Salient Object Detection: A Discriminative Regional Feature Integration Approach

Huaizu Jiang[†] Jingdong Wang[‡] Zejian Yuan[†] Yang Wu[§] Nanning Zheng[†] Shipeng Li[‡]
[†]Xi'an Jiaotong University [‡]Microsoft Research Asia [§]Kyoto University
https://sites.google.com/site/jianghz88/saliency_drfi

Abstract

Salient object detection has been attracting a lot of interest, and recently various heuristic computational models have been designed. In this paper, we regard saliency map computation as a regression problem. Our method, which is based on multi-level image segmentation, uses the supervised learning approach to map the regional feature vector to a saliency score, and finally fuses the saliency scores across multiple levels, yielding the saliency map. The contributions lie in two-fold. One is that we show our approach, which integrates the regional contrast, regional property and regional backgroundness descriptors together to form the master saliency map, is able to produce superior saliency maps to existing algorithms most of which combine saliency maps heuristically computed from different types of features. The other is that we introduce a new regional feature vector, backgroundness, to characterize the background, which can be regarded as a counterpart of the objectness descriptor [2]. The performance evaluation on several popular benchmark data sets validates that our approach outperforms existing state-of-the-arts.

1. Introduction

Visual saliency has been a fundamental problem in neuroscience, psychology, neural systems, and computer vision for a long time. It is originally a task of predicting the eye-fixations on images, and recently has been extended to identifying a region containing the salient object, which is the focus of this paper. There are various applications for salient object detection, including object detection and recognition [25, 46], image compression [21], image cropping [35], photo collage [17, 47], dominant color detection [51, 52] and so on.

The study on human visual systems suggests that the saliency is related to uniqueness, rarity and surprise of a scene, characterized by primitive features like color, texture, shape, etc. Recently a lot of efforts have been made to design various heuristic algorithms to compute the saliency [1, 6, 11, 15, 18, 27, 31, 34, 38].

In this paper, we regard saliency estimation as a regression problem, and learn a regressor that directly maps the regional feature vector to a saliency score. Our approach consists of three main steps. The first one is multi-level segmentation, which decomposes the image to multiple segmentations from a fine level to a coarse one. Second, we conduct a region saliency computation step with a random forest regressor that maps the regional features to a saliency score. Last, a saliency map is computed by fusing the saliency maps across multiple levels of segmentations.

The key contributions lie in the second step, region saliency computation. Unlike most existing algorithms that compute saliency maps heuristically from various features and combine them to get the saliency map, which we call saliency integration, we learn a random forest regressor that directly maps the feature vector of each region to a saliency score, which we call discriminative regional feature integration (DRFI). This is a principle way in image classification [19], but rarely studied in salient object detection. It turns out that the learnt regressor is able to automatically pick discriminative features rather than heuristically hand-crafting special features for saliency. On the other hand, we also introduce a new descriptor, called backgroundness, to discriminate the background from the object, which can be considered as a counterpart of the objectness descriptor [2].

1.1. Related work

The following gives a review of salient object detection (segmentation) algorithms that are related to our approach. A comprehensive survey of salient object detection can be found from [9]. The review on visual attention modeling [7] also includes some analysis on salient object detection.

The basis of most saliency detection algorithms can date back to the feature integration theory [43] which posits that different kinds of attention are responsible for binding various features into consciously experienced wholes. Later, a computational attention model built on a biologically-plausible architecture is proposed in [28] and completely implemented in [22]. It represents the input image from the color, intensity and orientation channels, and computes three conspicuity (saliency) maps using center-surround differences, which are combined together to form the final master saliency map.

Recently, a lot of research efforts have been made to design various saliency features characterizing salient objects or regions. Most works essentially follow the center-surround difference (or contrast) framework. The discriminant center-surround hypothesis is analyzed in [15, 16]. Color histograms, computed to represent the center and the surround, are used to evaluate the center-surround dissimilarity [31]. An information theory perspective is introduced to yield a sound mathematical formulation, computing the center-surround divergence based on feature statistics [27].

The center-surround difference framework is also investigated to compute the saliency from region-based image representation. In [23], the difference between the color histogram of a region and its immediately neighboring regions are used to evaluate the saliency score. The global contrast based approach [11], computing the saliency map by comparing each region with others, aims to directly compute the global uniqueness. Based on the regional contrast, element color uniqueness and spatial distribution are introduced to evaluate the saliency scores of regions [38]. The saliency map is generated by propagating the saliency scores of regions to the pixels.

Many other models are also proposed for saliency computation. Center-bias, i.e. the salient object usually lies in the center of an image, is investigated in [23, 50]. Object prior, such as connectivity prior [45], concavity context [34], auto-context cue [48], and the background prior [53] are also studied for saliency computation. Example-based approaches, searching for similar images of the input, are developed for salient object detection [35, 49]. A graphical model is proposed to fuse generic objectness and visual saliency together to detect objects [10]. A low rank matrix recovery scheme is proposed for salient object detection [41]. A top-down approach via joint conditional random fields and dictionary learning is introduced [54]. The stereopsis is leveraged for saliency analysis [37]. Besides, spectral analysis in the frequency domain is used to detect salient regions [1, 20]

Additionally, there are several works directly checking if an image window contains an object. The generic objectness measure is defined by combining several image cues to quantify the possibility that a window contains an object [2]. A category independent object detection cascade, which uses superpixel boundary integral, edge distribution and window symmetry to describe objectness, is learnt to rank a number of object window candidates [39]. Salient object detection by composition [13] checks if the content within an window can be composed by neighbor regions. A random forest regression approach is adopted to directly regress the object rectangle from the saliency map [50].

Eye fixation prediction, another visual saliency research direction, also attracts a lot of interests [7, 24]. Recent developments include using isocentric curvedness and color [44], adopting image histogram [32], quaternion-based spectral analysis [40], utilizing depth cues [30], multitask sparsity pursuit [29], statistically modeling [42], exploring patch rarities [6], combing bottom-up and top-down features [5], task-specific visual attention [8] and so on. There are some other saliency definitions, e.g. context-aware saliency detection [18] aiming to detect the image regions that represent the scene.

Our proposed approach differs from existing algorithms on two points. In term of the saliency features, we compute a contrast vector instead of a contrast value used in the existing algorithms for a region. Particularly, a novel feature vector is introduced to characterize the background. Our approach is also unique in the learning strategy. In contrast to existing learning algorithms that perform saliency integration by combining saliency maps computed from different types of features, e.g. [2, 10, 31], our approach learns to directly integrate feature vectors to compute the saliency map. The closely related approach [26] which also learns to integrate the saliency features is a pixel-based algorithm, while our approach is region-based that performs multi-level estimation and can capture non-local contrast. Moreover, we introduce a novel regional feature vector to characterize the background. Another one [36] touches the discriminative feature integration lightly without presenting a deep investigation and it only considers the regional property descriptor. The recent learning approach [33] aims to predict eye fixation, while our approach is for salient object detection and moreover, we solve the problem by introducing and exploring multi-level regional descriptors. The discriminative feature fusion has also been studied in image classification [14], which learns the adaptive weights of features according to the classification task to better distinguish one class from others. Our approach

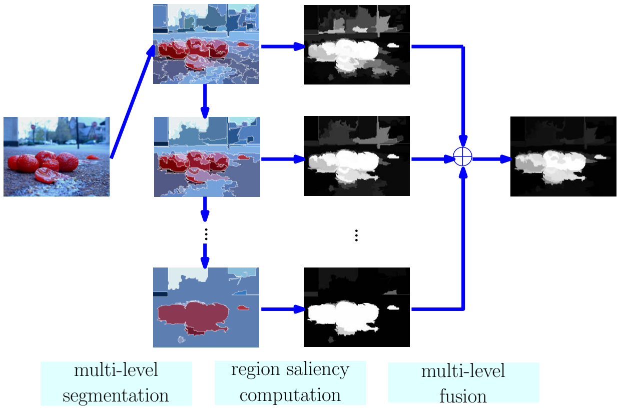


Figure 1. The framework of our proposed discriminative regional feature integration (DRFI) approach.

integrates three types of regional features in a discriminative strategy for the saliency regression on multiple segmentations.

2. Image saliency computation

The pipeline of our approach consists of three main steps: multi-level segmentation that decomposes an image into regions, region saliency computation that maps the features extracted from each region to a saliency score, and multi-level saliency fusion that combines the saliency maps over all the levels of segmentations to get the final saliency map. The whole process is illustrated in Figure 1.

Multi-level segmentation. Given an image I, we represent it by a set of M-level segmentations $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_M\}$, where each segmentation $\mathcal{S}_m$ is a decomposition of the image I and consists of K_m regions. $\mathcal{S}_1$ is the finest segmentation consisting of the largest number of regions, and $\mathcal{S}_M$ is the coarsest segmentation consisting of the smallest number of regions.

We apply the graph-based image segmentation approach [12] and compute the over-segmentation $\mathcal{S}_1 = \{R^1_1, R^1_2, \cdots, R^1_{K_1}\}$. Other segmentations $\{\mathcal{S}_2, \cdots, \mathcal{S}_M\}$ are computed based on $\mathcal{S}_1$, and specifically $\mathcal{S}_m$ is computed by merging the regions in $\mathcal{S}_{m-1}$. The regions in $\mathcal{S}_{m-1}$ are represented by a weighted graph, which connects the spatially neighboring regions. Pairs of regions are sequentially merged in the order of decreasing the weights of edges (the similarities of the corresponding regions) until the weight of two regions is greater than the specified threshold (controlled by the parameter k of the approach [12]. See details in [12]).

Region saliency computation. Our algorithm computes the saliency score for each region. It seems that the separate computation ignores the relation of neighboring regions. However, our algorithm essentially takes into consideration of such relations because we conduct the region saliency computation on multi-level segmentation. The spatial consistency of saliency scores for neighboring regions is imposed since the neighboring regions in the finer-level segmentation may form a single region in the coarser level.

Our approach represents each region using three types of features: regional contrast, regional property, and regional backgroundness, which will be described in Section 3. At present, we denote the feature as a vector $\mathbf{x}$. Then the feature $\mathbf{x}$ is passed into a random forest regressor f, yielding a saliency score. The random forest regressor is learnt from the regions of the training images, and integrates the features together in a discriminative strategy. The learning procedure will be given in Section 4.

Multi-level saliency fusion. After conducting region saliency computation, each region $R^m_n \in \mathcal{S}_m$ has a saliency value a^m_n. For each level, we assign the saliency value of each region to its contained pixels. As a result, we generate M saliency maps $\{\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_M\}$, and then fuse them together, $\mathbf{A} = g(\mathbf{A}_1, \cdots, \mathbf{A}_M)$, to get the final saliency map $\mathbf{A}$, where g is a combinator function introduced in section 4.

3. Regional features

3.1. Regional contrast descriptor

A region is likely thought to be salient if it is different from its surrounding regions. Unlike most existing approaches that directly compute the contrast values, e.g. the differences of region features like color and texture, and then combine them together directly forming a saliency score, our approach computes a contrast descriptor, which will be fed into a regressor to automatically calculate the saliency score.

To compute the contrast descriptor, we describe each region by a feature vector, including color and texture features, denoted by $\mathbf{v}$. The detailed description is given in Table 1. For a region $R \in \mathcal{S}_m$, we regard its immediately neighboring regions as a single one and compute the color and texture features $\mathbf{v}^N$ to represent the neighborhood. The regional contrast descriptor of R is computed as the differences $\text{diff}(\mathbf{v}^R, \mathbf{v}^N)$ between its features and the neighborhood features. Specifically, the difference of the histogram feature is computed as the distribution divergence, and the differences of other features are computed as the absolute elements differences of the vectors. As a result, we get a 26-dimensional feature vector. The details of the regional contrast descriptor are given in Table 1.

Table 1. Color and texture features describing the visual characteristics of a region which are used to compute the regional feature vector. $d(\mathbf{x}_1, \mathbf{x}_2) = (|x_{11} - x_{21}|, \cdots, |x_{1d} - x_{2d}|)$ where d is the number of elements in the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. $\chi^2(\mathbf{h}_1, \mathbf{h}_2) = \sum_{i=1}^{b} \frac{2(h_{1i} - h_{2i})^2}{h_{1i} + h_{2i}}$ with b being the number of histogram bins. The last two columns denote the symbols for regional contrast and backgroundness descriptors. (In the definition column, S corresponds to N for the regional contrast descriptor and B for the regional backgroundness descriptor, respectively.)

Color and texture features			Differences of features		Contrast	Backgroundness
	features	dim	definition	dim		
$\mathbf{a}_1$	the average RGB values	3	$d(\mathbf{a}_1^R, \mathbf{a}_1^S)$	3	$c_1 \sim c_3$	$b_1 \sim b_3$
$\mathbf{a}_2$	the average L*a*b* values	3	$d(\mathbf{a}_2^R, \mathbf{a}_2^S)$	3	$c_4 \sim c_6$	$b_4 \sim b_6$
$\mathbf{r}$	the absolute response of LM filters	15	$d(\mathbf{r}^R, \mathbf{r}^S)$	15	$c_7 \sim c_{21}$	$b_7 \sim b_{21}$
r	the max response among the LM filters	1	$d(r^R, r^S)$	1	c_{22}	b_{22}
$\mathbf{h}_1$	the L*a*b* histogram	$8 \times 16 \times 16$	$\chi^2(\mathbf{h}_1^R, \mathbf{h}_1^S)$	1	c_{23}	b_{23}
$\mathbf{h}_2$	the hue histogram	8	$\chi^2(\mathbf{h}_2^R, \mathbf{h}_2^S)$	1	c_{24}	b_{24}
$\mathbf{h}_3$	the saturation histogram	8	$\chi^2(\mathbf{h}_3^R, \mathbf{h}_3^S)$	1	c_{25}	b_{25}
$\mathbf{h}_4$	the texton histogram	65	$\chi^2(\mathbf{h}_4^R, \mathbf{h}_4^S)$	1	c_{26}	b_{26}

3.2. Regional property descriptor

In addition to regional contrast, we consider the generic properties of a region, including appearance and geometric features. The two features are extracted independently from each region like the feature extraction algorithm in image labeling [19]. The appearance features attempt to describe the distribution of colors and textures in a region, which can characterize the common properties of the salient object and the background. For example, the background usually has homogeneous color distribution or similar texture pattern. The geometric features include the size and position of a region that may be useful to describe the spatial distribution of the salient object and the background. For instance, the salient object tends to be placed near the center of the image while the background usually scatters over the entire image. In summary, we obtain a 34-dimensional regional property descriptor. The details are given in Table 2.

3.3. Regional backgroundness descriptor

There exist a few algorithms attempting to make use of the characteristics of the background (e.g. homogeneous color or textures) to heuristically determine if one region is background, e.g. [53]. In contrast, our algorithm extracts a set of features and adopts the supervised learning approach to determine the background degree (accordingly the object degree) of a region.

It has been observed that the background identification depends on the whole image context. Image regions with similar appearances might belong to the background in one image but belong to the salient object in some other ones. It is not enough to merely use the property features to check if one region is in the background or the salient object.

Therefore, we extract the pseudo-background region and compute the backgroundness descriptor for each region with the pseudo-background region as a reference. The pseudo-background region B is defined as the 15-pixel wide narrow border region of the image. To verify such a definition, we made a simple survey on the MSRA-B data set with 5000 images and found that 98% of pixels in the border area belongs to the background. The backgroundness feature of the region R is then computed as the differences $\text{diff}(\mathbf{v}^R, \mathbf{v}^B)$ between its features $\mathbf{v}^R$ and the features $\mathbf{v}^B$ of the pseudo-background region, resulting a 26-dimensional feature vector. See details in Table 1.

4. Learning

Learning the regional saliency regressor. We aim to learn the regional saliency estimator from a set of training examples. The training examples include a set of confident regions $\mathcal{R} = \{R_1, R_2, \cdots, R_Q\}$ and the corresponding saliency scores $\mathcal{A} = \{a_1, a_2, \cdots, a_Q\}$, which are collected from the multi-level segmentation over a set of images with the ground truth annotation of the salient objects. A region is considered to be confident if the number of the pixels belonging to the salient object or the background exceeds 80% of the number of the pixels in the region, and its saliency score is set as 1 or 0 accordingly. In experiments we find that few regions of all the training examples, around 6%, are unconfident and do not use them for training.

As aforementioned, each region is described by a feature vector $\mathbf{x}$, composed of the regional contrast, regional property, and regional backgroundness descriptors. We learn a random forest regressor f from the training data $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_Q\}$ and the saliency scores $\mathcal{A} = \{a_1, a_2, \cdots, a_Q\}$. Learning a saliency regressor can automatically combine the features and discover the most discriminative ones. Figure 3 presents the most important 20 features.

Learning the multi-level saliency fusor. Given the multi-level saliency maps $\{\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_M\}$ for

Table 2. The regional property descriptor.

description	notation	dim	description	notation	dim
the average normalized x coordinates	p_1	1	the average normalized y coordinates	p_2	1
the normalized perimeter	p_7	1	the $10th$ percentile of the normalized x coordinates	p_3	1
the aspect ratio of the bounding box	p_8	1	the $10th$ percentile of the normalized y coordinates	p_4	1
the variances of the RGB values	$p_9 \sim p_{11}$	3	the $90th$ percentile of the normalized x coordinates	p_5	1
the variances of the L*a*b* values	$p_{12} \sim p_{14}$	3	the $90th$ percentile of the normalized y coordinates	p_6	1
the variances of the HSV values	$p_{15} \sim p_{17}$	3	the variance of the response of the LM filters	$p_{18} \sim p_{32}$	15
the normalized area	p_{33}	1	the normalized area of the neighbor regions	p_{34}	1

an image, our aim is to learn a combinator $g(\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_M)$ to fuse them together to form the final saliency map $\mathbf{A}$. Such a problem has been already addressed in existing methods, such as the conditional random field solution [31]. In our implementation, we find that a linear combinator, $\mathbf{A} = \sum_{m=1}^{M} w_m \mathbf{A}_m$, performs well by learning the weights using a least square estimator, i.e. , minimizing the sum of the losses $(\|\mathbf{A} - \sum_{m=1}^{M} w_m \mathbf{A}_m\|_F^2)$ over all the training images.

5. Experimental results

5.1. Setup

We evaluate the performance over several data sets that are widely used in previous works, e.g. [9, 11, 23].

MSRA-B[1]. This data set [31] includes 5000 images, originally containing labeled rectangles from nine users drawing a bounding box around what they consider the most salient object. There is a large variation among images including natural scenes, animals, indoor, outdoor, etc. We manually segmented the salient object (contour) within the user-drawn rectangle to obtain binary masks. The ASD data set [1] is a subset (binary masks are provided) of the MSRA-B, and thus we no longer make the evaluation on it.

SED[2]. This data set [3] contains two subsets: SED1 that has 100 images containing only one salient object and SED2 that has 100 images containing two salient objects. Pixel-wise ground truth annotation for the salient objects in both SED1 and SED2 are provided.

SOD[3]. This data set is a collection of salient object boundaries based on the Berkeley segmentation data set. Seven subjects are asked to choose the salient object(s) in 300 images. We generate the pixel-wise annotation of the salient objects based on the boundary annotation. This data set contains many images with multiple objects making it challenging.

iCoSeg[4]. This is a publicly available co-segmentation data set [4], including 38 groups of totally 643 images. Each image is along with pixel-wise ground truth annotation, which may contain one or multiple salient objects. In this paper, we use it to evaluate the performance of salient object detection.

We randomly sample 2500 images from the MSRA-B data set to train our model, 500 images as the validation data set, and the remaining 2000 images as the testing data set. Rather than training a model for each data set, we use the model trained from the MSRA-B data set and test it over others. This is because other data sets are too small to train reliable models. More importantly, it can help test the adaptability to other different data sets of the model trained from one data set and avoid the model overfitted to a specific one.

We evaluate the performance using the measures used in [9], including the PR (precision-recall) curve, the ROC (receiver operating characteristic) curve and the AUC (Area Under ROC Curve) score. Precision corresponds to the percentage of salient pixels correctly assigned, and recall is the fraction of detected salient pixels belonging to the salient object in the ground truth. The PR curve is created by varying the saliency threshold that determines if a pixel is on the salient object. The ROC curve can also be generated based on true positive rates and false positive rates obtained during the calculation of PR curve.

5.2. Empirical analysis

Parameter analysis. We show how the level number of segmentations and the number of trees in the random forest regressor influence the performance in Figure 2. The quantitative results are obtained on the validation subset of the MSRA-B data set.

One can see in Figure 2(a) that the AUC score of the saliency maps increases when more levels of segmentations are adopted. The reason is that there may exist some confident regions that cover the most (even whole) part of an object in more levels of segmentations. However, a larger number of segmentations introduce more computational burden. Therefore, to balance the efficiency and the effectiveness, we set M to 15 segmentations in our experiments.

[1] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm

[2] http://www.wisdom.weizmann.ac.il/~vision/Seg_Evaluation_DB/index.html

[3] http://elderlab.yorku.ca/SOD/index.html

[4] http://chenlab.ece.cornell.edu/projects/touch-coseg

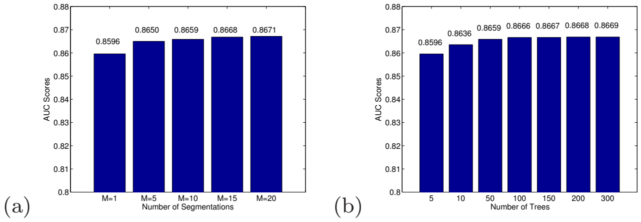


Figure 2. The AUC scores of the saliency maps of the validation set of MSRA-B using (a) different number of segmentations and (b) different number of trees in the random forest regressor.

As shown in Figure 2(b), the performance of our approach with more trees in the random forest saliency regressor is higher. The more trees there are, the less variances are among the weak classifiers, and thus the better performance can be achieved. We choose to use 200 trees to train the regressor to balance the efficiency and the effectiveness. When growing a tree, a node will be split until less than 5 training samples falling in it (i.e. forming a leaf node).

Feature importance. Our approach uses a wide variety of features. We empirically analyze the usefulness of various features. In training a random forest regressor, the feature importance can be estimated by adding the gini impurity decreases for each individual feature over all trees. Figure 3 shows the rank of the most important 20 regional features. The feature rank indicates that the backgroundness descriptor is the most critical one in our feature set (occupies 10 out of top 20 features). The regional contrast descriptor is the least important. The reason might be that it is a local contrast descriptor and less important compared with the regional backgroundness descriptor which is in some sense non-local. In the property descriptor, the geometric features are ranked higher as salient objects tend to lie in the center in most images.

Efficiency. It takes around 24h for training and 10s for testing given a typical 400×300 image on a PC with an Intel i5 CPU of 2.50GHz using our unoptimized MATLAB code. The most time-consuming step is the feature extraction on the multi-level segmentation, which can be accelerated using the parallel or GPU computing techniques as computation on each segmentation is independent on others.

5.3. Performance comparison

We report both quantitative and qualitative comparisons of our approach with state-of-the-art approaches. To save the space, we only consider the top four models ranked in the survey [9]: SVO [10], CA [18], CBsal [23], and RC [11] and recently-developed methods: SF [38] and LRK [41] that are not covered in [9].

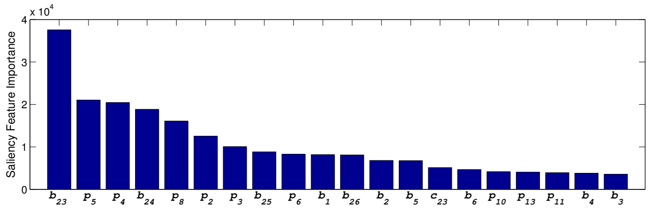
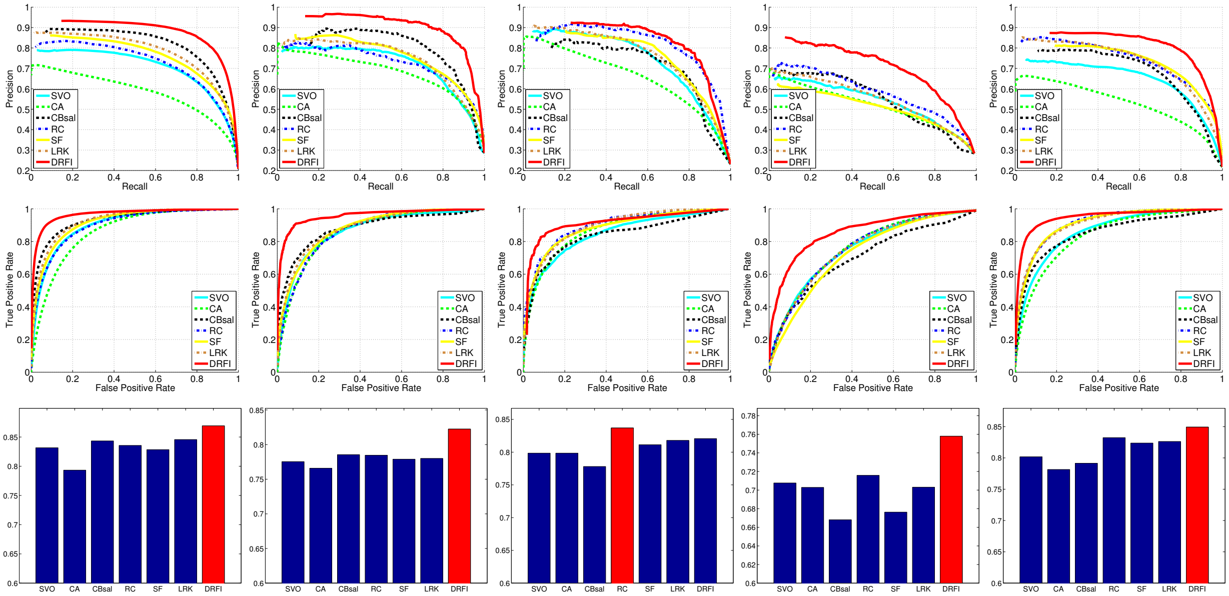


Figure 3. The most important 20 regional features. See Table 1 and Table 2 for the description of the features.

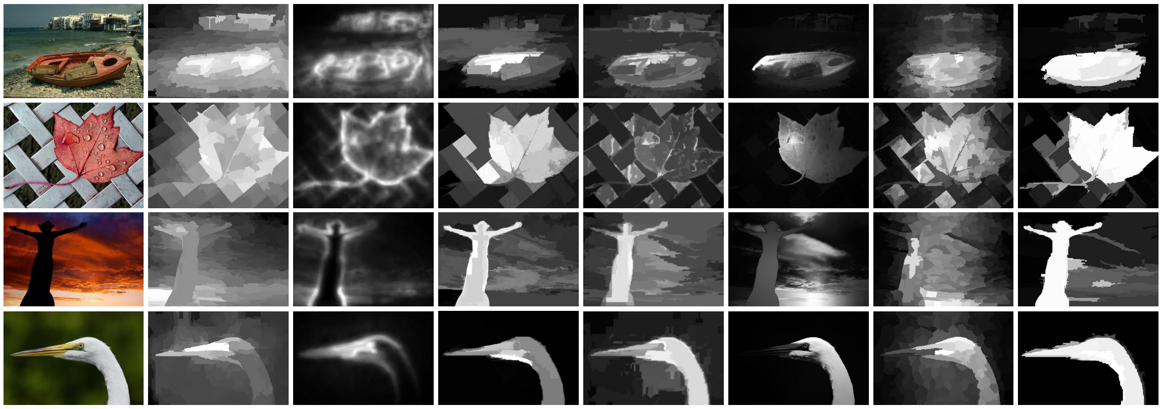
Quantitative comparison. The quantitative comparison is shown Figure 4. As can be seen, our approach (DRFI) achieves the best performance on the MSRA-B and SED1 data sets in which each image contains one single salient object. It improves by 2.77% and 4.70% over the second best algorithms, and 3.06% and 4.81% over the third best algorithms in terms of AUC scores. Additionally, the PR and ROC curves of our approach are consistently higher than others on these two data sets. Figure 4(c), corresponding to the SED2 data set, shows that the true positive rate of our approach is not very good for the high false positive rate, or equivalently, the precision is not so good for the high recall rate. Intuitively, our approach has limited ability when discovering all the salient objects within one image (higher recall). The reason might be that the position and size of the two objects in SED2 are very different from the training set of MSRA-B, where most of the images contain only one object. On other two data sets, SOD and iCoSeg, where an image may also contain one or multiple objects, our approach shows the best performance. It improves by 5.87% and 2.03% over the second best algorithms, and 7.11% and 2.79% over the third best algorithms in terms of the AUC scores. The improvement over state-of-the-arts are substantial when considering their performance and especially the adaptability of our model to different data sets.

Qualitative comparison. We also provide the visual comparison of different methods in Figure 5. As can be seen, our approach (shown in Figure 5 (h)) can deal well with the challenging cases where the background is cluttered. For example, in the first two rows, other approaches may be distracted by the textures on the background while our method almost successfully highlights the whole salient object. It is also worth pointing out that our approach performs well when the object touches the image border, e.g. the third and fourth rows in Figure 5, even though it violates the pseudo-background assumption.



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

Figure 4. Quantitative comparison of saliency maps produced by different approaches on different data sets. From left to right: (a) the MSRA-B data set, (b) the SED1 data set, (c) the SED2 data set, (d) the SOD data set, and (e) the iCoSeg data set. From top to bottom: the PR curves, the ROC curves, and the AUC scores.



(a) input (b) SVO [10] (c) CA [18] (d) CBsal [23] (e) RC [11] (f) SF [38] (g) LRK [41] (h) DRFI

Figure 5. Visual comparison of the saliency maps. Our method (DRFI) consistently generates better saliency maps.

6. Conclusions

In this paper, we address the salient object detection problem using a discriminative regional feature integration approach. The success of our approach stems from two key factors. One is that we learn to integrate a lot of regional descriptors to compute the saliency scores, rather than heuristically compute saliency maps from different types of features and combine them to get the saliency map. The other one is that we introduce the novel backgroundness descriptor, which is proved to be quite effective in our experiments. The groundtruth annotation of MSRA-B data set and our MATLAB implementation are available online.

Acknowledgements This work was supported in part by the National Basic Research Program of China under Grant No. 2012CB316400, and the National Natural Science Foundation of China under Grant No. 91120006.

References

[1] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009.

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, pages 73–80, 2010.

[3] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue

integration. In *CVPR*, 2007.

[4] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176. IEEE, 2010.

[5] A. Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *CVPR*, pages 438–445, 2012.

[6] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012.

[7] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, To Appear.

[8] A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *CVPR*, pages 470–477, 2012.

[9] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV (2)*, pages 414–429, 2012.

[10] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, pages 914–921, 2011.

[11] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[13] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, pages 1028–1035, 2011.

[14] B. Fernando, É. Fromont, D. Muselet, and M. Sebban. Discriminative feature fusion for image classification. In *CVPR*, pages 3434–3441, 2012.

[15] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*, 2007.

[16] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *ICCV*, pages 1–6, 2007.

[17] S. Goferman, A. Tal, and L. Zelnik-Manor. Puzzle-like collage. *Comput. Graph. Forum*, 29(2):459–468, 2010.

[18] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010.

[19] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005.

[20] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[21] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, 13(10):1304–1318, 2004.

[22] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. PAMI.*, 20(11):1254–1259, 1998.

[23] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *British Machine Vision Conference (BMVC)*, 2011.

[24] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. Technical report, MIT-CSAIL-TR-2012-001, 2012.

[25] C. Kanan and G. W. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *CVPR*, pages 2472–2479, 2010.

[26] P. Khuwuthyakorn, A. Robles-Kelly, and J. Zhou. Object of interest detection by saliency learning. In *ECCV (2)*, pages 636–649, 2010.

[27] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, pages 2214–2219, 2011.

[28] C. Kocn and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.

[29] C. Lang, G. Liu, J. Yu, and S. Yan. Saliency detection by multitask sparsity pursuit. *IEEE Transactions on Image Processing*, 21(3):1327–1338, 2012.

[30] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. S. Kankanhalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV (2)*, pages 101–115, 2012.

[31] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(2):353–367, 2011.

[32] S. Lu and J.-H. Lim. Saliency modeling from image histograms. In *ECCV (7)*, pages 321–332, 2012.

[33] Y. Lu, W. Zhang, C. Jin, and X. Xue. Learning attention map from images. In *CVPR*, pages 1067–1074, 2012.

[34] Y. Lu, W. Zhang, H. Lu, and X. Xue. Salient object detection using concavity context. In *ICCV*, pages 233–240, 2011.

[35] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, pages 2232–2239, 2009.

[36] P. Mehrani and O. Veksler. Saliency segmentation based on learning and graph cut refinement. In *British Machine Vision Conference (BMVC)*, 2010.

[37] Y. Niu, Y. Geng, X. Li, and F. Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012.

[38] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740, 2012.

[39] E. Rahtu, J. Kannala, and M. B. Blaschko. Learning a category independent object detection cascade. In *ICCV*, pages 1052–1059, 2011.

[40] B. Schauerte and R. Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *ECCV (2)*, pages 116–129, 2012.

[41] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, pages 853–860, 2012.

[42] X. Sun, H. Yao, and R. Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *CVPR*, pages 1552–1559, 2012.

[43] A. Treisman and G. Gelad. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[44] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, pages 2185–2192, 2009.

[45] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *CVPR*, 2008.

[46] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[47] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum. Picture collage. In *CVPR (1)*, pages 347–354, 2006.

[48] L. Wang, J. Xue, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue. In *ICCV*, pages 105–112, 2011.

[49] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. A. Rowley. Image saliency: From intrinsic to extrinsic context. In *CVPR*, pages 417–424, 2011.

[50] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li. Salient object detection for searched web images via global saliency. In *CVPR*, pages 3194–3201, 2012.

[51] P. Wang, D. Zhang, J. Wang, Z. Wu, X.-S. Hua, and S. Li. Color filter for image search. In *ACM Multimedia*, pages 1327–1328, 2012.

[52] P. Wang, D. Zhang, G. Zeng, and J. Wang. Contextual dominant color name extraction for web image search. In *ICME Workshops*, pages 319–324, 2012.

[53] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV (3)*, pages 29–42, 2012.

[54] J. Yang and M.-H. Yang. Top-down visual saliency via joint crf and dictionary learning. In *CVPR*, pages 2296–2303, 2012.