

Unsupervised Salient Object Segmentation Based on Kernel Density Estimation and Two-Phase Graph Cut

Zhi Liu, *Member, IEEE*, Ran Shi, Liquan Shen, Yinzhu Xue, King Ngi Ngan, *Fellow, IEEE*, and Zhaoyang Zhang

Abstract—In this paper, we propose an unsupervised salient object segmentation approach based on kernel density estimation (KDE) and two-phase graph cut. A set of KDE models are first constructed based on the pre-segmentation result of the input image, and then for each pixel, a set of likelihoods to fit all KDE models are calculated accordingly. The color saliency and spatial saliency of each KDE model are then evaluated based on its color distinctiveness and spatial distribution, and the pixel-wise saliency map is generated by integrating likelihood measures of pixels and saliency measures of KDE models. In the first phase of salient object segmentation, the saliency map based graph cut is exploited to obtain an initial segmentation result. In the second phase, the segmentation is further refined based on an iterative seed adjustment method, which efficiently utilizes the information of minimum cut generated using the KDE model based graph cut, and exploits a balancing weight update scheme for convergence of segmentation refinement. Experimental results on a dataset containing 1000 test images with ground truths demonstrate the better segmentation performance of our approach.

Index Terms—Color saliency, graph cut, kernel density estimation, saliency model, salient object segmentation, seed adjustment, spatial saliency.

I. INTRODUCTION

SALIENT object segmentation plays an important role in a variety of applications including content-based image retrieval [1], object-based image/video adaptation [2], [3], scene understanding [4], etc. A human observer can effortlessly identify salient objects even in a complex natural scene, but unsupervised segmentation of salient objects from images is nontrivial for a computer. In the last decade, many approaches have been

proposed for salient object segmentation, but it still remains a challenging problem up to now.

Salient objects in natural scenes generally stand out relative to its surrounding regions in terms of some features, and draw attention from a human observer. Therefore, the mechanism of human visual attention is useful for devising a feasible approach for unsupervised salient object segmentation. In practice, most salient object segmentation approaches exploit the so-called saliency map, which is generated using a saliency model, to provide the position and scale information of salient object as the useful segmentation cues. The quality of saliency map is a key factor that affects the reliability of salient object segmentation. In the following, we will first briefly introduce some related saliency models used for salient object segmentation (a recent comprehensive survey on saliency models for a wide range of applications can be found in [5]), and then review salient object segmentation approaches using different schemes.

Based on a biologically-plausible visual attention architecture [6] and feature integration theory [7], Itti *et al.* proposed a well-known saliency model [8], which computes feature maps of luminance, color and orientation using a center-surround operator across different scales, and performs normalization and summation to generate the saliency map. Salient regions showing high local contrast with their surrounding regions can be highlighted in the saliency map. Inspired by the centre-surround scheme used in Itti's saliency model, image saliency is measured using more features such as local contrast of color, texture and shape feature [9], multi-scale contrast [10], ordinal signatures of edge and color orientation histograms [11], oriented subband decomposition-based energy [12], and local regression kernel-based self-resemblance [13]. The key factor for realizing the center-surround scheme is the selection of surrounding region, which is selected as the whole image region in the frequency-tuned saliency model [14], and the maximum symmetric region in [15]. In [16], based on a region segmentation result, the center-surround differences on five features including color contrast, size, symmetry, orientation and eccentricity of regions are fully exploited to generate a region-level saliency map.

Except for the aforementioned center-surround scheme, there are various formulations for measuring saliency. In the frequency domain, both the spectral residual of Fourier transform [17] and the phase spectrum of quaternion Fourier transform [18] are exploited to evaluate the saliency at block level. Based on information theory, the rarity represented using self-information of local image features [19], and the average transferring information represented using entropy rate [20] are exploited to measure saliency. Conditional random field (CRF) learning is exploited in [21] to integrate a set of feature

Manuscript received June 17, 2011; revised October 25, 2011 and January 15, 2012; accepted February 21, 2012. Date of publication March 08, 2012; date of current version July 13, 2012. This work was supported by National Natural Science Foundation of China under Grant No. 61171144 and No. 60602012, Shanghai Natural Science Foundation (No. 11ZR1413000), Innovation Program of Shanghai Municipal Education Commission (No. 12ZZ086), and the Key (Key grant) Project of Chinese Ministry of Education (No. 212053). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Charles D. (Chuck) Creusere.

Z. Liu, L. Shen, and Z. Zhang are with the School of Communication and Information Engineering, Shanghai University, Shanghai, China, and also with the Key Laboratory of Advanced Display and System Application (Shanghai University), Ministry of Education, Shanghai, China (e-mail: liuzhisjtu@163.com; jsslq@163.com; zhyzhang@staff.shu.edu.cn).

R. Shi and Y. Xue are with the School of Communication and Information Engineering, Shanghai University, Shanghai, China (e-mail: dnasr@sohu.com; mhtymhty@163.com).

K. N. Ngan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, ShaTin, N. T. Hong Kong, China (e-mail: knngan@ee.cuhk.edu.hk).

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/TMM.2012.2190385

maps including multi-scale contrast, center-surround histogram and color spatial distribution into the saliency map. Recently, different statistical models and the global information of the image are efficiently utilized to improve the quality of the saliency map. In [22], the global color distribution represented using Gaussian mixture models (GMM), and both local and global orientation distribution are fully utilized to selectively generate the saliency map. In [23], the kernel density estimation (KDE)-based nonparametric model is constructed for each segmented region, and color and spatial saliency measures of KDE models are evaluated and exploited to measure the pixel's saliency. In [24], the histogram-based global contrast and the spatially weighted regional contrast are exploited to generate the saliency map at pixel-level and region-level, respectively.

It should be noted that some of the aforementioned saliency models generate spotlight saliency maps [8]–[13], [17]–[20], which generally can only highlight the center portion and/or the high-contrast boundaries of salient objects, but cannot suppress the high-contrast background regions. The other saliency models [14]–[16], [21]–[24] can highlight the salient regions more completely and suppress the background regions more sufficiently, and generally improve the quality of the generated saliency maps. Undoubtedly, the latter class of saliency maps is more suitable for salient object segmentation. If simple thresholding operations are performed on the saliency maps, the obtained salient objects using the latter class of saliency maps are generally better. In [25] and [26], convex hull analysis is performed on several binary object masks, which are generated by thresholding different feature-based saliency maps, to select the one with the most compact shape to represent the salient object. However, the thresholding operation is only sufficient for those clear saliency maps, in which the complete salient object is highlighted with well-defined boundaries and background regions are totally suppressed, to accurately extract the salient object. Therefore, more elaborate salient object segmentation approaches are needed for improving the segmentation quality and enhancing the applicability on various images.

Region segmentation can be used as a post-processing step to improve the accuracy of the segmented salient object boundaries. In [10], [12], and [14], region saliency is computed as the average saliency of all pixels in each segmented region, and is exploited to select some high-saliency regions to constitute the salient object. On the other hand, region segmentation can also be used as a pre-processing step for salient object segmentation. In [1], the contrasts of color and texture features are exploited to evaluate the saliency measures of segmented regions, and region combinations are iteratively popped out as salient objects by maximizing a global saliency index. However, the quality of salient object segmented using these approaches is highly dependent on the region segmentation result, and is severely degraded due to the problem of under-segmentation or over-segmentation.

Diverse methods from statistics, pattern recognition, and graph theory have been introduced into different salient object segmentation approaches. In [27], the attention GMM for salient object and background GMM are constructed on the image clustering result, and pixels are classified under the Bayesian framework to obtain the salient object. In [28], the

saliency map generated using Itti's model is exploited to select seed pixels for salient objects, and a Markov random field that integrates the features of color, texture and edge is utilized to grow salient object regions. In [29], a support vector machine is trained to select regions for clustering into the salient object. In [30], random walks on the weighted graph are exploited to select salient nodes and background nodes, and semi-supervised learning is further used to determine the labels of unlabelled nodes. However, its main limitation is that the generated binary mask of salient object only has block-level accuracy.

Generally, any problem of object segmentation can be formulated as a pixel-level binary labeling problem, which can be solved under the framework of graph cut [31]. In the context of salient object segmentation, the key issue is how to use the information of saliency map for graph cut. In [32] and [33], by performing binarization on the saliency map using the manually set threshold, the seeds for salient object/background are selected inside/outside a region with a pre-defined distance to the image center, and are exploited to define the data term for the graph. Differently, for constructing the graph in [15], the saliency map generated based on the maximum symmetric surrounding region is directly exploited to define the data term, and the smoothness term is defined to promote the label coherence among neighboring pixels with similar colors. However, as stated in [15], the quality of the segmented salient object strongly depends on the quality of the saliency map. Therefore, it is not desirable to obtain an acceptable quality of salient object segmentation if the salient object is not sufficiently highlighted or the background is not effectively suppressed in the saliency map. In [34], the saliency map is generated using the statistical formulation on the feature distribution contrast between the center and surrounding window. For constructing the graph, both saliency map and color similarity are used to define the two complementary data terms, and CRF learning is exploited to determine the weights for the two data terms and the smoothness term. However, the pre-determined scales for surrounding window, the manually set prior probability in statistical formulation, and the weights pre-determined using CRF learning may be not appropriate for some complicated images to achieve an acceptable segmentation quality.

Although various approaches mentioned above have been proposed for salient object segmentation, the segmentation quality achieved on complicated images, including cluttered background, highly textured regions, and low contrast between object and background, is severely degraded in most cases. In order to enhance the segmentation reliability especially for complicated images and improve the overall segmentation quality, we propose an efficient salient object segmentation approach using a KDE-based saliency model and a two-phase graph cut framework. Our approach, which is extended from our previous work [23], provides more appropriate saliency maps for salient object segmentation, and achieves a higher segmentation quality for a wide range of images using the proposed two-phase graph cut framework. Compared with previous salient object segmentation approaches, the main contributions of our approach are threefold. First, we propose to evaluate color saliency and spatial saliency of a set of KDE models, which are constructed based on the pre-segmentation result of the input image, and then generate the pixel-wise

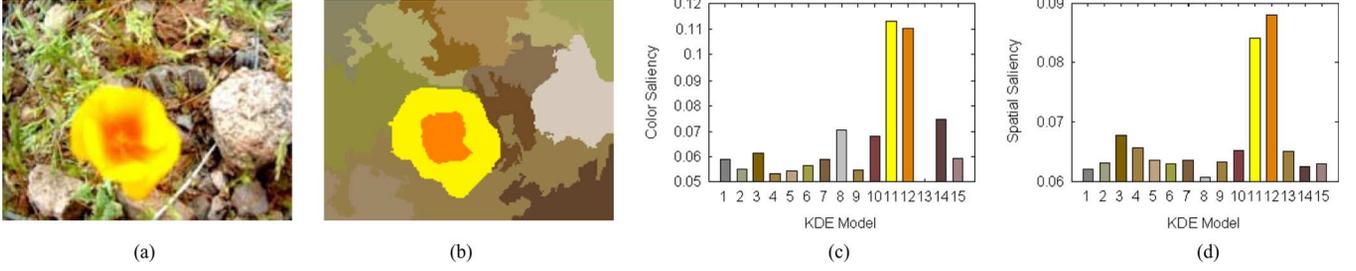


Fig. 1. Illustration of KDE modeling and saliency evaluation. (a) Original image. (b) Pre-segmentation result. (c) Normalized color saliency values of KDE models. (d) Normalized spatial saliency values of KDE models.

saliency map based on saliency measures of KDE models. Second, we propose a two-phase graph cut framework, in which the saliency map based graph cut in the first phase is generally sufficient to obtain a visually acceptable salient object segmentation result for some saliency maps, and the segmentation refinement based on the iterative seed adjustment in the second phase is exploited to further refine the unsatisfactory initial segmentation result. Third, we efficiently utilize the information of minimum cut obtained using the KDE model based graph cut for reasonable adjustments of object/background seeds, and ensure the convergence of segmentation refinements with the introduction of the balancing weight update scheme for graph cut. Experimental results show that our approach achieves considerable improvements on segmentation quality compared to two state-of-the-art salient object segmentation approaches [15], [34].

The rest of this paper is organized as follows. Section II describes the KDE-based saliency model, and Section III details the two-phase graph cut framework for salient object segmentation. Extensive experimental results are presented in Section IV, and conclusions are given in Section V.

II. KDE-BASED SALIENCY MODEL

For the purpose of efficient salient object segmentation, we expect to obtain a suitable saliency map that can effectively highlight salient object regions with well-defined boundaries and suppress background regions. In this section, we present a KDE-based saliency model, which first constructs KDE models based on pre-segmentation result, then evaluates the saliency measures of KDE models, and finally generates the pixel-wise saliency maps. The following three subsections will detail the KDE-based saliency model.

A. KDE Modeling Based on Pre-Segmentation

The original color image in the RGB color space is first transformed into the Luv color space, and then partitioned into a set of regions using the mean shift algorithm [35], in which the parameters of spatial bandwidth and range bandwidth are set to their default values. We only adjust the parameter of minimum allowable region area to control the degree between over-segmentation and under-segmentation, and set it to $\tau \cdot w \cdot h$, where w and h denotes the image width and height, respectively. We set τ to 0.03 for the following examples in Sections II and III, and we will show experimental analysis on how the parameter τ affects the quality of saliency maps and salient object segmentation results in Section IV. For the example image in Fig. 1(a), the pre-segmentation result using the mean shift algorithm is

shown in Fig. 1(b), in which each segmented region is represented using its mean color.

The pixels in each segmented region $r_i (i = 1, \dots, n)$ are then used as the samples to construct a KDE-based nonparametric model $\mathbf{K}_i (i = 1, \dots, n)$. As a nonparametric technique, which estimates the density function directly from the sample data without any assumptions about the underlying distribution, KDE can asymptotically converge to any density function [36]. This property makes KDE quite general and applicable to modeling the pixel samples from either homogenous region or textured region segmented using the mean shift algorithm. Specifically, for each pixel p , its likelihood to fit each KDE model \mathbf{K}_i is defined as

$$C_i(p) = \frac{1}{|r_i|} \sum_{q \in r_i} \kappa_i(\mathbf{c}_p - \mathbf{c}_q) \quad (1)$$

where $|r_i|$ denotes the number of pixels in r_i , i.e., the number of samples in \mathbf{K}_i . \mathbf{c}_p denotes the color feature of the pixel p , and \mathbf{c}_q denotes the color feature of any pixel q in r_i . Specifically, Gaussian distribution is selected as the kernel function κ_i for each KDE model \mathbf{K}_i due to its continuity, differentiability and locality properties [36], and is defined as

$$\kappa_i(\mathbf{c}_d) = \frac{1}{(2\pi)^{3/2} |\mathbf{H}_i|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{c}_d^t \mathbf{H}_i^{-1} \mathbf{c}_d\right) \quad (2)$$

where \mathbf{H}_i is the bandwidth matrix, and $\mathbf{c}_d = \mathbf{c}_p - \mathbf{c}_q$ is the color difference vector between \mathbf{c}_p and \mathbf{c}_q .

Since the chrominance channels are decoupled from the luminance channel in the Luv color space, we assume that the bandwidth for each channel has no correlation with the other two channels. Therefore, \mathbf{H}_i is simplified as a 3-D diagonal matrix, and (2) is simplified as

$$\kappa_i(\mathbf{c}_d) = \prod_{k=1}^3 \frac{1}{\sqrt{2\pi} \sigma_{i,k}} \exp\left(-\frac{1}{2} \frac{\mathbf{c}_{d,k}^2}{\sigma_{i,k}^2}\right) \quad (3)$$

where $\sigma_{i,k}$ denotes the bandwidth of the k th channel in \mathbf{H}_i , and $\mathbf{c}_{d,k}$ denotes the k th component of the color difference vector \mathbf{c}_d . The bandwidth for each channel is independently estimated using the fast binned kernel density estimator [37]. It can be seen from (1)–(3) that the likelihood measure $C_i(p)$ is higher when the color differences between the pixel p and the sample pixels in the KDE model \mathbf{K}_i are smaller, and vice versa.

B. Saliency Evaluation of KDE Models

The saliency of each KDE model is then evaluated based on its color distinctiveness and spatial distribution, and the color saliency and spatial saliency of KDE models are calculated in

turn. The color distance vector and the spatial distance vector between each pixel p and each KDE model \mathbf{K}_i are, respectively, defined as

$$\mathbf{d}_i^c(p) = \mathbf{c}_p - \boldsymbol{\mu}_i \quad (4)$$

$$\mathbf{d}_i^s(p) = [x_p - \tilde{x}_i, y_p - \tilde{y}_i]^T \quad (5)$$

where $\boldsymbol{\mu}_i$ is the mean color of sample pixels in \mathbf{K}_i , (x_p, y_p) is the spatial position of the pixel p , and $(\tilde{x}_i, \tilde{y}_i)$ is defined as the weighted spatial center position of the color distribution represented by \mathbf{K}_i

$$\tilde{x}_i = \frac{\sum_{p \in \mathcal{P}} x_p \cdot C_i(p)}{\sum_{p \in \mathcal{P}} C_i(p)}, \quad \tilde{y}_i = \frac{\sum_{p \in \mathcal{P}} y_p \cdot C_i(p)}{\sum_{p \in \mathcal{P}} C_i(p)} \quad (6)$$

where \mathcal{P} denotes the set of all pixels in the image. Using (6), the contribution of the pixels that show similar colors with the sample pixels in \mathbf{K}_i is substantially considered over the whole image.

Since the colors of salient objects are distinctive from background colors in natural images, the pixels belonging to salient objects have larger distances to other pixels in the color domain. Therefore, if the colors covered by a KDE model \mathbf{K}_i are far away from the colors covered by other KDE models in the color domain, the colors covered by \mathbf{K}_i are such distinctive colors. Between any pair of KDE models, \mathbf{K}_i and \mathbf{K}_j , the color distance with symmetrical form is defined in a probabilistic manner as follows:

$$D_c(i, j) = \frac{1}{2} \cdot \left[\frac{\sum_{p \in \mathcal{P}} C_i(p) \cdot \|\mathbf{d}_j^c(p)\|}{\sum_{p \in \mathcal{P}} C_i(p)} + \frac{\sum_{p \in \mathcal{P}} C_j(p) \cdot \|\mathbf{d}_i^c(p)\|}{\sum_{p \in \mathcal{P}} C_j(p)} \right] \quad (7)$$

where the former (resp. latter) term in the square bracket represents the color distance to \mathbf{K}_j (resp. \mathbf{K}_i) normalized over all pixels by considering the pixels' likelihoods to fit \mathbf{K}_i (resp. \mathbf{K}_j). The average of such two normalized color distances is then used to reasonably measure the color distance between \mathbf{K}_i and \mathbf{K}_j .

The color saliency for \mathbf{K}_i is then defined as the sum of weighted color distances between \mathbf{K}_i and all the other KDE models

$$KS_c(i) = \sum_{j=1}^n \alpha_j \cdot D_c(i, j) - \alpha_i \cdot D_c(i, i) \quad (8)$$

where the weight α_i is the ratio of the number of samples in \mathbf{K}_i to the total number of samples in all KDE models, and $\sum_{i=1}^n \alpha_i = 1$. The normalized color saliency values of all KDE models ($\sum_{i=1}^n KS_c(i) = 1$) calculated for the example image in Fig. 1(a) are shown in Fig. 1(c), in which each KDE model is represented using a bar with its mean color. We can see from Fig. 1(c) that the two KDE models (the 11th and 12th bar) have higher color saliency values, and they cover the colors of salient

object (the flower). On the other hand, the color saliency values of other KDE models that cover background colors are efficiently suppressed in Fig. 1(c).

Based on the center-surround scheme, which has been intensively explored using different representations in previous saliency models, salient objects are generally surrounded by background regions, and thus in the spatial domain, the colors of background regions usually have a wider distribution over the whole image than the colors of salient objects. In the following, the spatial distribution of KDE models is used to distinguish those models covering the colors of salient objects from other models. Similarly as (7), the spatial distance between any pair of KDE models, \mathbf{K}_i and \mathbf{K}_j , is defined as

$$D_s(i, j) = \frac{1}{2} \cdot \left[\frac{\sum_{p \in \mathcal{P}} C_i(p) \cdot \|\mathbf{d}_j^s(p)\|}{\sum_{p \in \mathcal{P}} C_i(p)} + \frac{\sum_{p \in \mathcal{P}} C_j(p) \cdot \|\mathbf{d}_i^s(p)\|}{\sum_{p \in \mathcal{P}} C_j(p)} \right] \quad (9)$$

Based on the above analysis, KDE models that mainly cover the colors of salient objects have shorter spatial distances to other KDE models. The spatial saliency for \mathbf{K}_i is thus defined as the reciprocal of the sum of weighted spatial distances between \mathbf{K}_i and all KDE models

$$KS_s(i) = \frac{1}{\sum_{j=1}^n \alpha_j \cdot D_s(i, j)}. \quad (10)$$

For the example image in Fig. 1(a), the normalized spatial saliency values for all KDE models ($\sum_{i=1}^n KS_s(i) = 1$) are shown in Fig. 1(d), in which the two KDE models (the 11th and 12th bar) also have higher spatial saliency values, while the spatial saliency values of other KDE models are suppressed.

By comparing (8) with (10), it should be noted that the intra-distance $D_s(i, i)$ is included in (10), while (8) only includes inter-distances $D_c(i, j)$, $\forall j \neq i$. The reason for such a difference is described as follows. The intra-distance $D_s(i, i)$ actually represents the spatial distribution of colors covered in \mathbf{K}_i , and thus it is considered for the evaluation of spatial saliency. However, the intra-distance $D_c(i, i)$ actually reflects the color homogeneity of the samples in \mathbf{K}_i . For the evaluation of color saliency, it is not reasonable to introduce such a bias that one KDE model covering more colors is more salient than another KDE model covering fewer colors, and thus $D_c(i, i)$ is excluded from the color saliency evaluation for KDE models.

C. Saliency Map Generation

Based on the color saliency values and spatial saliency values of KDE models, the pixel-wise color saliency map S_c and spatial saliency map S_s are generated as follows:

$$S_c(p) = \sum_{i=1}^n C_i(p) \cdot KS_c(i) \quad (11)$$

$$S_s(p) = \sum_{i=1}^n C_i(p) \cdot KS_s(i). \quad (12)$$

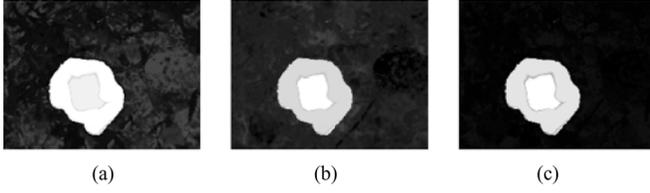


Fig. 2. Examples of saliency map generation. (a) Color saliency map. (b) Spatial saliency map. (c) Final saliency map.

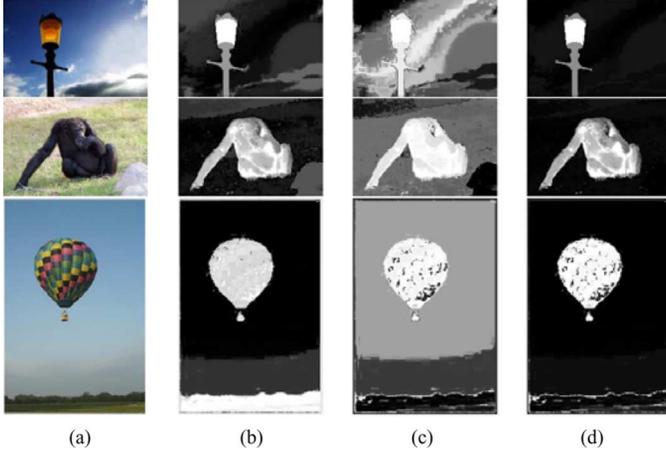


Fig. 3. More examples of saliency map generation. (a) Original images. (b) Color saliency maps. (c) Spatial saliency maps. (d) Final saliency maps.

Equation (11)–(12) indicates that the color/spatial saliency for each pixel is the sum of color/spatial saliency values of all KDE models weighted by its likelihoods to fit these KDE models. In this sense, the global color information of the image is actually incorporated into the saliency calculation for each local pixel. Based on Fig. 1(c) and (d), the pixel-wise color saliency map and spatial saliency map are, respectively, shown in Fig. 2(a) and (b), which are normalized into the range of $[0, 255]$ for display. By integrating color saliency map with spatial saliency map, the final saliency map S_f is generated as follows:

$$S_f(p) = S_c(p) \cdot S_s(p). \quad (13)$$

For the example image in Fig. 1(a), its final saliency map is shown in Fig. 2(c), which is also normalized into the range of $[0, 255]$ for display. Compared with Fig. 2(a) and (b), we can see that the salient object is completely highlighted, and background regions are more effectively suppressed in Fig. 2(c). Based on our observations on the saliency maps generated for a variety of images, we have found that color saliency map and spatial saliency map can complement each other to generate a more reasonable final saliency map, which can highlight salient object regions and suppress background regions more effectively (see more examples shown in Fig. 3).

III. TWO-PHASE GRAPH CUT

The saliency map generated using our KDE-based saliency model can provide useful cues for segmentation of salient objects, and a simple thresholding operation seems enough to extract the salient objects with acceptable quality for some saliency maps, in which salient object regions are sufficiently

highlighted and background regions are completely suppressed. Nonetheless, for segmentation reliability on a wide range of saliency maps and a higher segmentation quality, we propose a two-phase graph cut-based salient object segmentation approach. In the first phase, the saliency map based graph cut is exploited to efficiently obtain the initial salient object segmentation result. In the second phase, the object/background seeds are initialized using the initial salient object segmentation result, and the iterative seed adjustment-based graph cut is exploited to refine the salient object segmentation result using the gradually improved object/background seeds. Some basic terminologies of graph cut will be briefly reviewed in Section III-A. The first phase and the second phase of the proposed approach will be detailed in Section III-B and III-C, respectively.

A. Basic Terminologies of Graph Cut

Salient object segmentation is explicitly formulated as a binary pixel labeling problem, which can be solved under the framework of graph cut [31]. The input image is represented using an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where \mathcal{V} is a set of nodes and \mathcal{E} is a set of undirected edges connecting these nodes. Each node in the graph represents each pixel in the image, and there are two additional terminals in the graph, i.e., object terminal \mathcal{S} and background terminal \mathcal{T} . There are two types of edges in the graph. Specifically, edges between neighboring nodes are called *n-links* where *n* stands for “neighbor”, and edges connecting nodes to terminals are called *t-links* where *t* stands for “terminal”. All graph edges including *n-links* and *t-links* are assigned with some nonnegative costs. Formally, let \mathcal{N} denotes the set of all pairs of neighboring pixels in \mathcal{P} , which denotes the set of all pixels in the image. The two sets, \mathcal{V} and \mathcal{E} , are represented as follows:

$$\mathcal{V} = \mathcal{P} \cup \{\mathcal{S}, \mathcal{T}\} \quad (14)$$

$$\mathcal{E} = \mathcal{N} \cup_{p \in \mathcal{P}} \{(p, \mathcal{S}), (p, \mathcal{T})\} \quad (15)$$

where all *n-links* are included in \mathcal{N} , and (p, \mathcal{S}) and (p, \mathcal{T}) denote the *t-link* connecting with \mathcal{S} and \mathcal{T} , respectively.

A cut is defined as a subset of edges $\mathcal{C} \subset \mathcal{E}$, and nodes in the graph are separated by this subset of edges. Graph cut seeks to minimize a cost function with the following form to determine the optimal label configuration:

$$E(\mathbf{L}) = \sum_{p \in \mathcal{P}} R_{L_p}(p) + \lambda \sum_{(p,q) \in \mathcal{N}} B(p,q) \cdot \delta_{L_p \neq L_q} \quad (16)$$

where $\mathbf{L} = \{L_p\}$ is a binary vector denoting any possible label configuration of all pixels, L_p can be assigned with the label “bkg” for background or “obj” for salient object, and the Kronecker delta $\delta_{L_p \neq L_q}$ is defined as

$$\delta_{L_p \neq L_q} = \begin{cases} 1, & \text{if } L_p \neq L_q \\ 0, & \text{if } L_p = L_q. \end{cases} \quad (17)$$

$R_{L_p}(p)$ is the data term based on the label L_p , $B(p,q)$ is the smoothness term for neighboring pixels (p,q) , and λ is the weight for balancing the two terms. The data term $R_{L_p}(p)$ penalizes the inconsistency between a label L_p and the observed data such as saliency value and color feature of a pixel p , and the smoothness term $B(p,q)$ penalizes the label discontinuity

of neighboring pixels (p, q) . The minimum cut of the graph can be efficiently solved using the max-flow algorithm [38], and the corresponding binary labels of pixels are used to represent the salient object segmentation result.

B. Saliency Map Based Graph Cut

We observe that a considerable portion of saliency maps generated using our saliency model can generally highlight salient object regions and suppress background regions. Therefore, we first exploit the saliency map to define the cost function of graph cut, and obtain the initial segmentation result of salient objects. Based on the saliency map S_f , for each pixel p , the confidence belonging to the salient object is defined as

$$P_{\text{obj}}(p) = \frac{S_f(p)}{255} \cdot \frac{S_f(p)}{m_s} \quad (18)$$

where m_s is the average saliency value of S_f . In (18), the former term is used as the basic measure to estimate for each pixel p the confidence belonging to the salient object, i.e., a pixel with a higher saliency value is more likely to belong to the salient object. The latter term in (18) is introduced as an adjusting factor to reasonably enlarge the differences of the evaluated confidences between high-saliency pixels and low-saliency pixels in the saliency map with lower contrast. Based on the saliency map S_f , for each pixel p , the confidence belonging to the background is similarly defined as

$$P_{\text{bkg}}(p) = \frac{255 - S_f(p)}{255} \cdot \frac{255 - S_f(p)}{255 - m_s}. \quad (19)$$

Based on (18) and (19), the data term for each pixel p is defined as

$$R_l(p) = \frac{P_l(p)}{P_l(p) + P_{\bar{l}}(p)} \quad (20)$$

where the subscript l may denote ‘‘obj’’ or ‘‘bkg’’, and its complement \bar{l} denotes ‘‘bkg’’ or ‘‘obj’’, accordingly.

Based on the general observation that neighboring pixels with similar saliency values are highly likely to belong to the same salient object or background, the smoothness term for any pair of neighboring pixels (p, q) is defined as

$$B(p, q) = \frac{g_{p,q}}{1 + \|S_f(p) - S_f(q)\|^2} \quad (21)$$

where the coefficient $g_{p,q}$ is defined as

$$g_{p,q} = \max \left[\frac{S_f(p) + S_f(q)}{2}, 255 - \frac{S_f(p) + S_f(q)}{2} \right]. \quad (22)$$

The coefficient $g_{p,q}$ is actually used as a local balancing weight on the basis of neighboring pixels, to replace the role of the global balancing weight λ . Specifically, λ is set to 1 in the first phase. With the introduction of $g_{p,q}$, the smoothness term is further increased for those neighboring pixels that both have higher/lower saliency values, and thus the label smoothness, i.e., the probability that the neighboring pixels should be assigned with the same label, is increased reasonably.

The graph is constructed based on the above defined data term and smoothness term, and then graph cut is performed

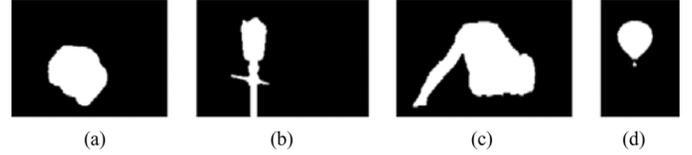


Fig. 4. Initial salient object segmentation results obtained using the saliency map based graph cut. The corresponding saliency maps are shown in Figs. 2(c) and 3(d).

to obtain the initial salient object segmentation result. For the saliency map in Fig. 2(c) and the three saliency maps in Fig. 3(d), the initial salient object segmentation results are shown in Fig. 4(a)–(d). It can be seen from Fig. 4 that the saliency map based graph cut can accurately segment the salient objects using the saliency maps with high contrast between salient objects and background regions, and can overcome the negative effect of some falsely highlighted/suppressed small background/object regions in the saliency map [see the bottom row in Fig. 3(d), and Fig. 4(d)].

However, for some images shown in Fig. 5(a), whose saliency maps in Fig. 5(b) show relatively low contrast between parts of the salient objects and the surrounding background regions, some redundant background regions are erroneously contained in the initial salient object segmentation results as shown in Fig. 5(c). We can see from Fig. 5 that only saliency map information may be insufficient to obtain an acceptable segmentation of salient objects. Therefore, in Section III-C, we will present the iterative seed adjustment-based segmentation refinement method to refine such unsatisfactory initial segmentation results.

C. Segmentation Refinement Based on Iterative Seed Adjustment

In the second phase, we refine the segmentation result based on the iterative seed adjustment, which efficiently utilizes the information of minimum cut with the introduction of balancing weight update scheme. The proposed segmentation refinement method consists of the following four steps.

Step 1) The object/background pixels in the initial salient object segmentation result is used as the object/background seeds, which will be updated in the following iterative seed adjustment process. Two KDE models are constructed based on object seeds and background seeds, respectively. For clarity of description, we use a trimap to represent the object seeds (white), background seeds (black), and uncertain pixels (gray), which are denoted by the three sets Ω_{obj} , Ω_{bkg} , and Ω_{un} , respectively (Section III-C1).

Step 2) For each pixel, the confidence belonging to the object/background is calculated using the KDE model constructed based on object/background seeds. The graph is then constructed by redefining the cost terms and the balancing weight, and graph cut is performed to obtain the minimum cut (Section III-C1).

Step 3) Based on the analysis of minimum cut, the object/background seeds are adjusted and used to update the trimap (Section III-C2), and the balancing weight is also adaptively updated (Section III-C3).

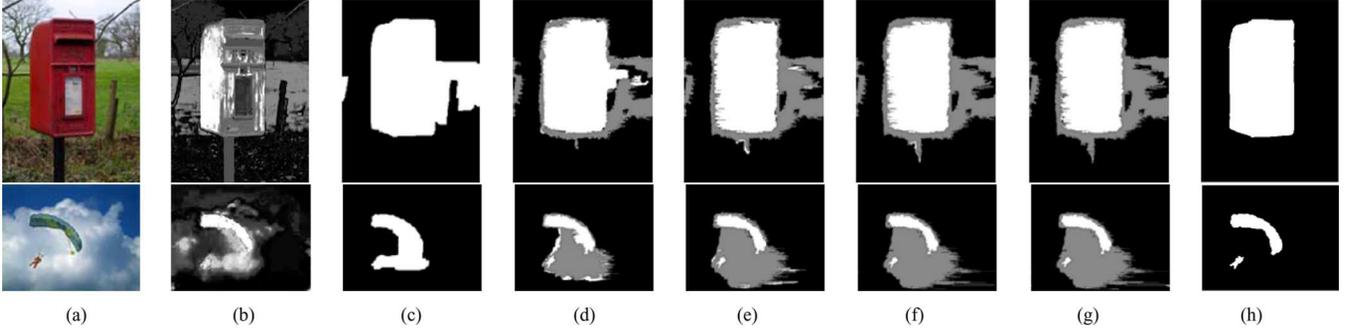


Fig. 5. Illustration of salient object segmentation process for the image *postbox* and *parachute jump*. (a) Original images. (b) Saliency maps. (c) Initial salient object segmentation results. (d)–(g) Updated trimaps representing object seeds (white), background seeds (black), and unknown pixels (gray) during the iterative seed adjustment process (from the 1st to the 4th iteration). (h) Final salient object segmentation results.

Step 4) The information of minimum cut is exploited to determine whether the iteration process from Step 2) to 3) should be repeated or not. If the iteration process is terminated, the graph is constructed based on the finally refined object/background seeds, and graph cut is performed to obtain the final salient object segmentation result (Section III-C4).

1) *KDE Model Based Graph Cut*: Using the seed pixels in Ω_{obj} and Ω_{bkg} as samples, two KDE models are constructed for salient object and background, respectively. For each pixel p , the confidence belonging to the object/background is redefined as

$$P_l(p) = \frac{1}{|\Omega_l|} \sum_{q \in \Omega_l} \kappa_l(\mathbf{c}_p - \mathbf{c}_q) \quad (23)$$

where the subscript “ l ” may denote “obj” or “bkg”, and Ω_l may denote $\Omega_{obj}/\Omega_{bkg}$ for the set of object/background seeds. Gaussian distribution is used as the kernel function κ_l , and the estimation of its bandwidth matrix is the same as that described in Section II-A.

Based on the redefined confidence belonging to the object/background, the data term $R_l(p)$ is recalculated using (20). The smoothness term $B(p, q)$ is redefined as

$$B(p, q) = \frac{1}{1 + \|\mathbf{c}_p - \mathbf{c}_q\|^2}. \quad (24)$$

The global balancing weight λ is used in the graph cut during the segmentation refinement process, and λ is iteratively updated based on the minimum cut obtained in each iteration (see Section III-C3). The initialization of λ is based on the analysis of the histogram with 256 bins generated for the saliency map S_f . Let N_s denote the number of bins whose values are greater than the average value of all bins, and σ_s denote the standard deviation of S_f . The balancing weight λ is initialized as

$$\lambda^{(1)} = N_s \cdot \frac{128}{\sigma_s} \quad (25)$$

where the superscript “1” denotes the initialization of λ used in the first iteration. For saliency maps that sufficiently highlight salient objects and suppress background regions, a smaller value of N_s and a larger value of σ_s result in a smaller value of $\lambda^{(1)}$, which puts relatively more confidence on the data term

for salient object segmentation. Therefore, the initial balancing weight is set adaptive to the quality of saliency map. The constant coefficient is set to 128, the possibly achieved maximum value of σ_s , for a reasonable range of the balancing weight.

The graph is constructed based on the updated data term, smoothness term and the balancing weight, and graph cut is performed to obtain the minimum cut, which is exploited to adjust the object/background seeds in the following subsection.

2) *Seed Adjustment*: The objective of seed adjustment process is to gradually refine object/background seeds by utilizing the information of minimum cut for a reliable segmentation of salient objects. The seed adjustment process in one iteration is detailed as follows. In the i th iteration, the possibly inaccurate background/object seeds are removed from $\Omega_{bkg}/\Omega_{obj}$ and considered as temporary object/background seeds, which are added into the temporary seed set $\Theta_{obj}/\Theta_{bkg}$ for further determination

$$\begin{aligned} p^{(i)} &\in \Theta_{obj}, \text{ if } p^{(i)} \in \Omega_{bkg} \text{ and } \{p^{(i)}, \mathcal{T}\} \in \mathcal{C}^{(i)} \\ p^{(i)} &\in \Theta_{bkg}, \text{ if } p^{(i)} \in \Omega_{obj} \text{ and } \{p^{(i)}, \mathcal{S}\} \in \mathcal{C}^{(i)} \end{aligned} \quad (26)$$

where $\mathcal{C}^{(i)}$ denotes the minimum cut obtained using the KDE model based graph cut, in which the KDE model for object/background is constructed using the seed pixels in $\Omega_{obj}/\Omega_{bkg}$ in the i th iteration. The rationality for (26) is explained as follows. For a background/object seed pixel $p^{(i)}$ in the i th iteration, if its t -link with the background/object terminal \mathcal{T}/\mathcal{S} is cut off, it is likely that such a pixel is not a reliable background/object seed, and thus is removed from $\Omega_{bkg}/\Omega_{obj}$ and correspondingly added into $\Theta_{obj}/\Theta_{bkg}$.

These temporary seeds in $\Theta_{obj}/\Theta_{bkg}$ are used to update each pixel’s confidence belonging to the object/background using (23). Then the graph is re-constructed by only updating the data terms, and graph cut is performed again to obtain a new minimum cut $\mathcal{C}_{new}^{(i)}$. The object/background seeds for the next iteration are determined based on the adjustment rules listed in Table I. In the case of $p^{(i)} \in \Theta_{obj}$, the rationality for the listed rules is explained as follows. If its t -link with the object terminal \mathcal{S} is not cut off, it further enhances the possibility that $p^{(i)}$ should be used as an object seed and added into Ω_{obj} in the next iteration. However, if its t -link with the object terminal \mathcal{S} is cut off, it indicates that $p^{(i)}$ is not a reliable seed and

TABLE I
RULES FOR SEED ADJUSTMENT

Temporary seed set	Condition	Determined seed set
$p^{(i)} \in \Theta_{obj}$	$\{p^{(i)}, \mathcal{S}\} \notin \mathcal{C}_{new}^{(i)}$	$p^{(i+1)} \in \Omega_{obj}$
	$\{p^{(i)}, \mathcal{S}\} \in \mathcal{C}_{new}^{(i)}$	$p^{(i+1)} \in \Omega_{un}$
$p^{(i)} \in \Theta_{bkg}$	$\{p^{(i)}, \mathcal{T}\} \notin \mathcal{C}_{new}^{(i)}$	$p^{(i+1)} \in \Omega_{bkg}$
	$\{p^{(i)}, \mathcal{T}\} \in \mathcal{C}_{new}^{(i)}$	$p^{(i+1)} \in \Omega_{un}$

should be put into the set Ω_{un} in the next iteration. The similar explanation is also applicable to the case of $p^{(i)} \in \Theta_{bkg}$.

Starting from the inaccurate initial segmentation results shown in Fig. 5(c), the seed adjustment processes are shown in Fig. 5(d)–(g). There are a total of 4 iterations for the two examples. We can observe from the trimaps in Fig. 5(d)–(g) that the object/background seeds become more and more suitable for salient object segmentation during the seed adjustment process.

3) *Balancing Weight Update*: The balancing weight used in each iteration is an important factor to control seed adjustments. Based on the balancing weight in the i th iteration, the balancing weight used in the next iteration is adaptively updated as follows:

$$\lambda^{(i+1)} = \min \left[1, \max \left(0.1, \beta \cdot \frac{N_c^{(i)}}{|\mathcal{P}|} \right) \right] \cdot \lambda^{(i)} \quad (27)$$

where $N_c^{(i)}$ denotes the number of t -links belonging to the minimum cut $\mathcal{C}^{(i)}$. The coefficient β is used to maintain a suitable range of the balancing weight, and is set to 10 by the experiments.

The form of (27) ensures that the balancing weight monotonically decreases during the whole iteration process. Using the max-flow algorithm for graph cut, a larger value of the balancing weight indicates that the capacity of t -links is easily saturated and likely to be cut off, while a smaller value indicates that the t -links are unlikely to be cut off. Therefore, $N_c^{(i)}$ will decrease during the whole iteration process, and it can be seen from (26) that the temporary object/background seeds selected in each iteration will become fewer and fewer. In this sense, the object/background seeds will become more and more stable. Besides, it can be seen from (27) that the degressive trend of the balancing weight is further enhanced due to the decrease of $N_c^{(i)}$ during the whole iteration process. Therefore, theoretically, the balancing weight update scheme can guarantee the convergence of the iterative seed adjustment process. Experimentally, we can observe from Fig. 5(d)–(g) that inaccurate seeds are gradually corrected as either accurate seeds or unknown pixels during several iterations. More examples of iterative seed adjustment are shown in Fig. 11.

4) *Final Segmentation*: By combining the seed adjustment with the adaptive update of balancing weight, we can obtain more reliable object/background seeds. We exploit the value of $N_c^{(i)}$ to determine whether the iteration process should be terminated or not. If the initial segmentation result obtained using saliency map based graph cut is already acceptable, the refinement only slightly improves the segmentation quality by seed

adjustments and not absolutely necessary. Therefore, we terminate the seed adjustment process after the 1st iteration if $N_c^{(i)}$ is less than 7.5% of the total number of t -links, a relatively larger value, which is effective to timely terminate the iteration process starting from an initial segmentation result with acceptable quality such as the examples in Fig. 4. The subsequent iterations are exploited to gradually refine the seeds for a reliable segmentation, and thus a rather smaller value, 0.5% of the total number of t -links, is used as the termination condition.

For the two examples in Fig. 5, their seed adjustment processes are terminated after 4 iterations, and the finally refined object/background seeds and the uncertain pixels are represented using the trimaps in Fig. 5(g). Based on the finally refined object/background seeds, the KDE model based graph cut is performed to obtain the binary labeling result, which is used as the final salient object segmentation result. As shown in Fig. 5(h), we can see that single or multiple salient objects can be accurately extracted with well-defined boundaries. A visual comparison between Fig. 5(h) and (c) obviously demonstrates the effectiveness of the iterative seed adjustment-based segmentation refinement method.

IV. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed salient object segmentation approach on an image test set [14] with manually segmented ground truths for salient objects in 1000 images (publicly available at http://ivrg.epfl.ch/supplementary_material/RK_CVPR09/GroundTruth/binarymasks.zip), which are selected from MSRA SOD (Microsoft Research Asia Salient Object Database, Image Set B) containing 5000 high-quality images [21]. First, we generate the saliency maps for all test images using the proposed KDE-based saliency model, and compare the saliency detection performance with five state-of-the-art saliency models, i.e., the most well-known Itti's model [8] and four recent saliency models including Zhang's model [19], Cheng's model [24], Achanta's model [15], and Rahtu's model [34] in Section IV-A. Then, we perform salient object segmentation using the proposed two-phase graph cut approach, and compare the segmentation performance with two state-of-the-art salient object segmentation approaches, i.e., Achanta's approach [15] and Rahtu's approach [34]. Subjective evaluation and objective evaluation of salient object segmentation are presented in Section IV-B and IV-C, respectively. Besides, we analyze how the performance of pre-segmentation using mean shift directly affects the quality of saliency maps in Section IV-A and finally affects the quality of salient object segmentation results in Section IV-C. Finally, we discuss some possible extensions based on our approach in Section IV-D.

A. Performance Evaluation of Saliency Models

For performance evaluation of different saliency models, we use the implementation code of Saliency Toolbox [39] for Itti's saliency model, and the authors' implementation codes for the other four saliency models. For comparison with other saliency models, the only parameter τ in our saliency model is set to 0.03. Using the six saliency models, we generate six classes of saliency maps for all 1000 test images. A subjective comparison of saliency maps generated using different saliency models is shown in Fig. 6. Compared with the other

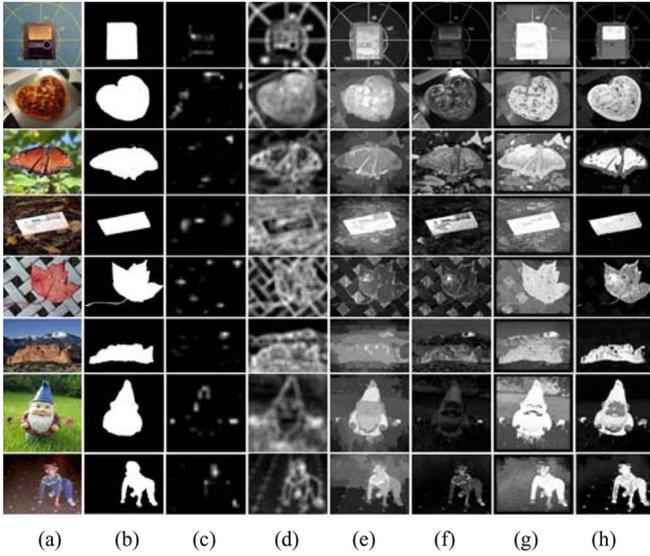


Fig. 6. Subjective comparison of saliency maps generated using different saliency models. (a) Original images. (b) Ground truths. Saliency maps generated using (c) Itti's model, (d) Zhang's model, (e) Cheng's model, (f) Achanta's model, (g) Rahtu's model, and (h) our model, respectively.

five saliency models, we can see from Fig. 6 that salient object regions can be more completely highlighted with well-defined boundaries, and background regions can be more effectively suppressed in the saliency maps generated using our saliency model. Therefore, we can anticipate that our saliency maps are generally more applicable to salient object segmentation.

In order to objectively evaluate the saliency detection performance of the six saliency models, we adopt the two commonly used objective measures, i.e., precision and recall, and plot the precision-recall curves for comparison. For each test image, the binary ground truth is denoted by G , and the binary salient object mask generated by thresholding the saliency map is denoted by M . In both G and M , each object pixel is labeled as "1" and each background pixel is labeled as "0", the precision and recall are defined as

$$\text{precision} = \frac{\sum_{(x,y)} M(x,y) \cdot G(x,y)}{\sum_{(x,y)} M(x,y)} \quad (28)$$

$$\text{recall} = \frac{\sum_{(x,y)} M(x,y) \cdot G(x,y)}{\sum_{(x,y)} G(x,y)}. \quad (29)$$

All the six classes of saliency maps are first normalized into the same range of $[0, 255]$. Then we use a series of fixed integer thresholds from 0 to 255, and obtain 256 binary salient object masks for each saliency map. At each threshold, the precision/recall measures for all 1000 saliency maps are averaged, and as shown in Fig. 7, the precision-recall curve of each saliency model plots the 256 average precision measures against the 256 average recall measures. The precision-recall curves present a robust comparison of saliency detection performance. These curves indicate how well different classes of saliency models can highlight salient objects and suppress background regions. We can see from Fig. 7 that the precision-recall curve of our saliency model is the highest one, and thus we can conclude that

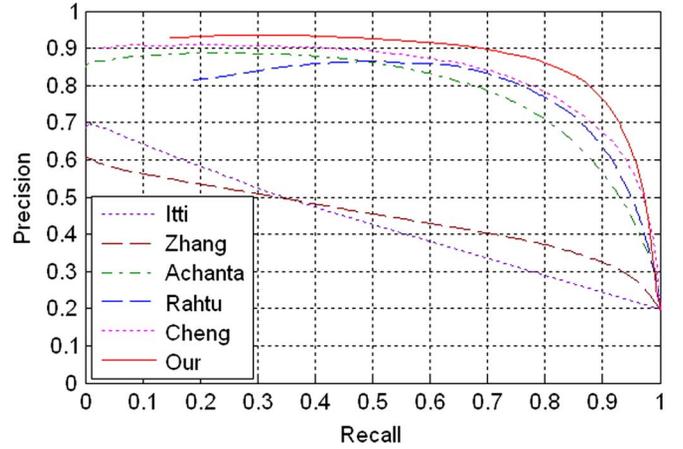


Fig. 7. Precision-recall curves of the six saliency models.

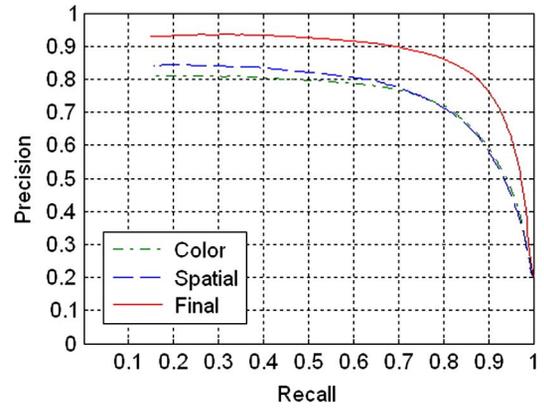


Fig. 8. Precision-recall curves generated using our color saliency maps, spatial saliency maps, and final saliency maps.

the quality of our saliency maps is generally better than the other five classes of saliency maps for salient object segmentation.

As stated in Section II-C, in our saliency model, the complementary effect of color saliency map and spatial saliency map contributes to the generation of more reasonable final saliency map. Similarly as Fig. 7, three precision-recall curves generated using our color saliency maps, spatial saliency maps, and final saliency maps are shown in Fig. 8. We can observe from Fig. 8 that the precision-recall curve generated using our final saliency maps is obviously higher than the other two precision-recall curves. Therefore, Fig. 8 objectively demonstrates the complementary effect of color saliency map and spatial saliency map.

We further evaluate how the pre-segmentation performance of the mean shift algorithm affects the quality of our saliency maps. We adjust the only parameter τ to control the degree between over-segmentation and under-segmentation in the pre-segmentation result, and generate a set of saliency maps with different values of τ . Similarly as Fig. 7, five precision-recall curves generated by setting τ from 0.008 (the finest pre-segmentation) to 0.07 (the coarsest pre-segmentation) are shown in Fig. 9. We can see from Fig. 9 that the quality of our saliency maps generally degrades with the increase of τ , but the three precision-recall curves with τ not greater than 0.03 are very close. Therefore, we can conclude that the performance of

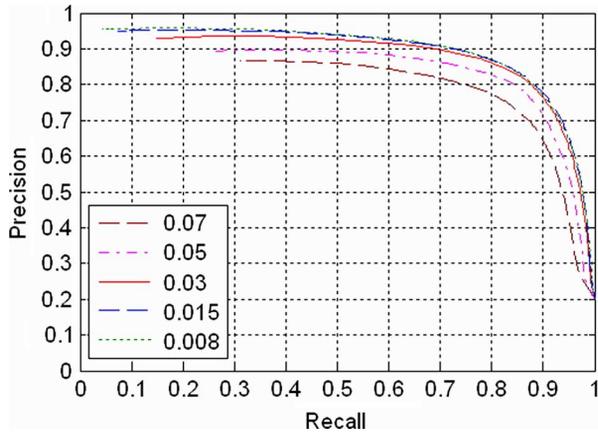


Fig. 9. Precision-recall curves generated using our saliency model with different values of τ .

our saliency model is consistently robust to τ with a value not greater than 0.03.

B. Subjective Evaluation of Salient Object Segmentation

In order to evaluate the performance of salient object segmentation, we perform experiments on all 1000 test images using the proposed two-phase graph cut approach, and compare the segmentation performance with Achanta's approach [15] and Rahtu's approach [34]. We use our saliency maps generated with $\tau = 0.03$ in our two-phase graph cut approach for the following comparisons with the other two approaches in Section IV-B.

The results for some test images are shown in Fig. 10, in which both saliency maps and salient object segmentation results generated using the three approaches are shown for subjective comparison. For images with obvious contrast between the salient object and a simple background (the 1st–4th rows in Fig. 10), the salient objects are highlighted with well-defined boundaries in the three classes of saliency maps, and the salient objects segmented using the three approaches are visually acceptable. However, for images with relatively low contrast between some parts of salient objects and the surrounding background regions (the 5th–7th rows in Fig. 10), the quality of both Achanta's and Rahtu's saliency maps is obviously degraded. For images with more complex backgrounds, which may contain strong structures and texture patterns (the 8th–10th rows in Fig. 10), the center-surround scheme exploited in both Achanta's and Rahtu's saliency model cannot effectively suppress such background regions with higher local contrast.

In contrast, our saliency model efficiently utilizes the global information of the image to evaluate the saliency measures of KDE models, and thus can efficiently suppress background regions in such images (the 5th–10th rows in Fig. 10). Therefore, the salient object segmentation results obtained using Achanta's approach and Rahtu's approach contain irrelevant background regions and/or incomplete salient objects, while our approach can completely segment the salient objects with well-defined boundaries due to the relatively high-quality saliency maps. From the observation of the two-phase graph cut performed on these images in Fig. 10, we have found that the initial segmentation results obtained in the first phase are

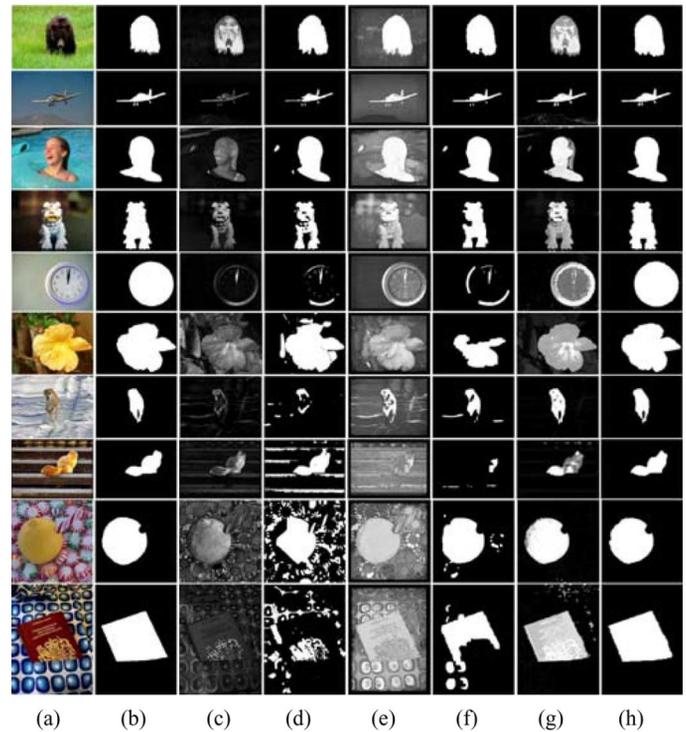


Fig. 10. Subjective comparison of some salient object segmentation results. (a) Original images. (b) Ground truths. (c) Achanta's saliency maps. (d) Achanta's segmentation results. (e) Rahtu's saliency maps. (f) Rahtu's segmentation results. (g) Our saliency maps. (h) Our segmentation results.

sufficiently accurate, and the seed adjustment processes in the second phase terminate after the first iteration. Therefore, we can conclude that a high-quality saliency map can generally guarantee a quick convergence of segmentation refinement.

More segmentation results on some complicated images with various scenes are shown in Fig. 11, which illustrates the iterative seed adjustment process of our approach, and also shows the saliency maps and segmentation results obtained using Achanta's approach and Rahtu's approach for subjective comparison. Compared with Fig. 10, the quality of our saliency maps in Fig. 11 is lower, i.e., the complex background regions cannot be efficiently suppressed and/or some parts of complex salient objects cannot be efficiently highlighted, and our initial segmentation results are not visually satisfactory. We can observe from the former four examples in Fig. 11 that the iterative seed adjustment method can gradually refine the object/background seeds, and guarantees the acceptable quality of final segmentation results. For these examples, we can see that Achanta's segmentation results are highly dependent on the saliency maps, and thus the relatively low-quality saliency maps significantly degrade the quality of segmentation results. Rahtu's approach also cannot efficiently overcome the insufficiency of their saliency maps, and some background regions that are highlighted in their saliency maps appear in their segmentation results. Therefore, with relatively low-quality saliency maps, it is not reliable that the one-shot graph cut used in Achanta's approach and Rahtu's approach guarantees acceptable segmentation results, while our two-phase graph cut approach can exploit the iterative seed adjustment process to obtain possibly refined segmentation results.

No.	Original image (top) and ground truth (bottom)	Achanta's approach	Rahtu's approach	Our approach			
		Saliency map (top) Segmentation result (bottom)	Saliency map (top) Segmentation result (bottom)	Pre-segmentation (top)	Saliency map (top)	Initial segmentation result (top)	Final segmentation result (top)
1							
							N/A
2							
							N/A
3							
4							
5							
					N/A	N/A	N/A
6							
					N/A	N/A	N/A
7							
							N/A

Fig. 11. Subjective comparison of salient object segmentation using Achanta's approach, Rahtu's approach, and our approach with the illustration of the iterative seed adjustment process. The symbol N/A (Not Available) in the top two examples and the bottom three examples indicates that the iterative seed adjustment process performs less than 4 iterations. For these five examples, the total number of iteration is 3, 3, 1, 1, and 3, respectively, from top to bottom.

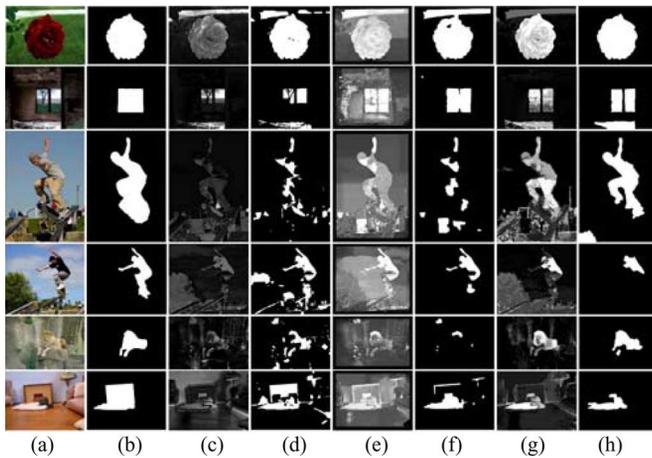


Fig. 12. Salient object segmentation results on some complicated images. (a) Original images. (b) Ground truths. (c) Achanta's saliency maps. (d) Achanta's segmentation results. (e) Rahtu's saliency maps. (f) Rahtu's segmentation results. (g) Our saliency maps. (h) Our segmentation results.

However, we also find that the proposed iterative seed adjustment method cannot achieve substantial segmentation refinements for some complicated images as shown in the latter three examples of Fig. 11. In such images, some background regions are visually salient against the main part of background (the 5th example), salient objects contain multiple heterogeneous regions (the 6th example), or there are very similar colors between some parts of the salient object and background regions (the last example). For such images, the main part of background is sufficiently suppressed, but such visually salient background regions are also highlighted and/or some regions of salient objects are also effectively suppressed in our saliency maps. It is unfeasible for the iterative seed adjustment method to effectively correct unsuitable object/background seeds in such cases, and thus the refinements on our initial segmentation results are not noticeable.

Fig. 12 shows more results on such complicated images with visually salient background regions (the 1st and 2nd examples), heterogeneous salient object (the 3rd and 4th examples), and similar colors between salient object and background regions (the 5th and 6th examples). We can see from Fig. 12 that our approach achieves a better segmentation quality for the 1st, 3rd, and 5th examples, while Rahtu's approach outperforms our approach on the other three examples. We can further observe from Fig. 12 that the contrast between salient object and background in the saliency map is the major factor to affect the segmentation quality for the three approaches.

Salient object segmentation results on more test images are shown in Fig. 13 for subjective comparison. We can see from Fig. 13 that the quality of Achanta's segmentation results is generally lower than Rahtu's results and our results. Compared with Rahtu's approach, our approach can generally segment the more complete salient objects with well-defined boundaries. The segmentation results shown in Fig. 13 as well as Figs. 10–12 demonstrate that our approach achieves an overall better subjective segmentation quality than Achanta's approach and Rahtu's approach. The examples shown in the bottom part of Fig. 13 further demonstrate that it is generally difficult to obtain satisfactory salient object segmentation results on

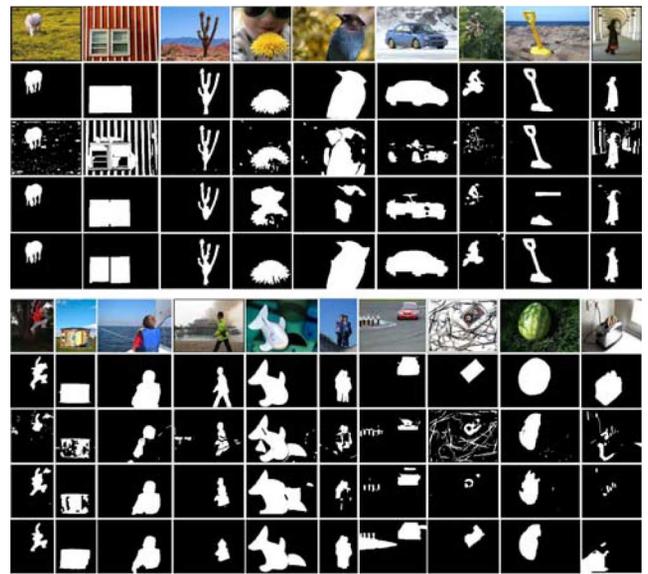


Fig. 13. Subjective comparison of more salient object segmentation results. From top to bottom: original images, ground truths, Achanta's segmentation results, Rahtu's segmentation results, and our segmentation results.

such complicated images as the examples in Fig. 12. For an unsupervised salient object segmentation approach, it is generally unreliable to compose many heterogeneous regions into a complete salient object, remove visually salient background regions, and separate salient object regions from background regions with very similar colors, since it is likely that such object (resp. background) regions show very low contrast with the correctly suppressed background regions (resp. highlighted object regions) in the saliency maps.

C. Objective Evaluation of Salient Object Segmentation

We further objectively evaluate the quality of salient object segmentation results using the measures of precision and recall, and an overall performance measure, F-measure, which is defined as

$$F_{\gamma} = \frac{(1 + \gamma) \cdot \text{precision} \cdot \text{recall}}{\gamma \cdot \text{precision} + \text{recall}} \quad (30)$$

where the coefficient γ is set to 0.5 in our experiments. The precision and recall are calculated for each image using (28) and (29), in which M denotes the binary salient object mask obtained using each approach. The three measures are averaged over all 1000 test images to evaluate the segmentation performance of each approach. Table II shows the three measures achieved using Achanta's approach, Rahtu's approach, and our approach with different pre-segmentation results, which are generated by adjusting the parameter τ in the mean shift algorithm. As already shown in Fig. 9, the overall quality of our saliency maps is similar when τ is not greater than 0.03. We can see from Table II that our approach achieves a consistently higher segmentation performance in terms of F-measure when τ is not greater than 0.03, and outperforms Achanta's approach and Rahtu's approach on all the three measures when τ is not greater than 0.05. Therefore, Table II not only demonstrates the overall better segmentation performance of our approach, but

TABLE II
OBJECTIVE COMPARISON ON SEGMENTATION PERFORMANCE OF ACHANTA’S APPROACH, RAHTU’S APPROACH, AND OUR APPROACH WITH DIFFERENT PRE-SEGMENTATION RESULTS

Segmentation approach	Achanta	Rahtu	Our (five different values of τ)				
			0.008	0.015	0.03	0.05	0.07
Precision	0.828	0.863	0.914	0.909	0.898	0.867	0.822
Recall	0.731	0.786	0.799	0.823	0.835	0.830	0.822
F-measure	0.793	0.836	0.872	0.879	0.876	0.855	0.822

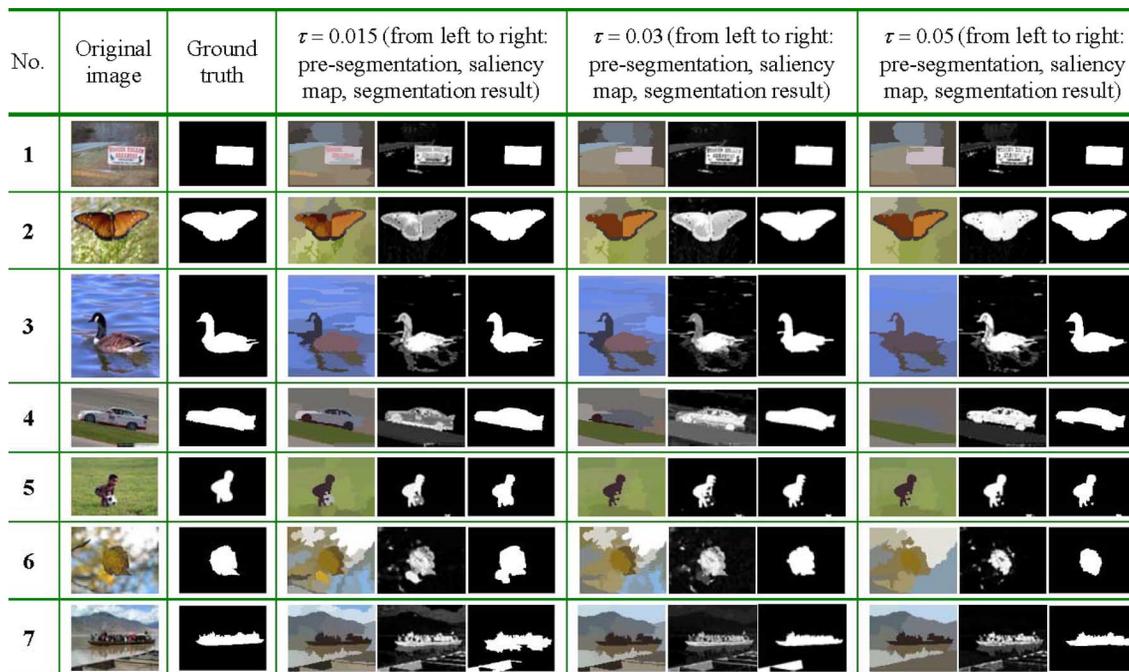


Fig. 14. Salient object segmentation results obtained with different pre-segmentation results.

also shows its overall robustness to different pre-segmentation results.

Based on the observation of our segmentation results, we have found that for most images, the quality of our segmentation results is not sensitive to different pre-segmentation results obtained with different values of τ . Fig. 14 shows a series of pre-segmentation results, saliency maps, and salient object segmentation results by setting τ to 0.015, 0.03, and 0.05, respectively. For most images such as the former four examples in Fig. 14, the saliency maps obtained with different pre-segmentation results have a similarly high quality, and thus the different salient object segmentation results are visually acceptable. Although a suitable pre-segmentation result, in which salient object boundaries are well preserved with a reasonable number of segmented regions cannot always be obtained using the mean shift algorithm, especially for $\tau = 0.05$ in these examples, our saliency maps show the robustness to τ and guarantee the quality of segmentation results. However, we also notice that for some complicated images such as the latter three examples in Fig. 14, the under-segmentation of a heterogeneous salient object (the 5th example) and the over-segmentation of the complex background (the last two examples) in the pre-segmentation results affect the quality of saliency maps, and finally degrade the quality of segmentation results. The best segmentation quality is achieved by setting τ to 0.015 for the 5th example, 0.03 for the

6th example, and 0.05 for the last example, respectively. Therefore, it is possible to improve the segmentation quality by tuning τ for some complicated images. In summary, we can conclude from Table II and Fig. 14 that the pre-segmentation performance of the mean shift algorithm does not affect the overall robustness of our approach, but may affect the segmentation quality of some complicated images.

However, we notice that for some complicated images, it is nontrivial to preserve well-defined boundaries between different regions in the pre-segmentation results by parameter tuning of the mean shift algorithm. In order to improve the pre-segmentation quality, which partly affects the quality of saliency map and salient object segmentation result, we will try to develop a more suitable image segmentation algorithm to replace the mean shift algorithm in our future work. Specifically, some superpixel segmentation algorithm [40] can be first used to obtain an over-segmentation result with uniform-sized regions, and then a scale-aware region merging algorithm using statistical models will be designed to obtain a moderate segmentation result.

D. Discussion

As demonstrated by previous two subsections, our approach shows an overall better segmentation performance both subjectively and objectively. It should be noted that unsupervised



Fig. 15. Some preliminary results of human object segmentation.

salient object segmentation approaches are designed to be general for a variety of images, and it is likely that any unsupervised approach cannot segment the user-desired salient objects from some complicated images. Our approach can serve as a base for developing an efficient interactive object segmentation tool, which can permit the user to flexibly refine the unsatisfied salient object segmentation results for some complicated images using simple user interactions. Besides, the proposed two-phase graph cut framework can serve as a general segmentation tool for different applications. For example, we are currently developing a frontal human object segmentation system for virtual video conference, and some preliminary results of human object segmentation are shown in Fig. 15. We incorporate the specific high-level knowledge about human object, i.e., a template-based human model, with our two-phase graph cut framework, and can efficiently segment single or multiple human objects.

V. CONCLUSION

In this paper, we have presented an efficient unsupervised salient object segmentation approach using KDE and the two-phase graph cut. The proposed saliency model utilizes the color saliency and spatial saliency of KDE models to generate a more appropriate saliency map for salient object segmentation. The proposed two-phase graph cut exploits the saliency map in the first phase, and combines the KDE model based graph cut, iterative seed adjustment based on the analysis of minimum cut, and the balancing weight update scheme in the second phase, to enhance the segmentation reliability for complicated images and improve the overall segmentation quality. Experimental results on a collection of 1000 test images demonstrate the better segmentation performance of our approach. We believe that the proposed salient object segmentation approach can be incorporated into object-based image retrieval and image adaptation systems to improve their performances. In our future work, we will extend the current framework to segment salient objects from videos by developing a spatiotemporal saliency model and incorporating an efficient object tracking method.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the associate editor for their valuable comments, which have greatly helped us to make improvements. The authors would also like to thank Dr. Achanta and Dr. Rahtu for providing their research codes for comparison.

REFERENCES

- [1] H. Fu, Z. Chi, and D. Feng, "Attention-driven image interpretation with application to image retrieval," *Pattern Recognit.*, vol. 39, no. 9, pp. 1604–1621, Sep. 2006.
- [2] W. H. Cheng, C. W. Wang, and J. L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.
- [3] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, "Retargeting images and video for preserving information saliency," *IEEE Comput. Graphics Appl.*, vol. 27, no. 5, pp. 80–88, Sep. 2007.
- [4] K. N. Ngan and H. Li, "Semantic object segmentation," *IEEE Communications Society Multimedia Communications Technical Committee E-Letter*, vol. 4, no. 6, pp. 6–8, Jul. 2009.
- [5] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.
- [6] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
- [7] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.
- [8] 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.
- [9] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2003, pp. 374–381.
- [10] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *Proc. IEEE ICME*, Jul. 2006, pp. 1477–1480.
- [11] W. Kim, C. Jung, and C. Kim, "Saliency detection: A self-ordinal resemblance approach," in *Proc. IEEE ICME*, Jul. 2010, pp. 1260–1265.
- [12] O. Le Meur and J. C. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2801–2813, Nov. 2010.
- [13] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proc. IEEE CVPR Workshops*, Jun. 2009, pp. 45–52.
- [14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.
- [15] R. Achanta and S. Susstrunk, "Saliency detection using maximum symmetric surround," in *Proc. IEEE ICIP*, Sep. 2010, pp. 2653–2656.
- [16] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Trans. Image Process.*, vol. 17, no. 5, pp. 633–644, May 2008.
- [17] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE CVPR*, Jun. 2007, p. 4270292.
- [18] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.
- [19] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vision*, vol. 8, no. 7, Dec. 2008, article 32.
- [20] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2368–2375.
- [21] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proc. IEEE CVPR*, Jun. 2007, p. 4270072.
- [22] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892–905, Aug. 2009.
- [23] Z. Liu, Y. Xue, L. Shen, and Z. Zhang, "Nonparametric saliency detection using kernel density estimation," in *Proc. IEEE ICIP*, Sep. 2010, pp. 253–256.
- [24] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Global contrast based salient region detection," in *Proc. IEEE CVPR*, Jun. 2011, pp. 409–416.
- [25] Y. Hu, X. Xie, W. Y. Ma, L. T. Chia, and D. Rajan, "Salient region detection using weighted feature maps based on the human visual attention model," in *Proc. Pacific Conf. Multimedia*, Nov. 2004, vol. 2, pp. 993–1000.
- [26] K. T. Park and Y. S. Moon, "Automatic extraction of salient objects using feature maps," in *Proc. IEEE ICASSP*, Apr. 2007, vol. 1, pp. 617–620.
- [27] W. Zhang, Q. M. J. Wu, G. Wang, and H. Yin, "An adaptive computational model for salient object detection," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 300–316, Jun. 2010.

- [28] J. Han, K. N. Ngan, M. Li, and H. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.
- [29] B. C. Ko and J. Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," *J. Opt. Soc. Amer. A*, vol. 23, no. 10, pp. 2462–2470, Oct. 2006.
- [30] 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.
- [31] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vision*, vol. 70, no. 2, pp. 109–131, Nov. 2006.
- [32] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Proc. IEEE ICPR*, Dec. 2008, pp. 1–4.
- [33] C. Jung, B. Kim, and C. Kim, "Automatic segmentation of salient objects using iterative reversible graph cut," in *Proc. IEEE ICME*, Jul. 2010, pp. 590–595.
- [34] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. ECCV*, Sep. 2010, pp. 366–379.
- [35] 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.
- [36] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, no. 7, pp. 1151–1163, Jul. 2002.
- [37] P. Hall and M. Wand, "On the accuracy of binned kernel density estimators," *J. Multivar. Anal.*, vol. 56, no. 2, pp. 165–184, Feb. 1996.
- [38] Y. Boykov and V. Kolmogorov, "An experimental comparison of mincut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
- [39] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, Nov. 2006.
- [40] S. Xiang, C. Pan, F. Nie, and C. Zhang, "TurboPixel segmentation using eigen-images," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 3024–3034, Nov. 2010.



Zhi Liu (M'07) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 2005.

Since 2005, he has been with the faculty of the School of Communication and Information Engineering, Shanghai University, where he currently serves as an Associate Professor and the Deputy Director of the image processing and transmission

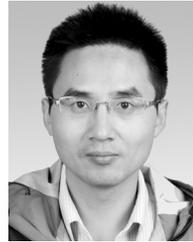
lab. His research interests include image/video segmentation, image/video retargeting, saliency model, video coding, and multimedia communication. He has authored or co-authored more than 90 refereed technical papers in international journals and conferences.

Dr. Liu served as TPC members in PCM 2010, ISPACS 2010, and IWVP 2011. He is a senior member of the Chinese Institute of Electronics.



Ran Shi received the B.E. degree from Changshu Institute of Technology, Changshu, China, in 2009. He is currently pursuing the M.E. degree at the School of Communication and Information Engineering, Shanghai University, Shanghai, China.

His research interests include salient object detection and interactive object segmentation.



Liquan Shen received the B.E. degree from Henan Polytechnic University, Henan, China, in 2001, and the M.E. and Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China, in 2005 and 2008, respectively.

Since 2008, he has been with the faculty of the School of Communication and Information Engineering, Shanghai University, where he is currently an Associate Professor. His research interests include scalable video coding, multiview video coding, high efficiency video coding (HEVC), and video codec optimization.



Yin Zhu Xue received the B.E. degree from Nanjing University of Information Science and Technology, Nanjing, China, in 2009. She is currently pursuing the M.E. degree at the School of Communication and Information Engineering, Shanghai University, Shanghai, China.

Her research interests include saliency model and image/video retargeting.



King Ngi Ngan (M'79–SM'91–F'00) received the Ph.D. degree in electrical engineering from the Loughborough University, Loughborough, U.K.

He is currently a Chair Professor at the Department of Electronic Engineering, The Chinese University of Hong Kong. He was previously a full Professor at the Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He holds honorary and visiting professorships of numerous universities in China, Australia, and South East Asia. He has published extensively including

three authored books, six edited volumes, over 300 refereed technical papers, and edited nine special issues in journals. In addition, he holds ten patents in the areas of image/video coding and communications.

Prof. Ngan was an associate editor of the *Journal on Visual Communications and Image Representation*, as well as an area editor of *EURASIP Journal of Signal Processing: Image Communication*, and he served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Journal of Applied Signal Processing*. He chaired a number of prestigious international conferences on video signal processing and communications, and he served on the advisory and technical committees of numerous professional organizations. He co-chaired the IEEE International Conference on Image Processing (ICIP) held in Hong Kong in September 2010. He is a Fellow of IET (U.K.) and IEAust (Australia), and an IEEE Distinguished Lecturer in 2006–2007.



Zhaoyang Zhang received the B.E. degree from Xi'an Jiaotong University, Xi'an, China, in 1962.

He is currently a Distinguished Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. He was the Director of the Key Laboratory of Advanced Display and System Application, Ministry of Education, China, and the Deputy Director of the Institute of China Broadcasting and Television and the Institute of China Consumer Electronics. He has published more than 200 refereed technical papers

and 10 books. His research interests include digital television, 2-D and 3-D video processing, image processing, and multimedia systems.