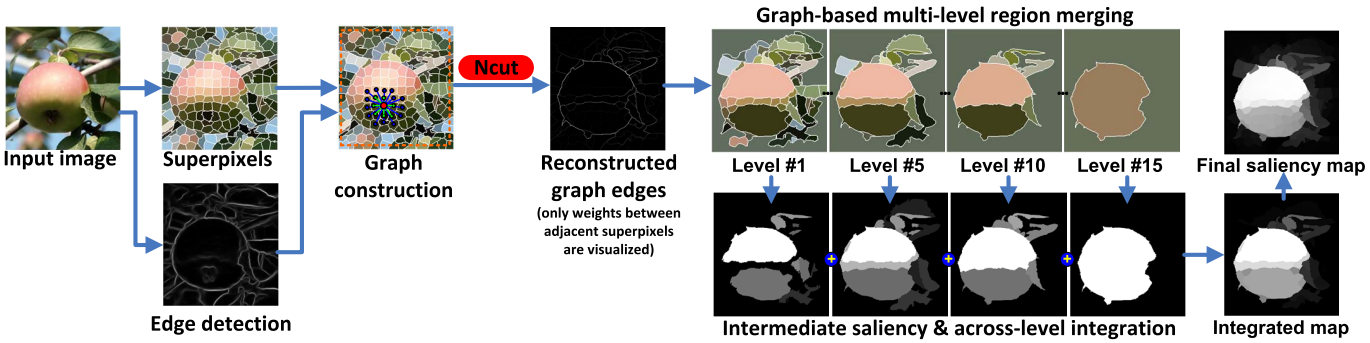


Fig. 7. The block diagram of proposed system.

where $S_{i,l}^{fg}$, $S_{i,l}^{cb}$, $S_{i,l}^{bc}$ 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 (Fig. 5 and Fig. 7).

E. Final Saliency Map Formulation

The final saliency map is formulated by linearly integrating intermediate saliency maps from all levels followed by graph-based manifold ranking [21] for smoothing:

$$\mathbf{f} = (\mathbf{D} - \beta\mathbf{W})^{-1}\mathbf{s} \tag{15}$$

where $\mathbf{W}$ is defined in (6), $\mathbf{D}$ is the degree matrix of $\mathbf{W}$ (see III), $\mathbf{s}$ and $\mathbf{f}$ are respectively saliency values of superpixels (in vector form) before and after the smoothing. β is set to 0.99 according to [21]. Finally, $\mathbf{f}$ is normalized to [0, 1]. Its components are further assigned to corresponding pixels for a final saliency map.

The justification for the cross-level integration is: as the merging proceeds, the cluster information in Ncut vectors is gradually discovered and is turned into regions, yielding to more accurate saliency estimation guided by the Ncut. Although one can choose one or several high levels to perform saliency estimation, the underlying challenge is to decide till which level a salient object would survive. An illustrative example is shown in Fig. 5 where the horizontal sea-sky line is very salient and only two regions (sky and beach) appear in top levels. The salient object (truck) is merged since a middle level. No regions are deemed as salient since level #13 due to (13). In contrary, robust performance for generic situations is achieved by integrating all levels.

Size Control for Salient Objects: In most cases of salient object detection, users wish to detect relatively large visual objects. Though Ncut favors cut of large regions, it is not a hard constraint on region sizes. To further eliminate impact of small-scale patterns, we consider to limit the size of a region. A straightforward way to consider this is by the number of superpixels in each region. Other metric for measuring region size can be considered as well, e.g. the one in [12]. We "inpaint" a region in an intermediate saliency map if it contains superpixels fewer than a pre-defined number N_{low}. This is done for a region by replacing its saliency score with the closest saliency of its neighbor regions. Fig. 6 shows an example where the size control is used to eliminate small-scale noise.

The block diagram of our complete scheme is shown Fig. 7.

V. Experiments and Results

In this section, we compare our scheme, Ncut saliency (NCS), with several state-of-the-art methods on four commonly used datasets.

Datasets and State-of-the-Art Methods: Four benchmark datasets for evaluation include commonly used MSRA-1000 [28] (1000 images), SOD [45] (300 images), SED [46] that consists of two parts, i.e. SED1 (one object set) and SED2 (two objects set) each containing 100 images, and CSSD [12] (200 images with texture background). We compare the proposed method (NCS) with 13 state-of-the-art salient region detection methods: CA (Context Aware) [7], FT (Frequency Tuned) [28], LC (Luminance Contrast) [27], HC (Histogram Contrast) [14], RC (Region Contrast) [14], SF (Saliency Filter) [15], LR (Low Rank) [20], GS (Geodesic Saliency) [35], HS (Hierarchical Saliency) [12], PCA (Principal Components Analysis) [33], DRFI (Discriminative Regional Feature Integration) [13], GC (Global Cue) [31], MR (Manifold Ranking) [21]. We have not compared with eye fixation models such as Itti's [9] and Hou's [1] due to different purposes of the methods.

Criteria: Precision, recall, F-measure (F_β) [15], [21], [28], [31], and mean absolute error (MAE) [15], [31] are used for the evaluation. Definition for these criteria are as follows:

$$Precision(T) = \frac{|M(T) \cap G|}{|M(T)|}, \quad Recall(T) = \frac{|M(T) \cap G|}{|G|} \tag{16}$$

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{17}$$

$$MAE = \frac{1}{|S_{map}|} \sum_{x \in S_{map}} |S_{map}(x) - G(x)| \tag{18}$$

In the above equations, $M(T)$ is the binary mask obtained by directly thresholding a saliency map S_{map} using threshold T, G is the ground truth map, $|\cdot|$ in (16) denotes the sum area of masks, β^2 is set as 0.3 as suggested in previous work to emphasize precision, x is a pixel location with a saliency value $S_{map}(x)$, and $|S_{map}|$ is the total size (width by height) of the map. The reason of using MAE as a compensation

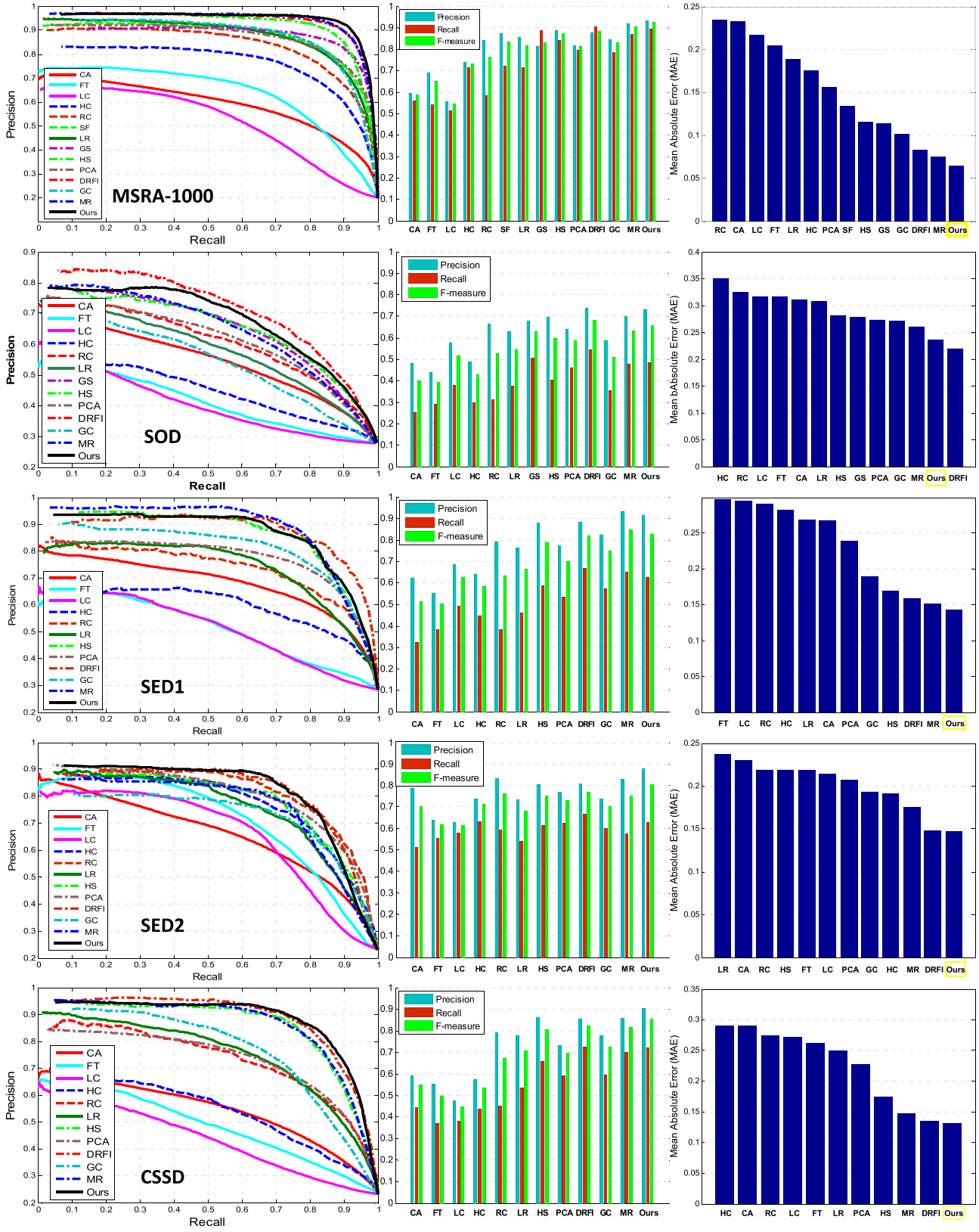


Fig. 8. Quantitative evaluations by precision-recall curves (left column), adaptive threshold (middle column) and mean absolute error (MAE) (right column) on four benchmark datasets: from top to bottom are MSRA-1000, SOD, SED (includes SED1 and SED2), and CSSD. Note because SF only provides results on MSRA-1000 while GS only provides results on MSRA-1000 and SOD, we did not compare with them on the rest datasets.

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

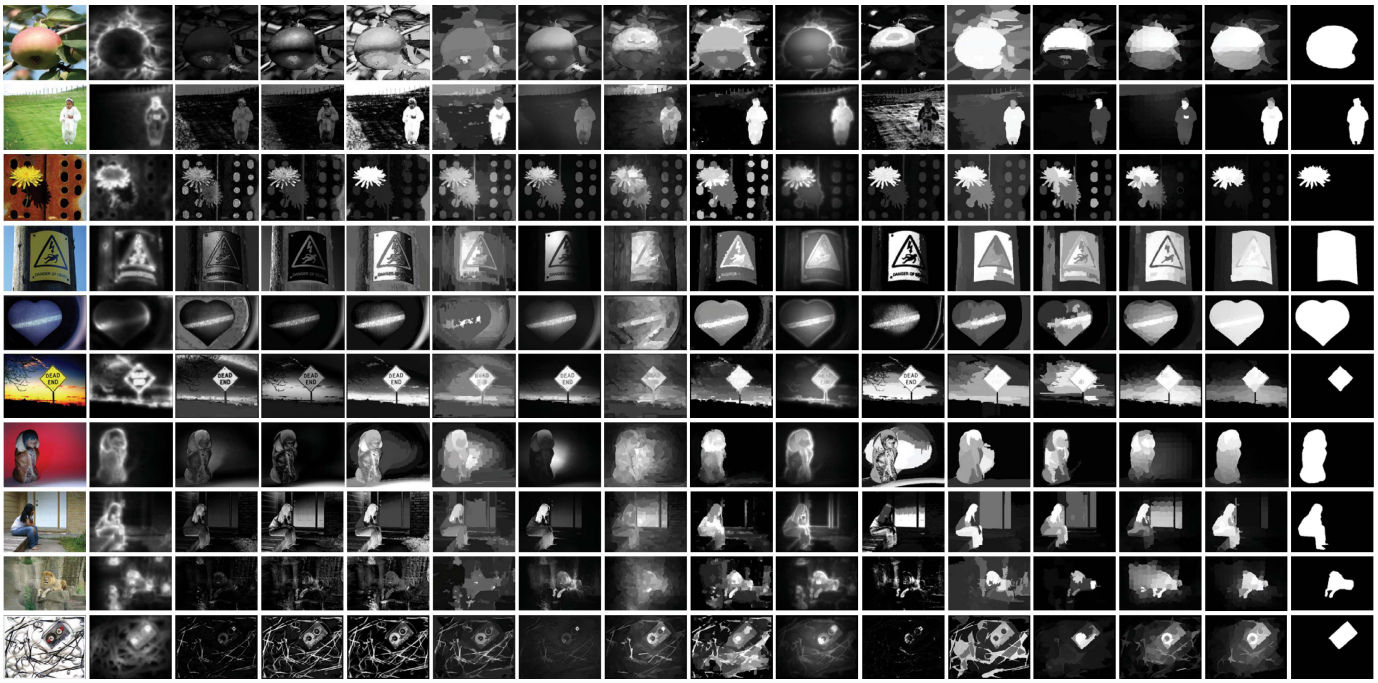


Fig. 9. Visual comparisons on MSRA-1000. The abbreviations have been listed at the beginning of Section V.

accumulation since it sums all pixel-wise errors. Under MAE, saliency maps after smoothing can obtain higher MAE values than those before smoothing (Fig. 12). In contrast, precision-recall curves are insensitive to this problem. A good saliency detection method should achieve high precision-recall curves meanwhile maintain low MAE.

Experiment Setup: The implementation of the full version of our method is: 2-ring graph topology, with edge detection ($\alpha = 0.5$), Ncut, size control ($N_{low} = 4$), and smoothing. Quantifying number n for adaptive region merging is 30, and σ^2 of the graph affinity in (6) is fixed as 0.1 for all experiments. All parameters are determined empirically and are not carefully tuned, though they can be optimized over moderate training images. The above configuration achieves good performance in our experiments.

A. Comparisons With the State-of-the-Art Methods

We compare the full version implementation of the proposed method to 13 existing methods on four benchmark datasets. Noting that size control for SED2 dataset is turned off as images in SED2 usually contain one large and one small objects. Small objects sometimes are as small as two or three superpixels. Abbreviations of all competitors have been listed in the beginning of this section. Precision-recall curves generated by using fixed threshold T from 0 to 1 are shown in Fig. 8. The performance of our method is comparable to the most recent techniques including HS, DRFI and MR. Our method significantly outperforms HS on MSRA-1000, SOD, SED2, and CSSD. Marginal improvement is observed on SED1. Besides, observing precision-recall curves, our method is comparable to DRFI [13] and MR [21] on all the four datasets.

Adaptive threshold experiments were carried out, where the adaptive threshold is defined as two times the mean value of a saliency map [14], [28]. Results are shown in the middle column in Fig. 8. Our method achieves both the highest precision and F-measure on MSRA-1000, SED2, CSSD 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 SED1. For SED1 whose images contain single objects in more complex scenarios, our method performs close to MR [21]. An observation on SED2 is since this dataset has many labeled objects which violate the boundary prior (e.g. 6th row in Fig. 11), MR performs less well than usual. In such cases, it may be better to keep a vague detection using only contrast rather than explicit boundary prior. This reveals why RC and HC that perform less well on other datasets, achieve relatively good performance on SED2.

To further evaluate the methods, we compute the MAE values [15]. As shown in Fig. 8, our method produces consistently the lowest error on MSRA-1000, SED and CSSD datasets, indicating more robustness against different datasets. Despite good performance in precision-recall curves and F-measure, RC [14], HC [14], FT [28] and LR [20] have the highest MAE due to the weak background suppression.

Fig. 9-Fig. 11 show visual comparisons on four datasets. Our method effectively suppresses background clutter and uniformly emphasizes the foreground objects. Better delineation between the object and background can be observed in complex scenes, such as the last row in Fig. 10. In most visual comparisons, much clearer object boundaries are obtained compared to other methods, e.g. 9th row in Fig. 9, 1st, 4th, 5th rows in Fig. 10, and 3rd, 7th, 8th rows in Fig. 11. In addition, the proposed method is able to deal with images

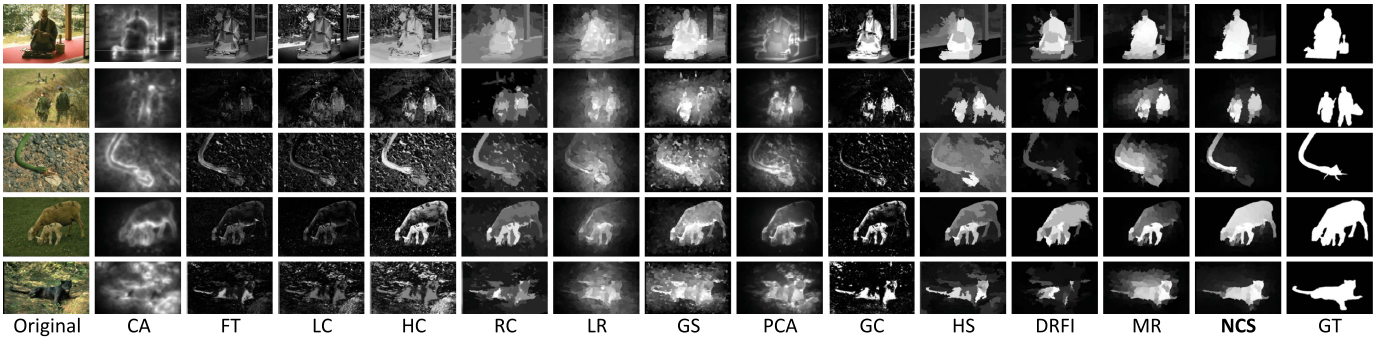


Fig. 10. Visual comparisons on SOD. The abbreviations have been listed at the beginning of Section V.

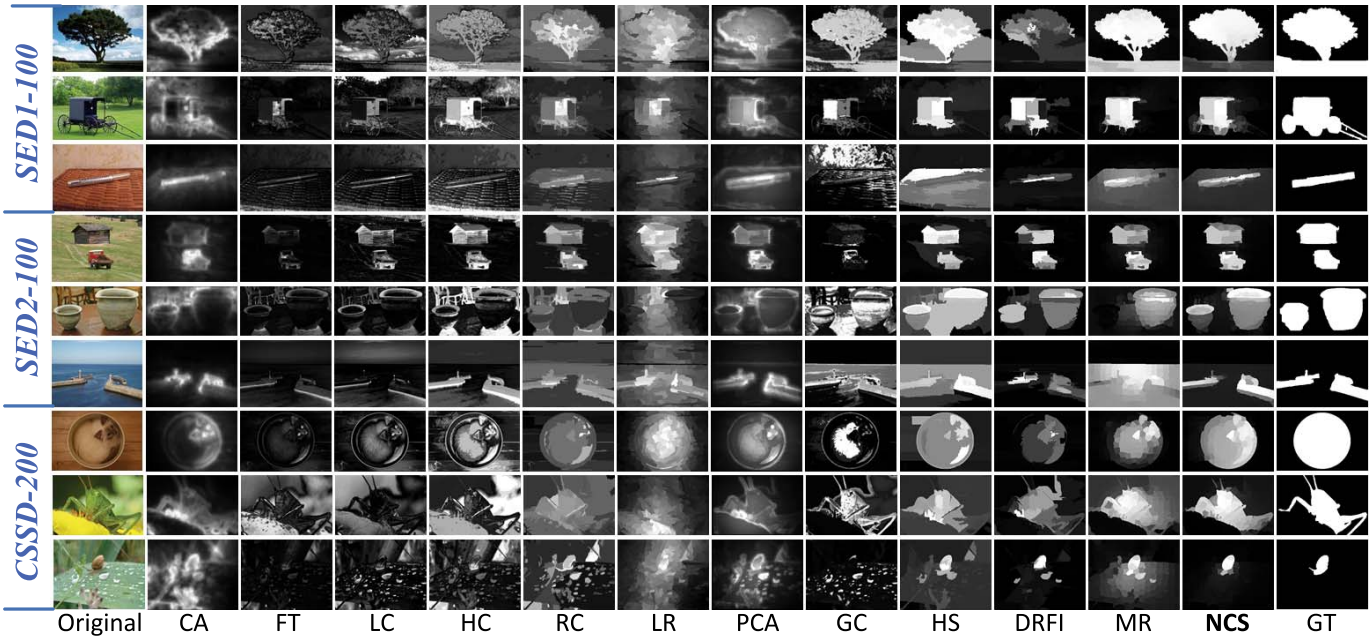


Fig. 11. Visual comparisons on SED1, SED2 and CSSD respectively. The abbreviations have been listed at the beginning of Section V.

containing "color ramps". Such effects are usually caused by shadow or lighting conditions (4th-7th rows in Fig. 9). Our region merging scheme effectively combines them into background, preserving perceptual homogeneity. In contrast, the contrast-based GC [31], SF [15] and geodesic based GS [35] methods that use over-partitioned image segments are less better due to color heterogeneity. Our method also handles challenging cases that the state-of-the-art methods fail. For example, a key procedure in MR [21] is intermediate thresholding and re-propagation to refine the results (called "second stage" in [21]). The operation is critical in achieving high performance. Since this operation depends highly on the threshold, once isolated cluttered regions are segmented, they are difficult to be absorbed into the background even with the help of re-propagation, e.g., the shadow of flower in 3rd row of Fig. 9. Another side-effect of this operation is the risk of missing useful object parts. This is consistently observed on SED2 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. 5th row in Fig. 11). In contrast, such risk is avoided in our method as no threshold is used to binarize the saliency map for performance boosting. By removing small size distracters, our method achieves good performance on cluttered background when other methods are less satisfactory, e.g. Fig. 11 last row.

B. Validation of Individual Modules

Alternatives of the proposed method include: smoothing on/off, size control on/off, edge detection on/off (by changing α). Our scheme can also work without the Ncut by using (3) to replace (7). The reason is that the graph-based merging which examines the average graph edges using (8) may discover silhouettes of objects [25]. Besides, one may also turn off the region merging and just compute saliency measures on initial superpixel regions. In this subsection, we show the test results on the effectiveness of each individual component by gradually removing these modules from the full version of our method. Quantitative results for this experiment on MSRA-1000 are shown in Fig. 12. Considering both gains

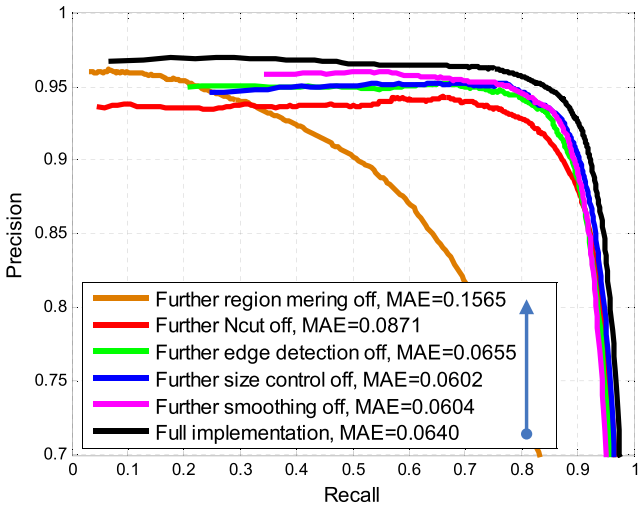


Fig. 12. Evaluations on alternatives of NCS including smoothing off, size control off, edge detection off ($\alpha = 0$), Ncut off, region merging off. Note curves are shown in the precision range between 0.7 and 1.

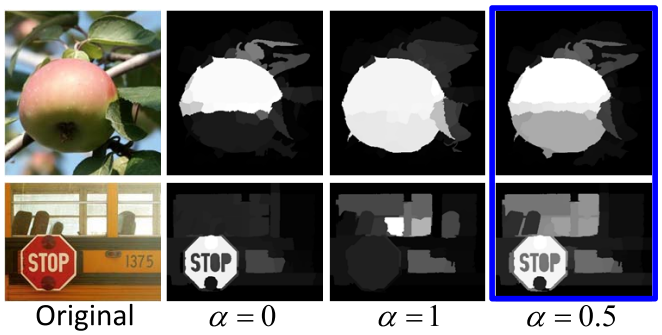


Fig. 13. Robust performance is obtained by combining appearance difference and edge detection, rather than using merely either. No smoothing at this point.

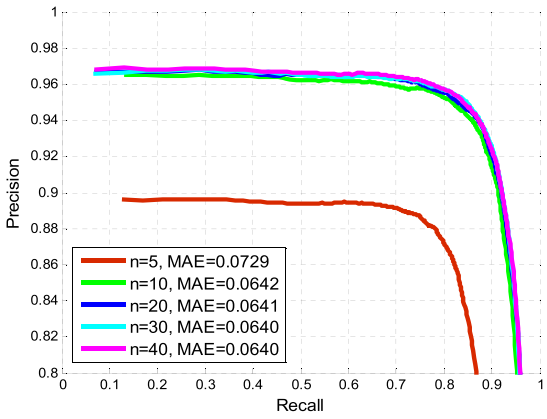


Fig. 14. Performance under varied quantifying number n. Note curves are shown in the precision range between 0.8 and 1.

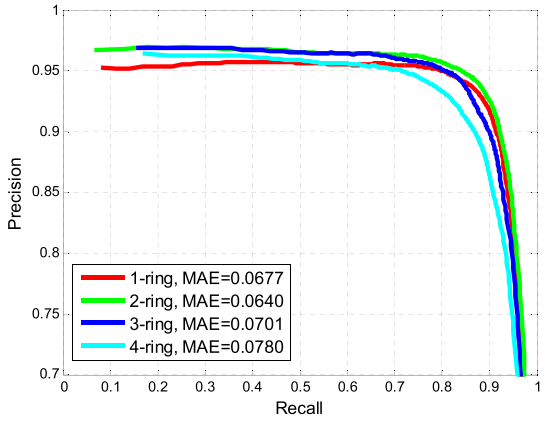


Fig. 15. Performance by changing the range of graph connection. Note curves are shown in the precision range between 0.7 and 1.

in precision-recall curves and MAE criterion, modules which have important contributions to the system performance are the region merging, Ncut, edge detection (Fig. 13), and final smoothing. Performance variation can be clearly observed in Fig. 12. The effect of size control is relatively minor on this dataset but we do observe individual cases where they play crucial roles (e.g. Fig. 6). Combining all these steps makes the system more robust and capable of achieving better performance. Fig. 13 shows the robustness of combining the edge cue with the color cue.

C. Quantitative Evaluation of Region Merging

We also find that our method is robust to the "quantifying number" n in region merging. Quantitative results generated in the experiments is by the full implementation with different n. Fig. 14 shows the results on MSRA-1000 dataset. When using an $n > 30$, our method produces similar results. Very minor effect on the MAE is observed when varying n from 10 to 40. Less good results in $n = 5$ are caused by missing object parts. Due to the large step T_s resulted from a small n, some desired salient regions can be directly merged with background without appearing in any intermediate level. Contrarily, large n results in more levels to compute. We set $n = 30$ as default for a good precision-recall curve although precision can be further sacrificed for the speed.

D. Sensitivity to the Range of Graph Connection

As clustering performance of the Ncut is related to the range of graph connection, we evaluate the detection performance by the full implementation with different ranges of graph connection in this part. We gradually increase the graph connection from 1-ring to 4-rings (Noting that the 1-ring graph is the local neighbor graph in [25]). Quantitative evaluation on MSRA-1000 is shown in Fig. 15. One can observe that 2-ring graph topology achieves the best results. It is better than 1-ring, whereas 1-ring is better than 3-rings and 4-rings. 4-ring case has the worst results. Extending the graph connection leads to more non-zero entries in the graph affinity matrix $\mathbf{W}$. This directly influences the performance of the Ncut. Though a long range connection can effectively propagate grouping cues to further regions, it can also confront inappropriate grouping that groups scattered background noise together or groups parts of object region with far away background regions, e.g. in Fig. 16 the "white cloth" on the girl is grouped with the "white door" in the background by using 4-rings. Hence both too short and too long range connection are inappropriate. In this experiment, we observe 2-ring graph

Fig. 16. Results generated by full implementation with different ranges of graph connection. From left to right: 1-ring, 2-ring, 3-ring, and 4-ring. The original image is shown in Fig. 1.

connection achieves the most optimal performance under the current system settings.

E. Efficiency and Speed

The average time cost for our full implementation on MSRA-1000 is 2.6 seconds for n equal to 30, where about 0.25s is taken by superpixel segmentation and 0.4s by random forest edge detection. Though the adopted regional saliency measures are computationally light, the main computation load lies in the Ncut (eigenvalue and eigenvector solving) and multi-level region merging. The computation time reported is acquired on an Intel i7-4720HQ 2.6GHz laptop with 8GB memory using unoptimized Matlab code. It is slower than Matlab code of MR which takes 0.5s but faster than DRFI which requires 5.5s under the same hardware condition.

F. Discussion on Future Work

Future work can be conducted in several aspects: 1) Graph cut that partitions a graph in a discriminative way, has shown its strong capability for clustering visual contents for saliency detection. As the Ncut is a classical technique, other advanced graph cut techniques which provides superior grouping performance can be considered. 2) Performance of integrating intervening contour cue and appearance difference of superpixels indicates that there is independent information in each cue to exploit for saliency detection. In the future, a learning-based approach can be used to study the optimal combination for constructing the graph affinity. Besides, more feature cues can be extracted from superpixels. 3) In our work, a fixed range of graph connection is empirically determined and used for all cases. In the future, an adaptive metric which determines the optimal connection range based on image properties is interesting to study. 4) Other effective regional saliency measures can be exploited and employed into our system, e.g. geometric information like convexity of a region.

VI. Conclusion

We have presented a new salient object detection framework based on the Normalized graph cut (Ncut) and adaptive multi-level region merging. The former guarantees good grouping of visual contents and the latter provides a feasible way of turning the cluster information into explicit regions, where region-based saliency measures can then be easily applied. When combined, they greatly improve the accuracy on detecting entire objects and effectively suppress the background. We also validated on combining intervening contour cue from edge detection to construct graph affinity. Results show better delineation between object and background, leading to better grouping results. Size control is a parameter in the proposed method that guarantees users' flexibility on deciding their own minimal size for object detection in some applications. Experiments have shown that our method performs well on enhancing objects holistically meanwhile suppressing the background. It achieves state-of-the-art performance on four commonly used benchmark datasets in terms of competent precision, recall and F-measure, meanwhile maintaining the lowest MAE.

References

[1] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[2] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 366–379.

[3] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 105–112.

[4] F. Stentiford, "Attention based auto image cropping," in *Proc. Workshop Comput. Attention Appl. ICVS*, 2007, pp. 1–9.

[5] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 2232–2239.

[6] Y. Ding, X. Jing, and J. Yu, "Importance filtering for image retargeting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 89–96.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2376–2383.

[8] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–10, Dec. 2006.

[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[10] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 18. Jun. 2006, pp. 155–162.

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

[12] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1155–1162.

[13] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2083–2090.

[14] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 409–416.

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

[16] K. Fu, C. Gong, J. Yang, and Y. Zhou, "Salient object detection via color contrast and color distribution," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2012, pp. 111–122.

[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[18] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[19] 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.

[20] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 853–860.

[21] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3166–3173.

[22] P. Mehrani and O. Veksler, "Saliency segmentation based on learning and graph cut refinement," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2010, pp. 1–12.

[23] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2008, pp. 1–4.

[24] H.-Y. Gao and K.-M. Lam, "Segmentation-enhanced saliency detection model based on distance transform and center bias," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 2803–2807.

[25] K. Fu *et al.*, "Adaptive multi-level region merging for salient object detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2014, pp. 1–11.

[26] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 414–429.

[27] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Multimedia*, 2006, pp. 815–824.

[28] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1597–1604.

[29] K. Shi, K. Wang, J. Lu, and L. Lin, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2115–2122.

[30] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[31] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1529–1536.

[32] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3232–3242, Dec. 2010.

[33] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1139–1146.

[34] Y. Xie and H. Lu, "Visual saliency detection based on Bayesian model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 645–648.

[35] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 29–42.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[37] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[38] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2010.

[39] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1841–1848.

[40] S. Belongie and J. Malik, "Finding boundaries in natural images: A new method using point descriptors and area completion," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 1998, pp. 751–766.

[41] T. Cour, T. Benezit, and J. Shi, "Spectral segmentation with multiscale graph decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 1124–1131.

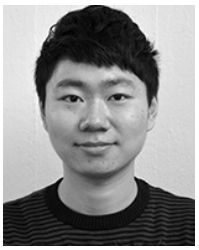
[42] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–4, 2007.

[43] K. Koffka, *Principles of Gestalt Psychology*. London, U.K.: Routledge & Kegan Paul, 1955.

[44] S. E. Palmer, *Vision Science: Photons to Phenomenology*. Cambridge, MA, USA: MIT Press, 1999.

[45] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Comput. Soc. Workshop Perceptual Org. Comput. Vis.*, Jun. 2010, pp. 49–56.

[46] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.



Keren Fu (S'14) received the bachelor's degree from the Huazhong University of Science and Technology, China, in 2011, and the Licentiate degree in engineering from the Chalmers University of Technology, Sweden, in 2014. He is currently pursuing the Ph.D. degrees with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and the Department of Signals and Systems, Chalmers University of Technology, Sweden, under the supervision of Prof. J. Yang and Prof. I. Y.-H. Gu. His research areas are computer vision and image/video modeling, with applications to, e.g., visual saliency detection, object tracking and detection, traffic analysis, and machine learning.

Chen Gong received the bachelor's degree from the East China University of Science and Technology, in 2010. He is currently pursuing the Ph.D. degrees with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, under the supervision of Prof. J. Yang and Prof. D. Tao. He has authored 21 technical papers at prominent journals and conferences, such as the IEEE T-NNLS, the IEEE T-CYB, CVPR, AAAI, and ICME. His research interests mainly include machine learning, data mining, and learning-based vision problems.

Irene Yu-Hua Gu (M'94–SM'03) 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 a Research Fellow with the Philips Research Institute IPO, Eindhoven, a Post-Doctoral Fellow with Staffordshire University, Staffordshire, U.K., and a Lecturer with the University of Birmingham, Birmingham, U.K. Since 1996, she has been with the Department of Signals and Systems, Chalmers University of Technology, Göteborg, Sweden, where she is currently a Full Professor. Her research interests include statistical image and video processing, object tracking and video surveillance, pattern classification, and signal processing with applications to electric power systems. She was an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS from 2000 to 2005. She was the Chair-Elect of the IEEE Swedish Signal Processing Chapter from 2002 to 2004. She has been an Associate Editor of the *EURASIP Journal on Advances in Signal Processing* since 2005, and on the Editorial Board of the *Journal of Ambient Intelligence and Smart Environments* since 2011.

Jie Yang received the Ph.D. degree from the Department of Computer Science, Hamburg University, Germany, in 1994. He is currently a Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China. He has led many research projects (e.g., the 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.