Saliency Tree: A Novel Saliency Detection Framework

Zhi Liu, *Member, IEEE*, Wenbin Zou, and Olivier Le Meur

***Abstract*— This paper proposes a novel saliency detection framework termed as saliency tree. For effective saliency measurement, the original image is first simplified using adaptive color quantization and region segmentation to partition the image into a set of primitive regions. Then, three measures, i.e., global contrast, spatial sparsity, and object prior are integrated with regional similarities to generate the initial regional saliency for each primitive region. Next, a saliency-directed region merging approach with dynamic scale control scheme is proposed to generate the saliency tree, in which each leaf node represents a primitive region and each non-leaf node represents a non-primitive region generated during the region merging process. Finally, by exploiting a regional center-surround scheme based node selection criterion, a systematic saliency tree analysis including salient node selection, regional saliency adjustment and selection is performed to obtain final regional saliency measures and to derive the high-quality pixel-wise saliency map. Extensive experimental results on five datasets with pixel-wise ground truths demonstrate that the proposed saliency tree model consistently outperforms the state-of-the-art saliency models.**

***Index Terms*— Saliency tree, saliency detection, saliency model, saliency map, regional saliency measure, region merging, salient node selection.**

I. Introduction

SALIENCY detection plays an important role in a variety of applications including salient object detection [1], [2], salient object segmentation [3]–[6], content-aware image/video retargeting [7]–[9], content-based image/video compression [10], [11], and content-based image retrieval [12], etc. Generally, saliency is defined as what captures human perceptual attention. Human vision system (HVS) has the ability to effortlessly identify salient objects even in a complex scene by exploiting the inherent visual attention mechanism. With the goal both to achieve a comparable saliency detection performance of HVS and to facilitate different saliency-based applications such as those mentioned above, a number of computational saliency models have been proposed in the past decades, and a recent benchmark for saliency models on saliency detection performance is reported in [13].

The early research on saliency model is motivated by simulating the visual attention mechanism of HVS, through which only the significant portion of the scene projected onto the retina is thoroughly processed by human brain for semantic understanding. Based on the biologically plausible visual attention architecture [14] and the feature integration theory [15], Itti et al. proposed a well-known saliency model [16], which first computes feature maps of luminance, color and orientation using a center-surround operator across different scales, and then performs normalization and summation to generate the saliency map. Salient regions showing high local contrast with their surrounding regions in terms of any of the three features are highlighted in the saliency map.

Since then, the center-surround scheme has been widely exploited in a variety of saliency models, due to its clear interpretation of visual attention mechanism and its concise computation form. The centre-surround scheme is implemented using a number of features including local contrasts of color, texture and shape features [17], oriented subband decomposition based energy [18], ordinal signatures of edge and color orientation histograms [19], Kullback-Leibler (KL) divergence between histograms of filter responses [20], local regression kernel based self-resemblance [21], and earth mover's distance (EMD) between the weighted histograms [22]. The selection of surrounding region is the key factor to suitably evaluate the saliency of the center pixel/region. Rather than using a fixed shape such as rectangle or circular region, the surrounding region is selected as the whole region of the blurred image in the frequency-tuned saliency model [23], and the maximum symmetric region in [24]. Besides, the center-surround differences are evaluated on the basis of segmented regions using several region-based features to generate the region-level saliency map in [25]. However, it is nontrivial to determine a suitable scale or to integrate multiple scales for surrounding regions, which can adapt well to salient objects and background with various scales and shapes for reasonable saliency evaluation.

Besides the widely exploited center-surround scheme, there are various formulations for measuring saliency based on different theories and principles such as information theory,

Manuscript received May 16, 2013; revised November 23, 2013 and February 6, 2014; accepted February 6, 2014. Date of publication February 20, 2014; date of current version March 26, 2014. This work was supported in part by the National Natural Science Foundation of China under Grant 61171144, in part by the Shanghai Natural Science Foundation under Grant 11ZR1413000, in part by the Innovation Program of Shanghai Municipal Education Commission under Grant 12ZZ086, in part by the Key (Key grant) Project of Chinese Ministry of Education under Grant 212053, and in part by a Marie Curie International Incoming Fellowship within the 7th European Community Framework Programme under Grant 299202. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. David Frakes.

Z. Liu is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with the Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes 35042, France (e-mail: liuzhisjtu@163.com).

W. Zou is with College of Information Engineering, Shenzhen University, Shenzhen, 518060, China; and he was with the National Institute of Applied Sciences of Rennes, European University of Brittany, Rennes 35708, France. (email: zouwb2008@gmail.com).

O. Le Meur is with the University of Rennes 1, Rennes 35042, France (e-mail: olemeur@irisa.fr).

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.2014.2307434

frequency domain analysis, graph theory and supervised learning. Based on information theory, the rarity represented using self-information of local image features [26], the complexity represented using local entropy [27], and the average transferring information represented using entropy rate [28] are exploited to measure saliency. Using frequency domain analysis methods, the spectral residual of the amplitude spectrum of Fourier transform [29], the phase spectrum of quaternion Fourier transform [10], and contrast sensitivity function in the frequency domain [30] are exploited to generate the saliency map. Based on graph theory, random walks on the weighted graph constructed at pixel level [31] and block level [32], and the stochastic graph model constructed on the basis of region segmentation [33] are exploited to generate saliency maps at different levels. Using supervised learning methods, a set of features including multi-scale contrast, center-surround histogram and color spatial distribution are integrated to generate the saliency map under the framework of conditional random field [34], and region feature vectors are mapped to saliency scores, which are fused across multiple levels to generate the saliency map [35]. Using support vector machine, eye tracking data is used to train the saliency model for selection of salient/non-salient pixel samples and feature extraction [36].

Recently, the global information of the image has been incorporated into saliency models with different forms. In the context-aware saliency model [37], the global uniqueness of color feature and some visual organization rules are combined with the local center-surround difference to generate the saliency map. In [38], the global color distribution represented using Gaussian mixture models (GMM), and both local and global orientation distribution are utilized to selectively generate the saliency map. In [39], GMMs are used to explicitly construct salient object/background model, and for each pixel, the ratio of posterior probability of object model to background model is calculated as the saliency measure.

Furthermore, in some saliency models [40]–[46], the image is partitioned into regions using either image segmentation methods or pixel clustering methods, and the global information is effectively incorporated at region level. Statistical models such as Gaussian model [40] and kernel density estimation based nonparametric model [41] are used to represent each region, and both color and spatial saliency measures of such statistical models are evaluated and integrated to measure the pixel's saliency. Using different formulations, global contrast and spatially weighted regional contrast [42], color compactness of over-segmented regions [43], distinctiveness and compactness of regional histograms [44], global contrast and spatial sparsity of superpixels [45], and two contrast measures for rating global uniqueness and spatial distribution of colors in the saliency filter [46] are exploited to generate saliency maps with well-defined boundaries. In the recently proposed hierarchical saliency model [47], saliency cues are calculated on three image layers with different scales of segmented regions, and then hierarchical inference is exploited to fuse them into a single saliency map.

Besides, some recent saliency models also exploit object/background priors and cues at different levels for a better saliency detection performance. For example, generic objectness measure [1] and object-level closed shape prior are effectively incorporated into saliency models presented in [48] and [49], respectively. Under the framework of low-rank matrix recovery, center prior, color prior and learnt transform prior [50] as well as region segmentation based object prior [51] are exploited for saliency detection. In the geodesic saliency model [52], background priors are exploited to formulate the saliency of patch/superpixel as the length of its shortest path to image borders. In the Bayesian saliency model [53], convex hull analysis on interest points and Laplacian sparse subspace clustering on superpixels are used as low-level and mid-level cues, respectively, to infer pixel's Bayesian saliency.

It should be noted that saliency detection performance has been progressively enhanced with the emerging saliency models, especially those recent models proposed in [40]–[53]. They can highlight salient object regions more completely with well-defined boundaries, and suppress background regions more effectively compared to previous saliency models. However, these state-of-the-art saliency models are still insufficient to effectively handle some complicated images with low contrast between objects and background, heterogeneous objects and cluttered background.

With the main motivation to improve the overall saliency detection performance and especially enhance the applicability on complicated images, we propose saliency tree as a novel saliency model in this paper. Our main contributions are fourfold. First, the proposed saliency tree model enables a hierarchical representation of saliency, which is different from the existing saliency models. Note that the recent model [47] exploits hierarchical inference for fusing multi-layer saliency cues, and takes the advantage of hierarchical saliency detection for improving the performance. The proposed model is considerably different from [47] in the complete framework of saliency tree generation and analysis, which selects the most suitable region representation by exploiting the hierarchy of tree structure, to effectively improve the saliency detection performance. Second, on the basis of our previous work [45], [51], we integrate three measures, i.e., global contrast, spatial sparsity and object prior, at region level to reasonably initialize regional saliency measures. Third, we propose a saliency-directed region merging approach with dynamic scale control scheme for saliency tree generation, which can preserve meaningful regions at different scales. Finally, we propose a systematic procedure of saliency tree analysis including regional center-surround scheme based node selection criterion, salient node selection, regional saliency adjustment and selection to generate high-quality regional saliency map and to derive the final pixel-wise saliency map. Both subjective observations and objective evaluations demonstrate that the proposed saliency tree model achieves a consistently higher saliency detection performance on five datasets compared to the state-of-the-art saliency models.

The flowchart of the proposed saliency tree model is shown in Fig. 1, and the following four sections from Section II to V describe image simplification, regional saliency measurement, saliency tree generation and saliency tree analysis, respectively. Extensive experimental results and analysis are

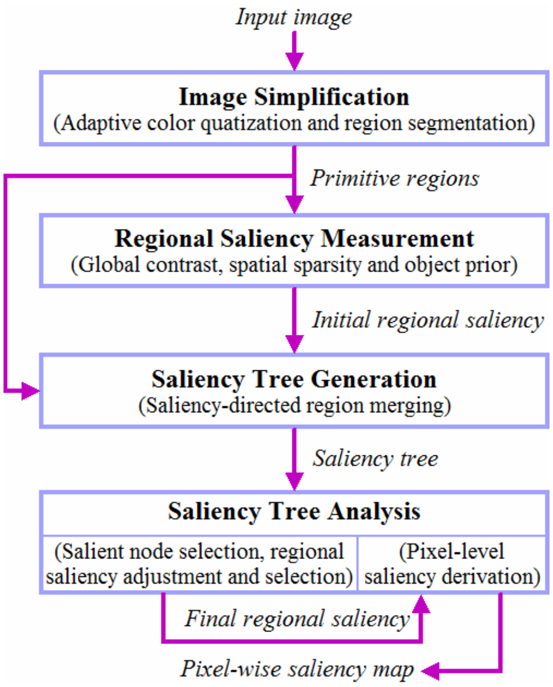

Input image
Image Simplification
(Adaptive color quatization and region segmentation)
Primitive regions
Regional Saliency Measurement
(Global contrast, spatial sparsity and object prior)
Initial regional saliency
Saliency Tree Generation
(Saliency-directed region merging)
Saliency tree
Saliency Tree Analysis
(Salient node selection, regional saliency adjustment and selection)
(Pixel-level saliency derivation)
Final regional saliency
Pixel-wise saliency map

Fig. 1. Flowchart of the proposed saliency tree model.

presented in Section VI, and conclusions are given in Section VII.

II. Image Simplification

Natural image generally contains thousands of pixels with a variety of colors. In order to effectively measure saliency, two simplification operations, i.e., adaptive color quantization and region segmentation, are performed on the original image to represent it using a reduced number of colors and regions. Before the following two simplification operations, the original color image is transformed into the *Lab* color space, in which the luminance channel and the two chrominance channels are well decorrelated.

For adaptive color quantization, each color channel is first uniformly divided into q bins, and the quantization step is defined as

$$\delta_z = \frac{z_{\max} - z_{\min}}{q} \tag{1}$$

where z denotes each channel L, a or b, and $z_{\max}$ and $z_{\min}$ denote the maximum and the minimum, of the channel z, respectively. The parameter q is set to a moderate value, 16, which is generally sufficient for color quantization of natural images. Then a color quantization table Q with $q \times q \times q$ entries is generated based on the colors of all pixels in the image. The quantized color of each entry, $\mathbf{qc}_k$, is calculated as the mean color of those pixels falling into the k^{th} entry of Q. Finally, by removing those entries with zero value, Q is updated to have a total of m (usually $m \ll q \times q \times q$) entries.

For region segmentation, we choose the gPb-owt-ucm method [54], which exploits the globalized probability of boundary (gPb) based contour detector and the oriented watershed transform (owt) to generate the real-valued ultrametric contour map (UCM). By thresholding the UCM, a set of closed boundaries are retained to form a boundary map, which can be converted into a region segmentation result.

Specifically, the thresholding operation is first performed on the UCM, which is normalized into the range of [0, 1], by increasing the threshold from 0 to 1 with an interval of 0.01, to obtain the boundary map when the corresponding region number decreases just below τ (let t denote the number of actually generated regions). Then for each region smaller than $\alpha M/t$, where M is the total number of pixels in the image and α is the coefficient for controlling removal of small regions, the pixels belonging to the weakest part of its boundary, in terms of UCM values, are set to zero in the UCM for elimination of this small region. The two parameters, τ and α are set to 200 and 0.2, respectively, to obtain an over-segmentation result with reasonable region sizes. After the above removal of small regions, the remaining regions $R_i (i = 1, \ldots, n)$ are termed as primitive regions, which are appropriate for regional saliency measurement and are used as the basis for generating the saliency tree in the following sections.

For the example image in Fig. 2(a), its UCM is shown in Fig. 2(b), in which strong boundaries are darker than weak boundaries. The primitive region segmentation result is shown in Fig. 2(c), in which each primitive region is represented using the mean color of the region.

III. Regional Saliency Measurement

Based on the image simplification result, regional similarity is measured based on regional histograms. Then initial regional saliency for each primitive region is evaluated by integrating three measures, i.e., global contrast, spatial sparsity and object prior. We observed from a variety of natural images that salient objects are generally different from background regions, and are surrounded by background regions, which usually touch image borders. Specifically, the above three measures are evaluated based on the following three aspects:

1) Salient object regions usually show contrast with background regions;

2) Spatial distribution of salient object colors is sparser than background colors, which usually scatter over the whole image;

3) Background regions generally have a higher ratio of connectivity with image borders than salient object regions.

The following five subsections will detail the whole process of regional saliency measurement.

A. Regional Similarity

For each primitive region $R_i (i = 1, \ldots, n)$, its regional histogram H_i is calculated using the quantized colors of all pixels in R_i, and then normalized to have $\sum_{k=1}^{m} H_i(k) = 1$. The regional similarity between each pair of primitive regions, R_i and R_j, is defined as

$$Sim(R_i, R_j) = Sim_c(R_i, R_j) \cdot Sim_d(R_i, R_j) \tag{2}$$

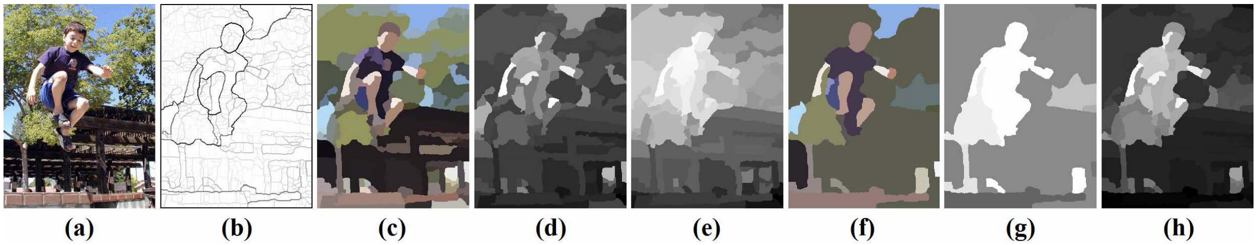


Fig. 2. Illustration of region segmentation and regional saliency measurement. (a) original image; (b) ultrametric contour map; (c) primitive region segmentation result; (d) global contrast map; (e) spatial sparsity map; (f) coarse region segmentation result; (g) object prior map; (h) initial regional saliency map.

The color similarity between R_i and R_j is defined based on the chi-square distance between H_i and H_j as follows:

$$Sim_c(R_i, R_j) = \exp\left(-\frac{1}{2}\sum_{k=1}^{m}\frac{\left[H_i(k) - H_j(k)\right]^2}{H_i(k) + H_j(k)}\right) \tag{3}$$

The spatial similarity between R_i and R_j is defined as

$$Sim_d(R_i, R_j) = 1 - \frac{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|_2}{d} \tag{4}$$

where d denotes the diagonal length of the image, and the spatial center position of R_i is defined as

$$\boldsymbol{\mu}_i = \frac{\sum_{p \in R_i} \mathbf{x}_p}{|R_i|} \tag{5}$$

where $\mathbf{x}_p$ denotes the spatial coordinates of each pixel p, and $|R_i|$ denotes the number of pixels in R_i.

Both $Sim_c(R_i, R_j)$ and $Sim_d(R_i, R_j)$ fall into the normalized range of [0, 1]. $Sim(R_i, R_j)$ is evaluated higher when the color distributions of R_i and R_j are similar and the spatial distance between them is shorter.

B. Global Contrast

The global contrast of each primitive region R_i is measured using the weighted color differences with all the other regions as follows:

$$GC(R_i) = \sum\nolimits_{j=1}^{n} \left|R_j\right| \cdot Sim_d(R_i, R_j) \cdot \left\|\mathbf{mc}_i - \mathbf{mc}_j\right\|_2 \tag{6}$$

where $\mathbf{mc}_i$ (resp. $\mathbf{mc}_j$) denotes the mean color of R_i (resp. R_j). The weight $\left|R_j\right| \cdot Sim_d(R_i, R_j)$ indicates that those regions, which are larger and spatially closer to R_i, have a relatively larger contribution to the evaluation of global contrast of R_i. Then the normalized global contrast measure for R_i is calculated as follows:

$$NGC(R_i) = \frac{GC(R_i) - GC_{\min}}{GC_{\max} - GC_{\min}} \tag{7}$$

where $GC_{\max}$ and $GC_{\min}$ are the maximum and the minimum, respectively, in the global contrast measures of all primitive regions.

Based on a reasonable assumption that region pairs with a high regional similarity should be evaluated with similar values on the global contrast measures, the regional similarities are used to refine the normalized global contrast measures as follows:

$$RGC(R_i) = \frac{\sum_{j=1}^{n} Sim(R_i, R_j) \cdot NGC(R_j)}{\sum_{j=1}^{n} Sim(R_i, R_j)} \tag{8}$$

C. Spatial Sparsity

For each primitive region R_i, the spatial spread of its color distribution is defined as follows:

$$SS(R_i) = \frac{\sum_{j=1}^{n} Sim(R_i, R_j) \cdot D(R_j)}{\sum_{j=1}^{n} Sim(R_i, R_j)} \tag{9}$$

where $D(R_j)$ denotes the Euclidean spatial distance from the center position of R_j to the image center position. Then an inverse normalization operation is performed on the spatial spread measures to obtain the normalized spatial sparsity measures as follows:

$$NSS(R_i) = \frac{SS_{\max} - SS(R_i)}{SS_{\max} - SS_{\min}} \tag{10}$$

where $SS_{\max}$ and $SS_{\min}$ are the maximum and the minimum, respectively, in the spatial spread measures of all primitive regions. Similarly, the normalized spatial sparsity measures are refined as follows:

$$RSS(R_i) = \frac{\sum_{j=1}^{n} Sim(R_i, R_j) \cdot NSS(R_j)}{\sum_{j=1}^{n} Sim(R_i, R_j)} \tag{11}$$

D. Object Prior

The connectivity ratio between each region and image borders can be used to indicate the prior probability belonging to a salient object, because salient objects usually do not connect with image borders or connect with image borders less than background regions in a variety of images. Following our previous work [51], it is suitable to evaluate the object prior on the basis of a coarse region segmentation, so as to obtain more uniform object prior values for those homogenous background regions, which connect with image borders and are less partitioned in the coarse segmentation result. For this purpose, the UCM is thresholded using a relatively higher value, 0.25, to obtain a coarse segmentation result with n_c regions, and the object prior for each region $\Re_j (j = 1, \ldots, n_c)$ is defined as follows:

$$OP(\Re_j) = \exp\left(-\lambda\frac{\left|\Re_j \cap B\right|}{\partial\Re_j}\right) \tag{12}$$

where B denotes the image borders, and $\partial\Re_j$ denotes the perimeter of region $\Re_j$. The coefficient λ is set to 2.0 for a moderate attenuation effect on object priors of those regions touching image borders. Then for each primitive region $R_i (i = 1, \ldots, n)$, its object prior is assigned as follows:

$$OP(R_i) = OP(\Re_j), \ \forall R_i \subseteq \Re_j \tag{13}$$

E. Initial Regional Saliency

By integrating the aforementioned three measures using the multiplication operation, the initial regional saliency measure for each primitive region R_i is defined as follows:

$$S_I(R_i) = RGC(R_i) \cdot RSS(R_i) \cdot OP(R_i) \tag{14}$$

The initial regional saliency measures of all primitive regions are normalized into the range of [0, 1] for the latter use in the saliency tree generation. For the example image in Fig. 2(a), which has a human object with heterogeneous regions and a cluttered background, the regional global contrast map and spatial sparsity map are shown in Fig. 2(d) and (e), respectively. On the basis of coarse region segmentation in Fig. 2(f), the generated object prior map is shown in Fig. 2(g). Note that for such a complicated image in Fig. 2(a), either Fig. 2(d) or (e) reasonably but partly highlights salient object regions and suppresses background regions to some extent. Fig. 2(g) moderately suppresses a large part of background regions, but those background regions without touching image borders cannot be suppressed.

By integrating the three maps in Fig. 2(d), (e), and (g), the initial regional saliency map shown in Fig. 2(h) can highlight salient object regions and suppress background regions more effectively than any of the three maps, but it is still insufficient to suppress some background regions. To effectively improve the saliency detection performance on such complicated images, we propose the following saliency tree generation in Sec. IV and saliency tree analysis in Sec. V.

For reference, the saliency maps generated using the state-of-the-art saliency models for this image are shown from column (c) to (m) in the bottom row of Fig. 7. We can observe that such a complicated image is difficult for the state-of-the-art saliency models to generate high-quality saliency maps, while the above initial regional saliency map visually outperforms most saliency maps generated using the state-of-the-art saliency models. Besides, in Sec. VI-B, a performance analysis on five datasets is used to objectively evaluate the contribution of each measure, i.e., global contrast, spatial sparsity and object prior, to the initial regional saliency.

IV. Saliency Tree Generation

Starting from the primitive regions with their initial regional saliency measures, a saliency-directed region merging approach is proposed to generate the saliency tree, which is a binary partition tree [55] with saliency measures. Specifically, the region merging sequence is recorded by exploiting the structure of binary partition tree, in which each node is assigned with regional saliency measure. Each primitive region is represented by a leaf node in the saliency tree, and each non-primitive region, which is generated during the region merging process, is represented by a non-leaf node in the saliency tree. The following two subsections first describe the merging criterion and merging order exploited in the region merging process, and then detail the saliency-directed region merging approach for saliency tree generation.

A. Merging Criterion and Merging Order

In order to direct the region merging process, the merging criterion for each pair of adjacent regions, R_i and R_j, is evaluated based on color similarity and saliency similarity between them as follows:

$$Mrg(R_i, R_j) = Sim_c(R_i, R_j) \cdot Sim_s(R_i, R_j) \tag{15}$$

where the saliency similarity is defined as

$$Sim_s(R_i, R_j) = 1 - \left| S_I(R_i) - S_I(R_j) \right| \tag{16}$$

It is obvious that the merging criterion $Mrg(R_i, R_j)$ achieves a higher value when R_i and R_j show similar color distributions and similar regional saliency measures. Since region merging is always performed on adjacent regions, the merging criterion for each non-adjacent region pair is set to zero, for clarity of the following description.

Based on the above defined merging criterion, the region merging process is iteratively performed based on the determined merging order. Specifically, at each merging step, one pair or multiple pairs of adjacent regions are selected to merge by checking the following two conditions (a) and (b). Without loss of generality, assume there are a total of t regions at the beginning of the current merging step.

(a) The out-of-scale regions, which are too small in view of the region number at the current merging step, are first searched to merge with a higher priority. The set of out-of-scale regions is denoted as $\Omega = \{R_i\}, \forall |R_i| \leq \beta M/t$, where the coefficient β is set to a moderate value, 0.2, for selection of small regions. In case that Ω is not empty, for each region R_i in Ω, its most similar region R_j is selected to constitute a region pair (R_i, R_j) for merging, i.e.,

$$R_j = \arg\max_{k=1\ldots t} Mrg(R_i, R_k), \ \forall R_i \in \Omega \tag{17}$$

The condition (a) is exploited to dynamically control the region scale dependent on the current region number t. If more than one pair of regions are selected based on the condition (a), these region pairs are merged simultaneously. However, in case that Ω is empty (usually at the latter merging steps), the condition (b) will be used to select one region pair to merge.

(b) The region pair (R_i, R_j) with the highest merging criterion, which is actually the general condition to determine the merging order in conventional region merging algorithms, is selected as follows:

$$(R_i, R_j) = \arg\max_{k_1=1\ldots t,\, k_2=1\ldots t} Mrg(R_{k_1}, R_{k_2}) \tag{18}$$

When a region pair (R_i, R_j) is used to merge into a new region R_k, its regional histogram H_k and initial regional saliency measure $S_I(R_k)$ are calculated as follows:

$$H_k = \frac{|R_i| \cdot H_i + \left|R_j\right| \cdot H_j}{|R_i| + \left|R_j\right|} \tag{19}$$

$$S_I(R_k) = \frac{|R_i| \cdot S_I(R_i) + \left|R_j\right| \cdot S_I(R_j)}{|R_i| + \left|R_j\right|} \tag{20}$$

where $|R_i| \cdot S_I(R_i)$ is termed as the saliency gross in R_i.

Algorithm 1 Pseudo Code of Saliency-Directed Region Merging Approach

```
Input: Primitive regions $R_i (i = 1,...n)$ and their initial regional saliency measures $S_I(R_i)$.
Output: Saliency tree with $2n-1$ nodes.
Begin
  Calculate merging criteria for all pairs of adjacent primitive regions using Eqs. (15)-(16).
  Initialize the saliency tree with $n$ leaf nodes, which represent the $n$ primitive regions.
  Set the current region number $t \leftarrow n$.
  While $t > 1$
    Find $s$ pair(s) of adjacent regions to merge by checking the condition (a) in Eq. (17).
    If $s = 0$ Then
      Find one pair of adjacent regions to merge by checking the condition (b) in Eq. (18), and set $s \leftarrow 1$.
    End
    Merge the above found region pairs to generate a total of $s$ new region(s).
    For $k \leftarrow 1$ to $s$
      Denote the region pair for generating each new region $R_k$ as $(R_i, R_j)$.
      Calculate for $R_k$ its regional histogram $H_k$ and initial regional saliency measure $S_I(R_k)$ using Eq. (19) and Eq. (20), respectively.
      Calculate the merging criteria between $R_k$ and its adjacent regions using Eqs. (15)-(16).
      Add a non-leaf node representing $R_k$ into the saliency tree. The left/right child node of this non-leaf node represents $R_i$ / $R_j$.
    End
    Decrease $t \leftarrow t - s$.
  End
End
```

B. Saliency-Directed Region Merging

Based on the above defined merging criterion and merging order, the proposed saliency-directed region merging approach for saliency tree generation is summarized in Algorithm 1. Starting from n primitive regions, the total times of merging two regions into a new region is $n-1$ during the whole region merging process. Therefore, in the saliency tree, there are n leaf nodes representing primitive regions $R_i(i = 1, \ldots, n)$ and $n - 1$ non-leaf nodes representing during the region merging process the generated regions $R_i(i = n + 1, \ldots, 2n - 1)$, which are termed as non-primitive regions. The root node in the saliency tree represents the complete image region.

For the example in Fig. 2, the saliency tree generated using the proposed saliency-directed region merging approach is shown in Fig. 3(a), which only shows the nodes at the top 8 levels and some meaningful regions generated during the region merging process for a clear display.

V. Saliency Tree Analysis

Saliency tree provides for an image a hierarchical saliency representation, which can be fully exploited to generate high-quality regional saliency map and pixel-wise saliency map. In the following, a systematic saliency tree analysis including the definition of node selection criterion, salient node selection, regional saliency adjustment and selection, and pixel-wise saliency map derivation, will be described in the following four subsections.

A. Node Selection Criterion

Node selection criterion is defined based on a regional center-surround scheme, in which the region with certain saliency gross and higher saliency difference from its regional surround is considered as more salient. For each region $R_i(i = 1, \ldots, 2n - 1)$ represented by each node N_i in the saliency tree, its regional surround C_i is defined as the set of primitive regions adjacent to R_i. An example of the region R_a and its regional surround C_a is shown in Fig. 3(b). The saliency measure of each regional surround C_i is defined as follows:

$$S_C(C_i) = \frac{\sum_{R_j \in C_i} \log\left(\left|R_j\right|\right) \cdot S_I(R_j)}{\sum_{R_j \in C_i} \log\left(\left|R_j\right|\right)} \tag{21}$$

where R_j is each primitive region covered in C_i. The logarithm of region area is used as the weight to reasonably attenuate the contribution of large-sized regions, which have a higher percentage of pixels far away from the boundary of R_i, and to make the contribution comparable among the surrounding primitive regions with variable areas.

Based on the regional center-surround scheme, the node selection criterion for the node N_i representing R_i is then defined as follows:

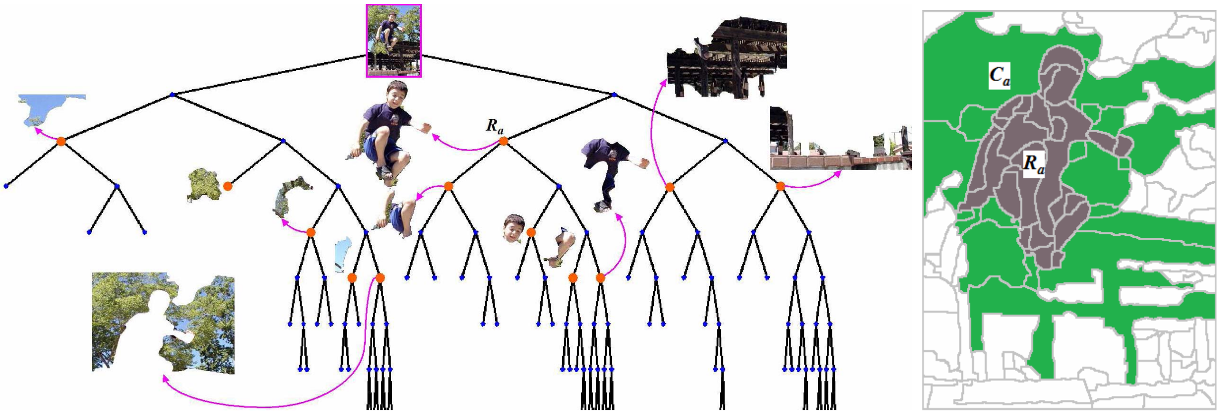
$$SC(N_i) = [S_I(R_i) - S_C(C_i)] \cdot |R_i| \cdot S_I(R_i) \tag{22}$$

where the former term represents the saliency difference between R_i and C_i, and the latter term $|R_i| \cdot S_I(R_i)$ representing the saliency gross in R_i is introduced to assign a higher node selection criterion for reasonable-sized regions with a higher saliency difference from their surrounds.

B. Salient Node Selection

From the saliency tree, in which all nodes are now assigned with node selection criteria, different sets of salient nodes are selected at a series of saliency visibility levels $V_\ell(\ell = 1, \ldots, \zeta)$, which determine the minimum saliency gross contained in the correspondingly selected regions. In our implementation, the number of total levels ζ is set to 10, and the saliency visibility levels are set from 1% to 10%, with an interval of 1%, of the total saliency gross of the image, i.e., $\sum_{i=1}^{n} |R_i| \cdot S_I(R_i)$. The purpose of salient node selection is to preserve a set of regions, which are considered as salient and the most representative at a certain saliency visibility level, and will be used for regional saliency adjustment and selection in Sec. V-C.

The proposed salient node selection procedure is summarized in Algorithm 2. Using the salient node



(a) (b)

Fig. 3. (a) Example of saliency tree; (b) (better viewed in color) Illustration of regional surround.

Algorithm 2 Pseudo Code of Salient Node Selection Procedure

```
Input: Saliency tree with node selection criteria.
Output: A set of region selection results.
Begin
  Initialize the saliency visibility levels $V_\ell (\ell = 1,...,\zeta)$.
  For $\ell \leftarrow 1$ to $\zeta$
    Set all nodes in the saliency tree to active status.
    Set the accumulated saliency gross $ASG_\ell \leftarrow 0$.
    Set the node selection list $\Phi_\ell \leftarrow \varnothing$.
    Repeat
      Select the active node with the highest node selection criterion, $N_k$, from the saliency tree.
      Add the saliency gross of region $R_k$ represented by $N_k$, i.e., $|R_k| \cdot S_I(R_k)$, into $ASG_\ell$.
      Set $N_k$ to inactive status.
      If $|R_k| \cdot S_I(R_k) \geq V_\ell$ Then
        Mark $R_k$ as salient at saliency visibility level $V_\ell$.
        Add $N_k$ into $\Phi_\ell$.
        Set all descendant nodes of $N_k$ to inactive status due that the regions represented by these descendant nodes are completely included in $R_k$.
        Set all ascendant nodes of $N_k$ to inactive status due that the regions represented by these ascendant nodes completely contain $R_k$.
      End
    Until all nodes are inactive or $ASG_\ell$ is not less than the total saliency gross of the image.
    If $\ell = 1$ or $\Phi_\ell \neq \Phi_{\ell-1} (\ell > 1)$ Then
      Output the regions represented by nodes in $\Phi_\ell$ as the region selection result $\Gamma_\ell$.
    End
  End
End
```

selection procedure on the saliency tree in Fig. 3(a), the region selection results are output at seven saliency visibility levels, i.e., V_1~V_6 and V_{10}, and shown in the top row of Fig. 4. We can observe that the region, which represents the salient object more completely, is selected at different saliency visibility levels.

C. Regional Saliency Adjustment and Selection

For each region selection result Γ_ℓ, the remaining areas (white areas in the top row of Fig. 4) uncovered by any region in Γ_ℓ are labeled using the connected component analysis to generate one or multiple regions, which constitute a complementary region set $\overline{\Gamma}_\ell$. Then the primitive regions, which are covered by each region in Γ_ℓ (resp. $\overline{\Gamma}_\ell$), constitute the region set Ψ_ℓ (resp. $\overline{\Psi}_\ell$). The two sets, Ψ_ℓ and $\overline{\Psi}_\ell$, are complementary to each other on the basis of primitive regions.

The regions in Γ_ℓ and $\overline{\Gamma}_\ell$ constitute a partition of the image at each saliency visibility level V_ℓ. As shown in the top row of Fig. 4 and Fig. 3(a), salient object regions as well as other meaningful background regions, which are generated in the saliency tree, can be more completely preserved in such a partition. Therefore, we can exploit the object prior evaluated on the basis of regions in Γ_ℓ and $\overline{\Gamma}_\ell$ to adjust the initial regional saliency measures, in order to highlight salient object regions and suppress background regions more effectively.

Specifically, for each region $\Re_k \in \Gamma_\ell \cup \overline{\Gamma}_\ell$, its object prior $OP_\ell(\Re_k)$ is calculated using Eq. (12). Then for each primitive region $R_i \in \Psi_\ell \cup \overline{\Psi}_\ell$, its object prior is assigned as follows:

$$OP_\ell(R_i) = OP_\ell(\Re_k),\ \forall R_i \subseteq \Re_k \ and\ \Re_k \in \Gamma_\ell \cup \overline{\Gamma}_\ell \qquad (23)$$

and then its regional saliency measure at the saliency visibility level V_ℓ is adjusted as follows:

$$AS_\ell(R_i) = OP_\ell(R_i) \cdot S_I(R_i) \qquad (24)$$

From a set of the adjusted regional saliency measures at different saliency visibility levels, the optimal set is selected as the one that can maximize the saliency difference between regions in Ψ_ℓ and regions in $\overline{\Psi}_\ell$, since such an adjustment of regional saliency measures shows that the corresponding region selection result is more confident and rational for regional saliency measurement. Specifically, the optimal set of the adjusted regional saliency measures, $AS_{\ell *}$, is selected

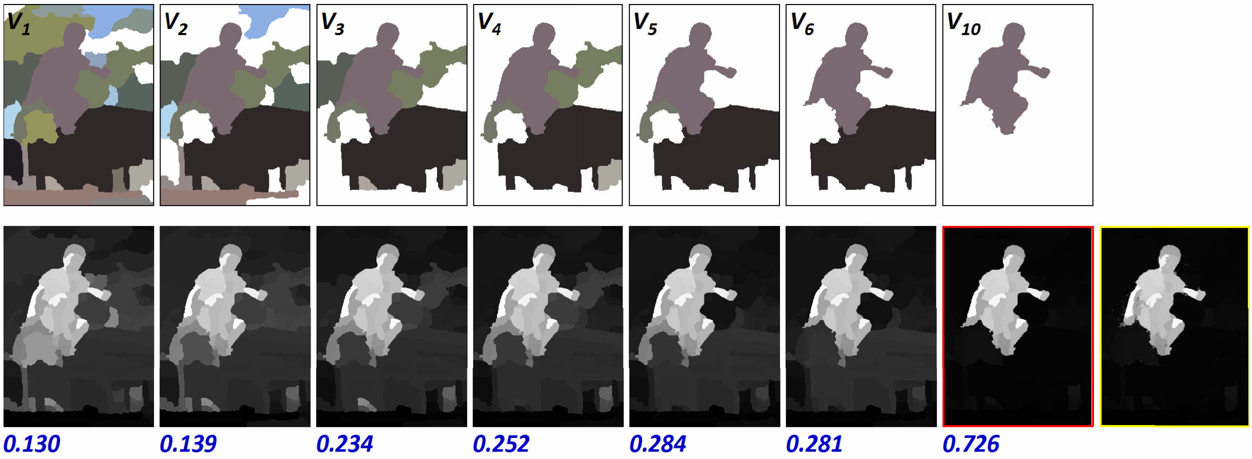


Fig. 4. (better viewed in color) Illustration of regional saliency adjustment and selection, and pixel-wise saliency map derivation. Top row: seven region selection results (each selected region is shown using the region's mean color, and the unselected regions are shown using white areas); bottom row: regional saliency maps with saliency difference measures, final regional saliency map (marked with the red border) and pixel-wise saliency map (marked with the yellow border).

by maximizing the following saliency difference measure:

$$AS_{\ell *} = \arg\max_{\ell} \left[\frac{\sum\limits_{R_i \in \Psi_\ell} |R_i| \cdot AS_\ell(R_i)}{\sum\limits_{R_i \in \Psi_\ell} |R_i|} - \frac{\sum\limits_{R_j \in \overline{\Psi}_\ell} |R_j| \cdot AS_\ell(R_j)}{\sum\limits_{R_j \in \overline{\Psi}_\ell} |R_j|} \right] \tag{25}$$

and is used as the final regional saliency measures, i.e., $S_F(R_i) = AS_{\ell *}(R_i)$ for each primitive region R_i.

For the region selection results shown in the top row of Fig. 4, the correspondingly adjusted regional saliency maps with the saliency difference measures are shown in the bottom row of Fig. 4. The adjusted regional saliency map at the saliency visibility level V_{10} is selected as the final regional saliency map, in which the salient object is highlighted and background regions are suppressed more effectively.

D. Pixel-Wise Saliency Map Derivation

Finally, a pixel-wise saliency map is derived based on final regional saliency measures of primitive regions. For each pixel $p \in R_i$, its local neighborhood Θ_p includes the primitive region R_i and the adjacent primitive regions of R_i. The saliency measure of the pixel p is defined as the weighted sum of final regional saliency measures of its neighboring primitive regions, i.e.,

$$S_P(p) = \frac{\sum_{R_j \in \Theta_p} \omega_j \cdot S_F(R_j)}{\sum_{R_j \in \Theta_p} \omega_j} \tag{26}$$

where the weight ω_j is defined as follows:

$$\omega_j = \begin{cases} H_j(b_p), \; if \; R_j = R_i \\ H_j(b_p) \cdot \exp\left(-\left\|\mathbf{x}_p - \mu_j\right\|_2 / \left\|\mathbf{x}_p - \mu_i\right\|_2\right), \; otherwise \end{cases} \tag{27}$$

where b_p denotes the entry number for the quantized color of the pixel p in the color quantization table Q. Using Eq. (27), a higher weight is given to the primitive region, which is closer to p and has a higher probability of the pixel's color in its regional histogram. It is reasonable that those primitive regions showing a higher color similarity with p and a shorter distance to p have a higher contribution to the saliency of p.

As shown in the rightmost column of Fig. 4, the derived pixel-wise saliency map better highlights the complete salient object region with well-defined boundaries, which are more natural and smoother compared to the final regional saliency map. The pixel-wise saliency map is also shown in the rightmost column of the bottom row in Fig. 7, for a visual comparison with saliency maps generated using the state-of-the-art saliency models, and we can see that the quality of our pixel-wise saliency map is better than other saliency maps.

VI. Experimental Results

To evaluate the performance of the proposed saliency tree (ST) model, we performed extensive experiments on the following five datasets and made a comparison with eleven state-of-the-art saliency models. We used all datasets in the benchmark [13] on saliency models, i.e., ASD, MSRA, SED and SOD, and one recently introduced dataset, PASCAL-1500 [51]. We choose the top five models that achieve the best performance in the benchmark [13], i.e., region contrast (RC) [42], kernel density estimation (KDE) [41], context and shape prior (CS) [49], fusion of saliency and generic objectness (SVO) [48], and context-aware (CA) [37] model, and six recently proposed saliency models including low-rank matrix recovery (LR) [50], saliency filter (SF) [46], regional histogram (RH) [44], Bayesian saliency (BS) [53], segmentation driven low-rank matrix recovery (SLR) [51], and hierarchical saliency (HS) [47]. Note that we use more informative abbreviations, KDE, CS and CA to replace the corresponding abbreviations, LiuICIP, CBsal and Goferman, respectively, used in the benchmark [13]. We used the executables or source codes with default parameter settings provided by the authors for the eleven saliency models. For a fair comparison, all saliency maps generated using different saliency models are normalized into the same range of [0, 255] with the full resolution of original images. The results of the proposed ST model are available at http://people.irisa.fr/Olivier.Le_Meur/shivpro/index.html.

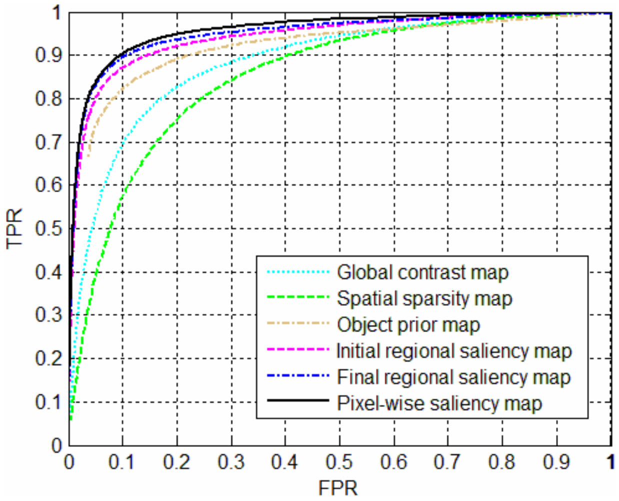


Fig. 5. (better viewed in color) ROC curves generated using global contrast maps, spatial sparsity maps, object prior maps, initial regional saliency maps, final regional saliency maps and pixel-wise saliency maps of MSRA dataset.

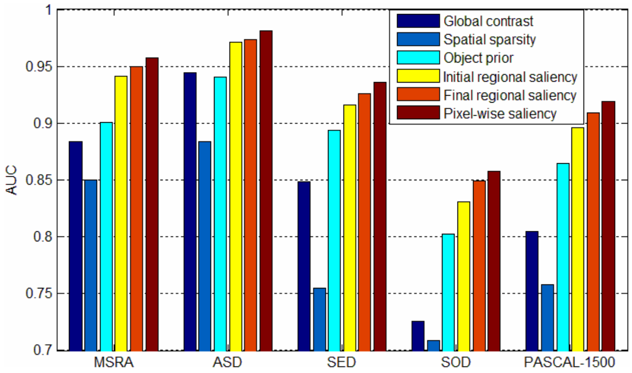


Fig. 6. AUC values achieved using global contrast maps, spatial sparsity maps, object prior maps, initial regional saliency maps, final regional saliency maps and pixel-wise saliency maps of all the five datasets.

In the following, Sec. VI-A introduces the five datasets, and Sec. VI-B analyzes the performance and contribution of different parts in the proposed ST model. The comparison of saliency detection performance with the eleven state-of-the-art saliency models, both subjectively and objectively, are presented in Sec. VI-C and VI-D, respectively. Some failure cases are analyzed in Sec. VI-E, and the complexity issue of ST model is discussed in Sec. VI-F.

A. Datasets

The five datasets used in the following experiments are described as follows:

1) **MSRA** [34] dataset contains 5,000 images from the image set B of Microsoft Research Asia salient object database. The pixel-wise binary masks of salient objects [35] are provided as the ground truths. There is a large variation among images including natural scenes, animals, indoor, outdoor, etc.

2) **ASD** [23] dataset contains 1,000 images selected from the above MSRA dataset with pixel-wise binary ground truths for salient objects. Note that ASD is the most commonly used dataset for evaluation of saliency detection performance in the recent years, but the images in this dataset are relatively simpler than the other four datasets.

3) **SED** [56] dataset contains 100 images with one salient object and the other 100 images with two salient objects. Pixel-wise ground truth annotations for salient objects in all 200 images are provided.

4) **SOD** [57] dataset contains 300 images from the Berkeley segmentation dataset (BSD) [58], for which salient object boundaries are marked by seven users. A unique binary ground truth for each image is generated using the marked boundaries which receive a majority of user votes. This dataset contains many images with different natural scenes making it challenging for saliency detection.

5) **PASCAL-1500** [51] dataset contains 1500 real-world images from PASCAL VOC 2012 segmentation challenging [59], in which only images intuitively deemed to have salient objects are selected. The binary ground truths for evaluation of saliency detection performance are adapted from the pixel-wise annotated segmentation ground truths in PASCAL VOC, by labeling object pixels as “1” and other pixels as “0”. In PASCAL-1500, many images contain multiple objects with various locations and scales, and/or highly cluttered background, which make this dataset also challenging for saliency detection.

B. Performance Analysis

The proposed ST model first generates global contrast map, spatial sparsity map, object prior map and initial regional saliency map in Sec. III, and then generates final regional saliency map and pixel-wise saliency map in Sec. V. In order to objectively evaluate the contribution of different parts in the proposed ST model to the saliency detection performance, we adopted the commonly used receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) and presents a robust evaluation of saliency detection performance. Specifically, the above mentioned six classes of maps generated using the ST model are first normalized into the same range of [0, 255]. Then thresholding operations using a series of fixed integers from 0 to 255 are performed on each map to obtain 256 binary salient object masks, and a set of TPR and FPR values are calculated by comparing to the corresponding binary ground truth. Finally, for each class of map, at each threshold, the TPR/FPR values of all images in the dataset are averaged, and the ROC curve for each class of map plots the 256 average TPR values against the 256 average FPR values.

Fig. 5 only shows the ROC curves on the largest dataset, i.e., MSRA, due to the page limit. As shown in Fig. 5, the ROC curve for initial regional saliency map is higher than the three ROC curves for global contrast map, spatial sparsity map and object prior map. This demonstrates the complementary effect of global contrast, spatial sparsity and object prior for a reasonable estimate of initial regional saliency. Furthermore, compared to the ROC curve for initial regional saliency map, the ROC curve for final regional saliency map is elevated and the ROC curve for pixel-wise saliency map is further elevated. This demonstrates the contribution of saliency tree analysis

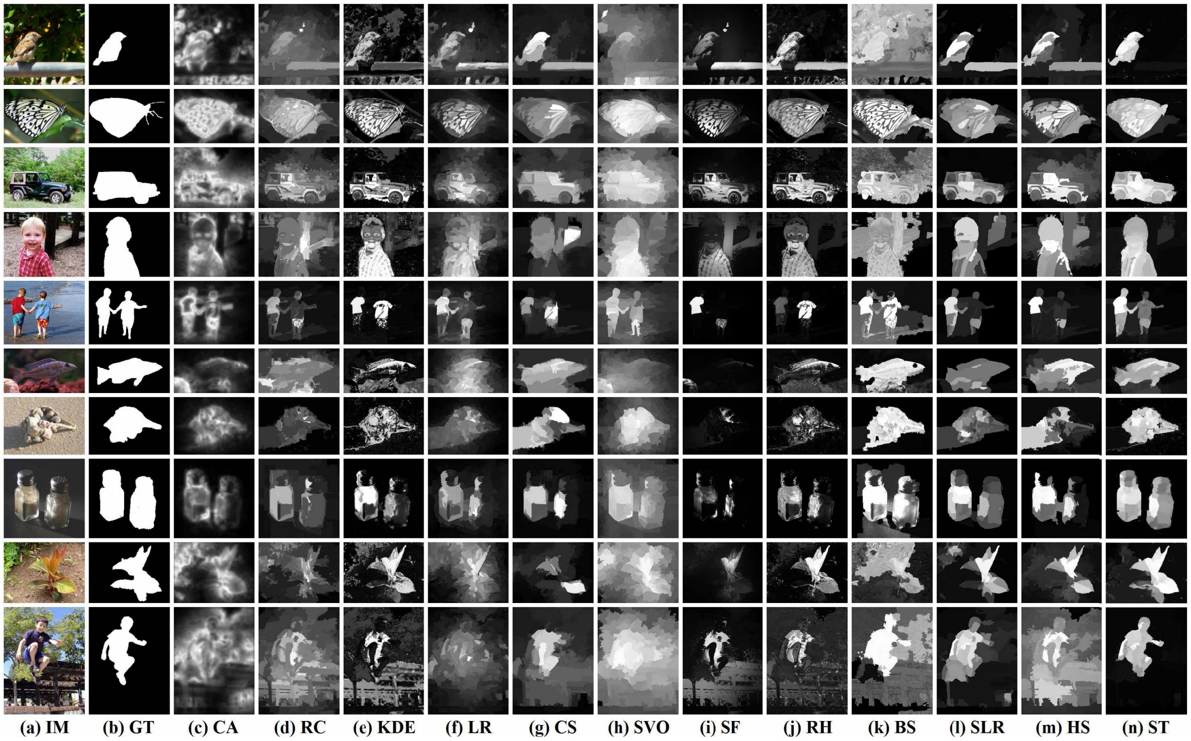


Fig. 7. Examples of saliency detection on MSRA dataset. (a) images (IM), (b) ground truths (GT) and (c)–(n) saliency maps generated using different models.

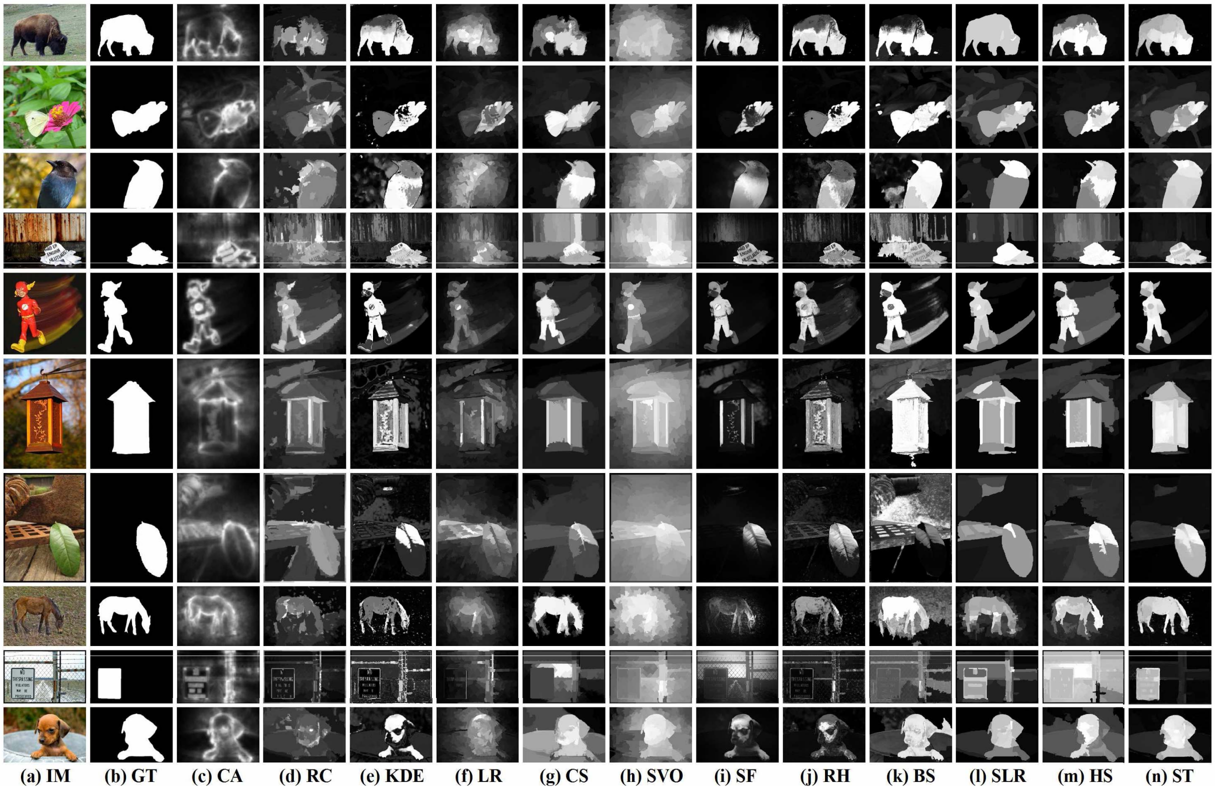


Fig. 8. Examples of saliency detection on ASD dataset. (a) images (IM), (b) ground truths (GT) and (c)–(n) saliency maps generated using different models.

for improving the saliency detection performance. We also observed similar trends of such ROC curves on the other four datasets.

For a more intuitive evaluation, we calculated the area under each ROC curve (AUC) on all the five datasets as a quantitative metric. As shown in Fig. 6, the AUC values clearly

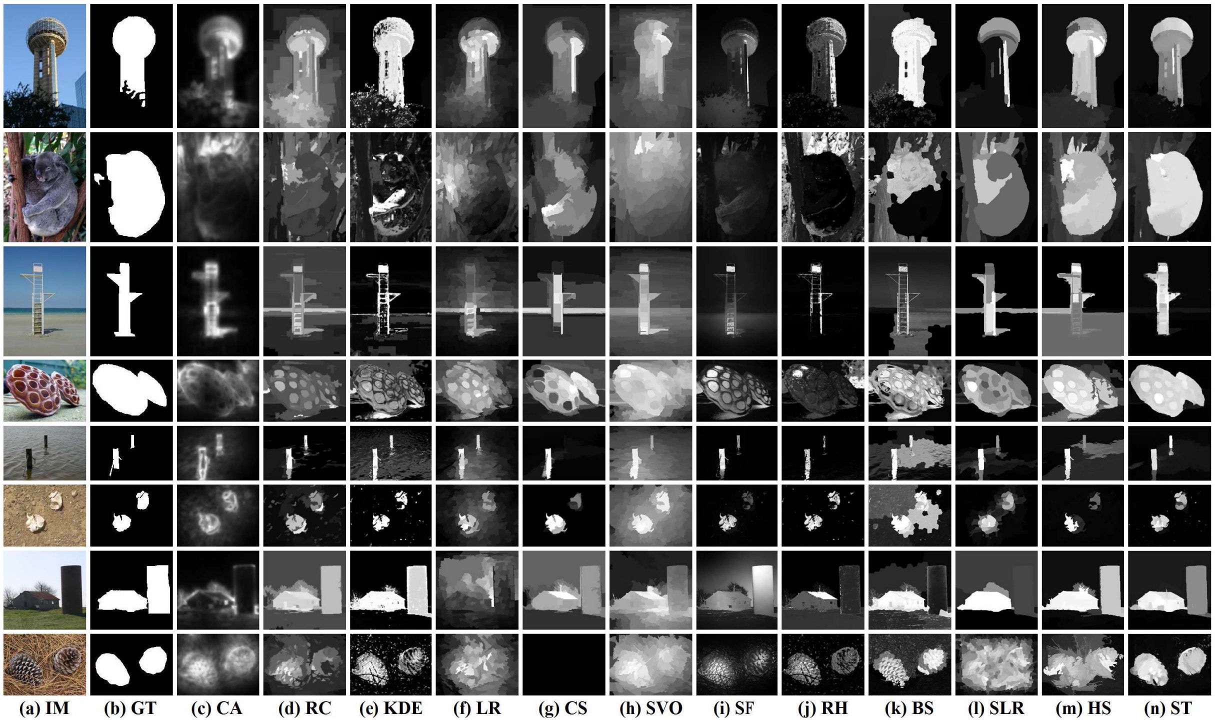


Fig. 9. Examples of saliency detection on SED dataset. (a) images (IM), (b) ground truths (GT) and (c)–(n) saliency maps generated using different models.

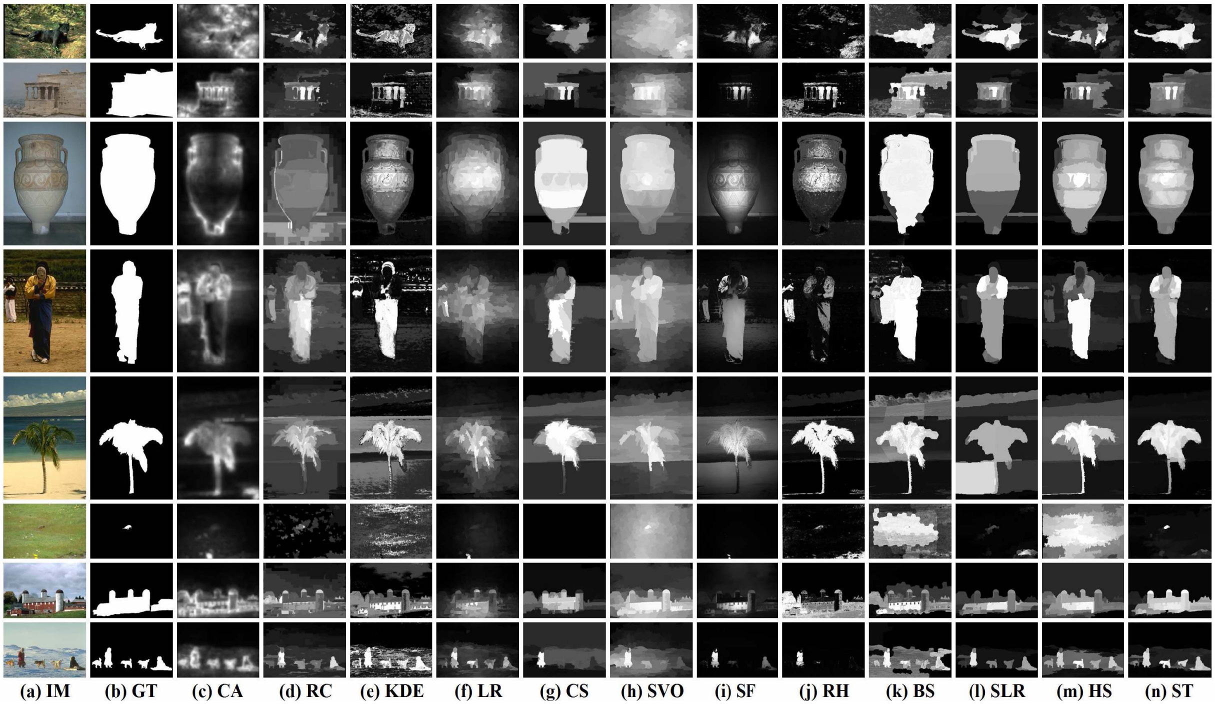


Fig. 10. Examples of saliency detection on SOD dataset. (a) images (IM), (b) ground truths (GT) and (c)–(n) saliency maps generated using different models.

demonstrate the effectiveness of initial regional saliency measurement and the contribution of saliency tree analysis to improve the saliency detection performance on all the five datasets.

C. Subjective Evaluation

Some saliency maps generated using the proposed ST model and the eleven state-of-the-art saliency models on the five datasets are shown in Figs. 7–11 for a subjective comparison. We can observe that most saliency models can effectively handle images with relatively simple background and homogenous objects, such as the top two examples in Fig. 8, to generate high-quality saliency maps. However, for some complicated images containing heterogeneous objects (such as human objects, vehicles in Figs. 7 and 11, and buildings in Fig. 10), showing a low contrast between objects and background (such as row 6 and 7 in Fig. 7, the bottom two

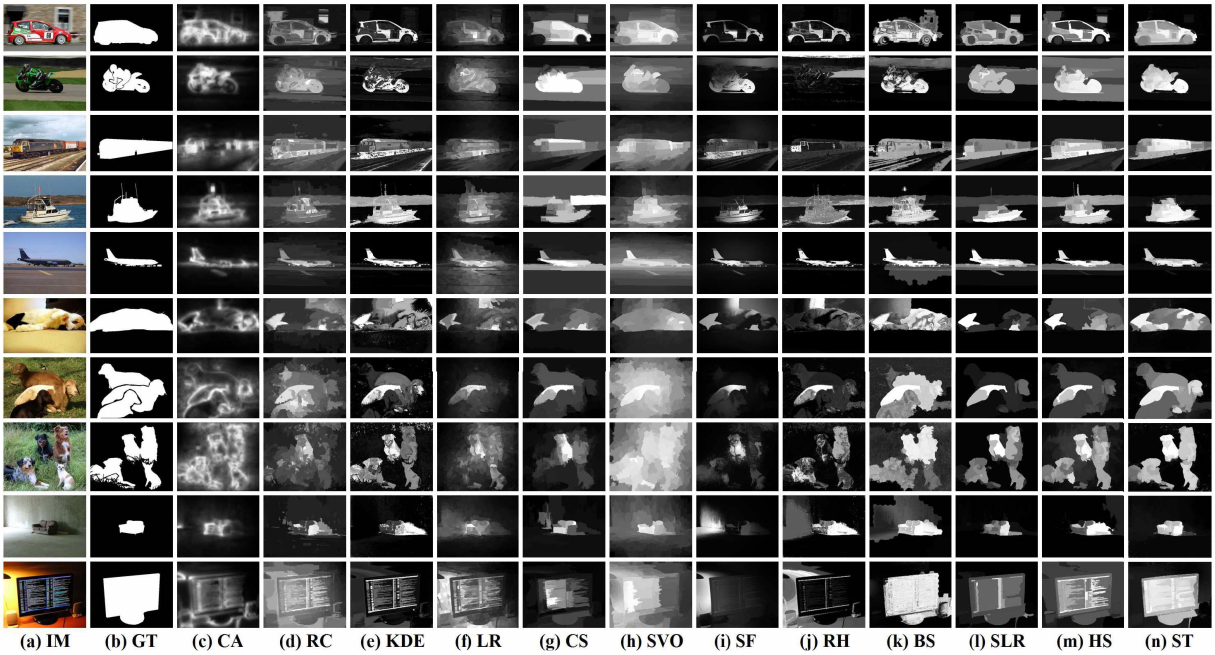


Fig. 11. Examples of saliency detection on PASCAL-1500 dataset. (a) images (IM), (b) ground truths (GT) and (c)–(n) saliency maps generated using different models.

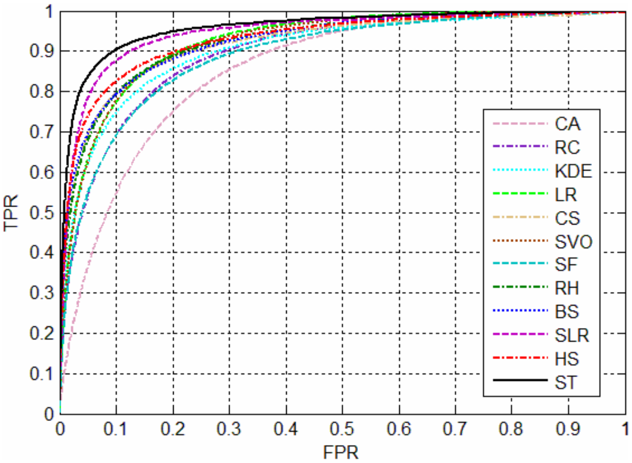


Fig. 12. (better viewed in color) ROC curves of different saliency models on MSRA dataset.

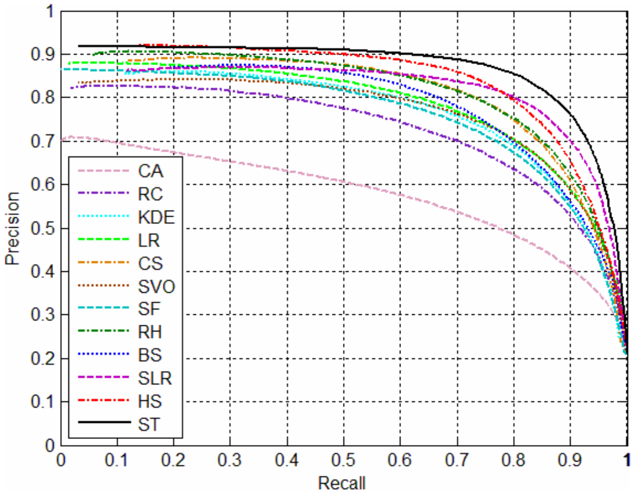


Fig. 13. (better viewed in color) PR curves of different saliency models on MSRA dataset.

rows in Fig. 8, row 6 and 9 in Fig. 11), and having a cluttered background (such as row 1, 4 and 10 in Fig. 7, row 7 in Fig. 8, row 2 in Fig. 9, row 1 and 4 in Fig. 10, and the top five rows in Fig. 11), the proposed ST model can suppress background regions and highlight the complete salient object regions with well-defined boundaries more effectively than the other saliency models. Thanks to the use of tree structure and the systematic saliency tree analysis process, ST model can better handle the problems of heterogeneous objects, cluttered background and low contrast between object and background more effectively compared to other saliency models.

Besides, ST model can highlight both large-scale salient objects (such as row 2 in Fig. 9, row 3 in Fig. 10, and the bottom row in Fig. 11) and tiny-scale salient object (such as row 6 in Fig. 10) more effectively compared to other saliency models, due that the regional center-surround scheme used in ST model flexibly addresses the issue of object scale compared to single or several fixed scales used in other saliency models. Note that the issue of multiple objects itself is not challenging for most saliency models in case that multiple objects are well contrasted with the background (such as row 5 and 6 in Fig. 9). However, if images containing multiple objects are coupled with the above mentioned problems of heterogeneous objects, cluttered background, low contrast and object scale (such as row 5 and 8 in Fig. 7, the bottom row in Fig. 10, and row 7 in Fig. 11), ST model is more effective to highlight multiple salient objects with well-defined boundaries.

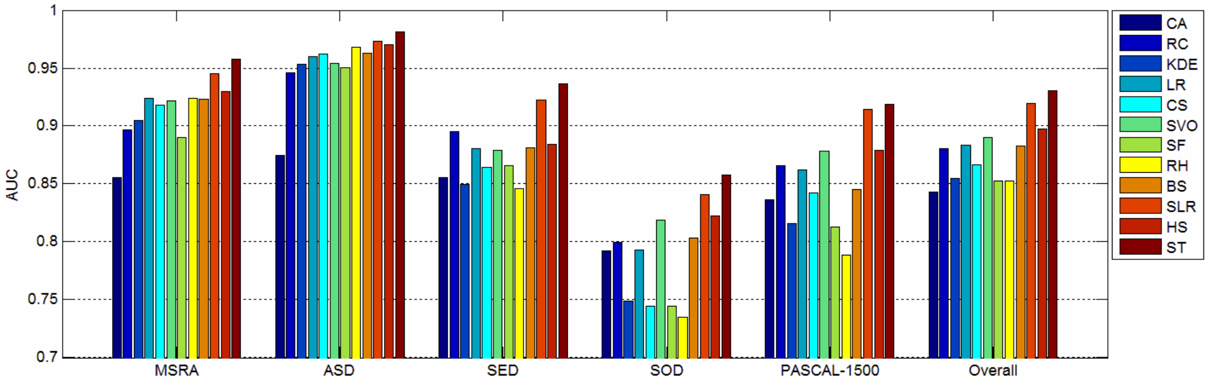


Fig. 14. AUC values achieved using different saliency models on all the five datasets.

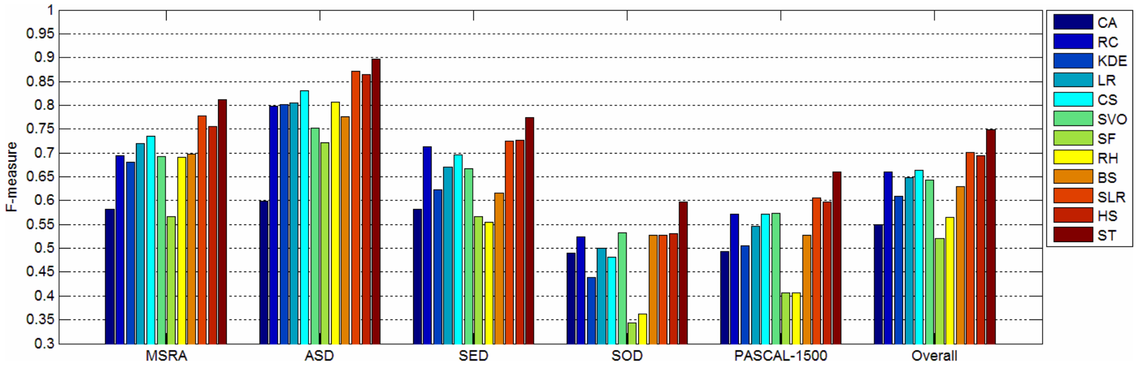


Fig. 15. F-measures achieved using different saliency models on all the five datasets.

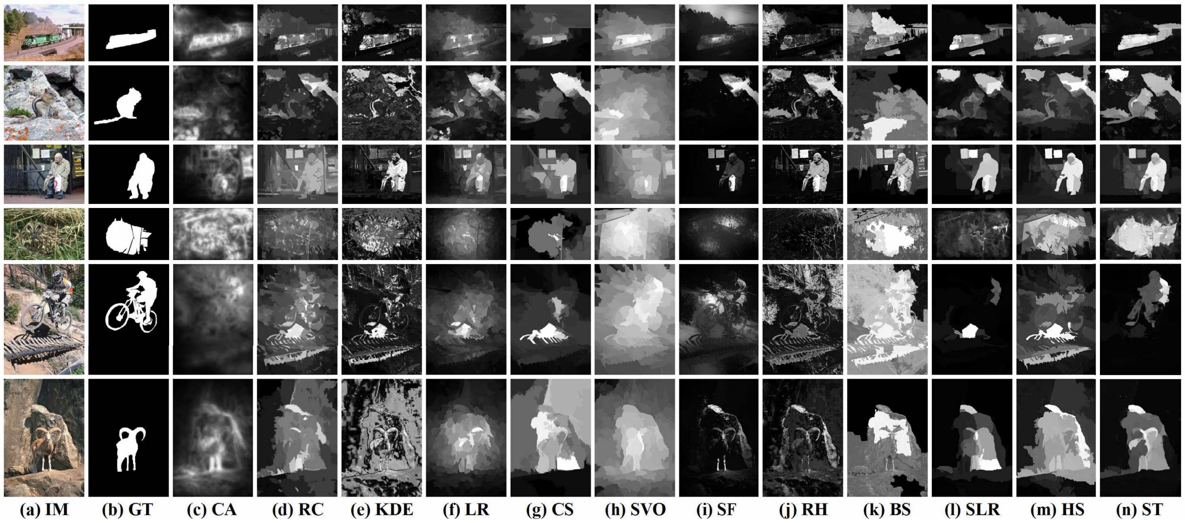


Fig. 16. Some failure examples. (a) images (IM), (b) ground truths (GT) and (c)–(n) saliency maps generated using different models.

D. Objective Evaluation

In order to objectively evaluate the saliency detection performance of different saliency models, ROC curves are generated for each saliency model on each dataset using the same method in Sec. VI-B. Similarly as the generation process of ROC curves, we also generate for each saliency model on each dataset the precision-recall (PR) curve, which plots the precision measure against the recall measure to characterize the saliency detection performance. Due to the page limit, only the ROC curves and PR curves on the largest dataset, i.e., MSRA, are shown in Figs. 12 and 13, respectively. We can see from Figs. 12 and 13 that both ROC curve and PR curve of ST model are the highest one, which demonstrates the better saliency detection performance of ST model.

For a quantitative and intuitive comparison, Fig. 14 shows the AUC values, which are calculated for ROC curves of all saliency models on all the five datasets, and objectively demonstrates that ST model consistently outperforms all the other saliency models on all the five datasets. Besides, we can observe from Fig. 14 that for each saliency model, the highest and the lowest AUC value are consistently achieved on ASD dataset and SOD dataset, respectively. This indicates that the widely used ASD dataset is relatively simpler for the state-of-the-art saliency models (note that 10 models achieve an AUC value higher than 0.95), while SOD dataset, which is originally designed for evaluation of image segmentation performance and contains a variety of natural scenes, are the most challenging for saliency detection.

In order to evaluate the quality of saliency maps and the applicability for salient object detection and segmentation more explicitly, we performed adaptive thresholding operation on each saliency map using the well-known Otsu's method [60], which is simple yet effective, to obtain a binary salient object mask. We calculate the measures of precision and recall by comparing each binary salient object mask with the corresponding binary ground truth, and then calculate F-measure, which is the harmonic mean of precision and recall, to evaluate the overall performance as follows:

$$F_{\gamma} = \frac{(1+\gamma)\cdot precision \cdot recall}{\gamma \cdot precision + recall} \tag{28}$$

where the coefficient γ is set to 1 indicating the equal importance of precision and recall. For each dataset, the average F-measure on all saliency maps generated using each saliency model is calculated and shown in Fig. 15. We can see from Fig. 15 that on all the five datasets, ST model consistently achieves the highest F-measure, which objectively demonstrates the overall better quality of saliency maps generated using ST model.

E. Failure Cases and Analysis

As shown in the previous two subsections, the proposed ST model outperforms the state-of-the-art saliency models on both subjective and objective evaluation. However, some difficult images are still challenging for ST model as well as other state-of-the-art saliency models. If an image contains a part of background regions, which are visually salient against the major part of background, such as rows 1-3 in Fig. 16, it is difficult to suppress such visually salient background regions. In addition, if a part of salient object shows a very similar color with its nearby background regions in a cluttered scene, such as rows 3-6 in Fig. 16, the salient object cannot be completely highlighted or/and the nearby background regions are erroneously highlighted in the generated saliency maps. The proposed ST model as well as the state-of-the-art saliency models are still not effective to handle such difficult cases mentioned above. It should be noted that some class-specific knowledge about human object and vehicles (such as motorcycle and train) can be incorporated into saliency models to improve the saliency detection performance on such images in the row 1, 3 and 5 of Fig. 16, for specific applications such as detection of human objects and vehicles.

F. Complexity Analysis

We implemented the proposed ST model using Matlab R2012b, and used the source code of the gPb-owt-ucm method [54], which is written mostly in C++ and coordinated by Matlab scripts, to generate the UCM for region segmentation. Our experiments are performed on a laptop with Intel Core i7-3720QM 2.6GHz CPU and 8GB RAM.

TABLE I

AVERAGE PROCESSING TIME AND MEMORY CONSUMPTION OF EACH COMPONENT IN THE PROPOSED ST MODEL PER IMAGE IN THE SOD DATASET

Component		Time (Sec.)	Memory (MB)
UCM generation	gPb	78.170	1630
	owt-ucm	1.608	
Image simplification (excluding UCM generation)		0.382	170
Regional saliency measurement		0.250	
Saliency tree generation		0.582	
Saliency tree analysis	Salient node selection, regional saliency adjustment and selection	0.636	
	Pixel-wise saliency map derivation	0.234	
Total		81.862 (2.084*)	/

* Total processing time excluding UCM generation.

The average processing time and memory consumption of each component in the proposed ST model per image in the SOD dataset (all images have a resolution of either 481 × 321 or 321 × 481, equivalent to about 0.15 Megapixel) are shown in Table I. Note that in the current implementation, the resizing factor for eigenvector computation in the gPb is set to 0.5 to reduce the time complexity and memory consumption. Even so, the gPb still occupies a large amount of time and a large memory. In contrast, all the own components of ST model (excluding UCM generation) only take 2.084 seconds in total, and the memory consumption is also lower.

Therefore, in order to make the proposed ST model more practical for applications with runtime requirements, the computational efficiency of gPb, which is the bottleneck of runtime, should be improved with the highest priority. Fortunately, as reported in [61], the gPb method can be effectively parallelized and accelerated using a GPU implementation, which can process the image with a resolution of 0.15 Megapixel in the BSD dataset (note that all 300 images in the SOD dataset are from the BSD dataset as mentioned in Sec. VI-A) within 1.8 seconds on a NVidia GTX 280 GPU. The three own components of ST model can also be parallelized using a GPU implementation. Specifically, regional saliency measurement can be parallelized on the basis of primitive region; salient node selection, regional saliency adjustment and selection can be parallelized on the basis of saliency visibility level; pixel-wise saliency map derivation can be parallelized on the basis of pixel. We believe that a parallel GPU implementation of ST model will substantially improve the computational efficiency.

VII. CONCLUSION

In this paper, we have presented saliency tree as a novel saliency detection framework, which provides a hierarchical representation of saliency for generating high-quality regional and pixel-wise saliency maps. Initial regional saliency is measured by integrating global contrast, spatial sparsity and object prior of primitive regions to build a reasonable basis for generating the saliency tree. Then saliency-directed region merging, regional center-surround scheme, salient node selection, regional saliency adjustment and selection, and pixel-wise saliency map derivation are proposed and systematically integrated into a complete saliency tree model. Both subjective and objective evaluations on five datasets demonstrate that saliency tree achieves a consistently higher saliency detection performance compared to the state-of-the-art saliency models, and especially enhances the applicability on complicated images.

In our future work, we will extend the current saliency tree model with the incorporation of motion fields and inter-frame spatiotemporal correlations for effective saliency detection in videos. Specifically, saliency detection using the current saliency tree model is only performed on some key frames, which are selected on the basis of video shot with a constraint of maximum interval. Then a regional motion trajectory based temporal saliency measure will be designed to modulate the current final regional saliency measure for each primitive region in the key frame, as its spatiotemporal saliency measure. Finally, we will investigate an inter-frame regional saliency propagation scheme using motion fields, to estimate for each non-key frame its region partition and the spatiotemporal saliency measures of regions, based on the results available in the preceding and the following key frame. The pixel-wise saliency map derivation method can be adapted for estimating pixel's saliency from its spatiotemporal neighboring regions.

Acknowledgment

Zhi Liu and Wenbin Zou contributed equally to this paper. 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 Ms. Shuhua Luo and Dr. Wanlei Zhao for their help on performing experiments.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object," in *Proc. IEEE CVPR*, Jun. 2010, pp. 73–80.

[2] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, "Region diversity maximization for salient object detection," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 215–218, Apr. 2012.

[3] 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.

[4] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *Proc. ECCV*, Sep. 2010, pp. 366–379.

[5] C. Jung, and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, Mar. 2012.

[6] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[7] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, "Retargeting images and video for preserving information saliency," *IEEE Comput. Graph. Appl.*, vol. 27, no. 5, pp. 80–88, Sep. 2007.

[8] A. Shamir, and S. Avidan, "Seam carving for media retargeting," *Commun. ACM*, vol. 52, no. 1, pp. 77–85, Jan. 2009.

[9] Y. Luo, J. Yuan, P. Xue, and Q. Tian, "Saliency density maximization for efficient visual objects discovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1822–1834, Dec. 2011.

[10] 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.

[11] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.

[12] 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.

[13] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. ECCV*, Oct. 2012, pp. 414–429.

[14] 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.

[15] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[16] 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.

[17] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Multimedia*, Nov. 2003, pp. 374–381.

[18] 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.

[19] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 446–456, Apr. 2011.

[20] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Proc. NIPS*, Dec. 2007, pp. 497–504.

[21] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, Nov. 2009.

[22] Y. Lin, Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth Mover's distance based saliency measurement and nonlinear feature combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 314–328, Feb. 2013.

[23] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.

[24] R. Achanta and S. Susstrunk, "Saliency detection using maximum symmetric surround," in *Proc. IEEE ICIP*, Sep. 2010, pp. 2653–2656.

[25] 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.

[26] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, Dec. 2008.

[27] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, Nov. 2001.

[28] 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.

[29] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.

[30] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[31] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, Dec. 2006, pp. 545–552.

[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] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[34] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.

[35] 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 CVPR*, Jun. 2013, pp. 2083–2090.

[36] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE ICCV*, Sep. 2009, pp. 2106–2113.

[37] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE CVPR*, Jun. 2010, pp. 2376–2383.

[38] 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.

[39] 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.

[40] Z. Liu, Y. Xue, H. Yan, and Z. Zhang, "Efficient saliency detection based on Gaussian models," *IET Image Process.*, vol. 5, no. 2, pp. 122–131, Mar. 2011.

[41] 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.

[42] 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.

[43] X. Zhang, Z. Ren, D. Rajan, and Y. Hu, "Salient object detection through over-segmentation," in *Proc. IEEE ICME*, Jul. 2012, pp. 1033–1038.

[44] Z. Liu, O. Le Meur, S. Luo, and L. Shen, "Saliency detection using regional histograms," *Opt. Lett.*, vol. 38, no. 5, pp. 700–702, Mar. 2013.

[45] Z. Liu, O. Le Meur, and S. Luo, "Superpixel-based saliency detection," in *Proc. IEEE WIAMIS*, Jul. 2013, pp. 1–4.

[46] F. Perazzi1, P. Krähenbül, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE CVPR*, Jun. 2012, pp. 733–740.

[47] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE CVPR*, Jun. 2013, pp. 1155–1162.

[48] K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. IEEE ICCV*, Nov. 2011, pp. 914–921.

[49] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. BMVC*, Aug. 2011, pp. 1–12.

[50] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE CVPR*, Jun. 2012, pp. 853–860.

[51] W. Zou, K. Kpalma, Z. Liu, and J. Ronsin, "Segmentation driven low-rank matrix recovery for saliency detection," in *Proc. BMVC*, Sep. 2013, pp. 1–13.

[52] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, vol. 3. Oct. 2012, pp. 29–42.

[53] Y. Xie, H. Lu, and M. H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.

[54] 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 2011.

[55] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 561–576, Apr. 2000.

[56] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.

[57] V. Movahedi, and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE POCV*, Jun. 2010, pp. 49–56.

[58] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE ICCV*, vol. 2. Jul. 2001, pp. 416–423.

[59] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[60] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[61] B. Catanzaro, B.-Y. Su, N. Sundaram, Y. Lee, M. Murphy, and K. Keutzer, "Efficient, high-quality image contour detection," in *Proc. IEEE ICCV*, Sep. 2009, pp. 2381–2388.

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

He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, China. Since Aug. 2012, he has also been a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. His current research interests include saliency model, image/video segmentation, image/video retargeting, video coding, and multimedia communication.

Dr. Liu has published more than 100 refereed technical papers in international journals and conferences. He served as a TPC Member with ICME 2014, WIAMIS 2013, IWVP 2011, PCM 2010, and ISPACS 2010. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014.

Wenbin Zou received the Ph.D. degree from the National Institute of Applied Sciences (INSA), Rennes, France, in 2014. He received the M.E. degree in Software Engineering, with a specialization in Multimedia Technology, from Peking University, China, in 2010. He was a Visiting Research Student with the Hong Kong University of Science and Technology from 2008 to 2009.

He is currently on the faculty of the College of Information Engineering, Shenzhen University, China. His current research interests include saliency detection, object segmentation, and semantic segmentation.

Olivier Le Meur received the Ph.D. degree from the University of Nantes in 2005. From 1999 to 2009, he was involved in the media and broadcasting industry. In 2003, he joined the Research Center of Thomson-Technicolor, Rennes, where he supervised a research project concerning the modeling of human visual attention.

He has been an Associate Professor of image processing with the University of Rennes 1 since 2009. In the SIROCCO Team of IRISA/INRIA-Rennes, his current research interests include the understanding of the human visual attention, which includes computational modeling of the visual attention and saliency-based applications (video compression, objective assessment of video quality, and retargeting).