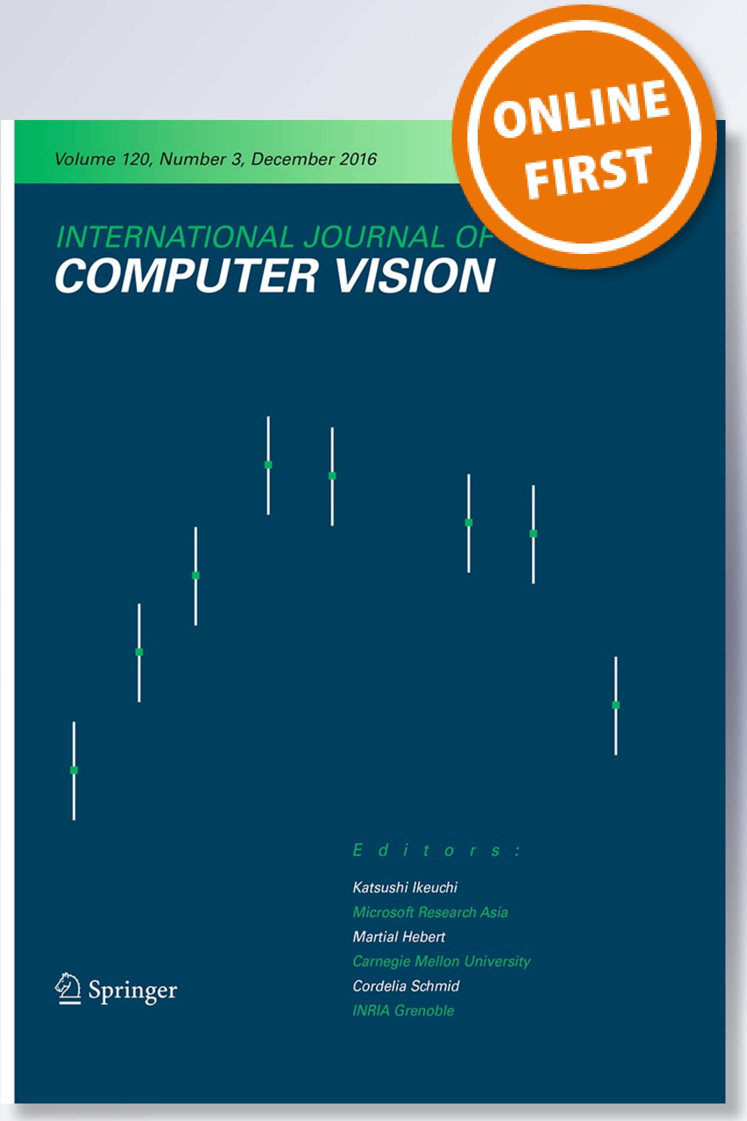
Salient Object Detection: A Discriminative Regional Feature Integration Approach

Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Xiaowei Hu & Nanning Zheng

International Journal of Computer Vision

ISSN 0920-5691

Int J Comput Vis
DOI 10.1007/s11263-016-0977-3



Int J Comput Vis
DOI 10.1007/s11263-016-0977-3

Salient Object Detection: A Discriminative Regional Feature Integration Approach

Jingdong Wang[1] · Huaizu Jiang[2] · Zejian Yuan[3] · Ming-Ming Cheng[4] · Xiaowei Hu[4] · Nanning Zheng[3]

Received: 15 February 2016 / Accepted: 19 November 2016

Abstract Feature integration provides a computational framework for saliency detection, and a lot of hand-crafted integration rules have been developed. In this paper, we present a principled extension, supervised feature integration, which learns a random forest regressor to discriminatively integrate the saliency features for saliency computation. In addition to contrast features, we introduce regional object-sensitive descriptors: the objectness descriptor characterizing the common spatial and appearance property of the salient object, and the image-specific backgroundness descriptor characterizing the appearance of the background of a specific image, which are shown more important for estimating the saliency. To the best of our knowledge, our supervised feature integration framework is the first successful approach to perform the integration over the saliency features for salient object detection, and outperforms the integration approach over the saliency maps. Together with fusing the multi-level regional saliency maps to impose the spatial saliency consistency, our approach significantly outperforms state-of-the-art methods on seven benchmark datasets. We also discuss several followup works which jointly learn the representation and the saliency map using deep learning.

Keywords Salient object detection · Data-driven

1 Introduction

Visual saliency has been a fundamental problem in neuroscience, psychology, neural systems, and computer vision for a long time. It is originally defined as a task of predicting the eye-fixations on images (Itti and Niebur 1998). Recently it is extended to identifying a region (Ma and Zhang 2003; Liu et al. 2007) containing the salient object, known as *salient object detection* or *salient region detection*. Applications of salient object detection include object detection and recognition (Kanan and Cottrell 2010; Walther and Koch 2006), image compression (Itti 2004), image cropping (Marchesotti et al. 2009), photo collage (Goferman et al. 2010; Wang et al. 2006), dominant color detection (Wang et al. 2012b, c) 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. Built upon the feature integration theory (Itti and Niebur 1998; Treisman and Gelad 1980) suggesting a computational framework: compute conspicuity (feature) maps from different saliency cues and then combine them together to form the final saliency map, various heuristic algorithms (Achanta et al. 2009; Borji and Itti 2012; Cheng et al. 2014; Gao et al. 2007; Goferman et al. 2010; Klein and Frintrop 2011; Liu et al. 2011; Lu et al. 2011; Perazzi et al. 2012) have been developed. Hand-crafted integration rules, however, are fragile and poor to generalize. As studied in Borji et al. (2012) and 2015, none of the hand-crafted

Communicated by Cordelia Schmid.

✉ Jingdong Wang
jingdw@microsoft.com

Huaizu Jiang
hzjiang@cs.umass.edu

[1] Microsoft Research, Beijing, People's Republic of China

[2] University of Massachusetts, Amherst, MA, USA

[3] Xi'an Jiaotong University, Xi'an, Shaanxi, People's Republic of China

[4] CCCE & CS, Nankai University, Tianjin, People's Republic of China

Published online: 05 December 2016

integration algorithms can consistently outperform others on the benchmark data sets.

In this paper, we present a discriminative feature integration approach, a supervised extension of the conventional unsupervised feature integration (Treisman and Gelad 1980). Our approach learns a regressor, random forest in this paper, to integrate the saliency features, obtaining a saliency score. The benefit from learning to integrate is to automatically learn the integration rule from the training data without referring to the testing data.

In addition to the contrast descriptor, the difference of a region from its surrounding region, which is widely used as *saliency* (uniqueness) indicators, we introduce extra *object*-sensitive features into the discriminative feature integration framework, to indicate the degrees that a region belongs an object and the background: the objectness descriptor that characterizes the spatial and appearance properties of an object, and the image-specific background descriptor that characterizes the appearance of the background of a specific image.

Considering that a salient object is usually formed from spatially-connected regions and a single image segmentation might be not reliable enough, we propose to fuse the saliency maps computed from the multi-level image segmentations, in order to compute a spatially smooth saliency map and remedy the possible inaccuracy due to unreliable segmentation. Experimental results show that our approach significantly outperforms state-of-the-art methods on seven benchmark datasets.

The main contributions are summarized as follows.

1. We introduce a supervised feature integration approach to salient object detection. The conference version of this paper (Jiang et al. 2013b) is the first to demonstrate the successes of supervised feature integration over the saliency features for salient object detection, though it is simple, and inspires many subsequent data-driven approaches for salient object detection.
2. The generalization capability of the supervised integration approach is demonstrated by the observation that the feature integration model learnt from one dataset achieves the superior performance over other datasets.
3. In addition to the contrast or center-surrounding difference that most previous approaches rely on, our paper introduce two object-sensitive descriptors, the objectness and image-specific backgroundness descriptors, which are found to be more significant than the contrast feature for salient object detection.

This paper is an extension of our previous conference version (Jiang et al. 2013b). The extension includes the following aspects. (1) We point out that in the training phase a supervised approach to multi-level segmentation is superior and that in the testing phase an unsupervised approach to multi-level segmentation is more efficient and effective. (2) We present more empirical analysis, including parameters analysis, features importance, robustness analysis, and compare our approach with more algorithms. (3) We discuss several follow-up works that extend our approach with the deep learning technique. (4) We rewrote our code using C++ to get an efficient implementation and published our C++ code in the project page, http://supermoe.cs.umass.edu/~hzjiang/drfi.

2 Related Work

Saliency detection includes two basic tasks, eye fixation prediction and salient object detection. The research on eye fixation prediction started from a general computational framework (Treisman and Gelad 1980) and became hot since an implementation of the framework (Itti and Niebur 1998). Since the pioneering work (Liu et al. 2011), saliency analysis moves from predicting the eye fixation points to detecting the salient objects which has been attracting more and more research interests in computer vision, driven by real applications such as image recognition, search, and editing, and is also the interest of this paper. The readers may refer to Borji et al. (2015) and Borji and Itti (2013) for the comprehensive survey and review. In the following, we present a brief review on salient object detection and divide existing algorithms into two categories: unsupervised and supervised.

2.1 Unsupervised

Most unsupervised salient object detection algorithms follow the center-surrounding contrast framework (Itti and Niebur 1998), where different kinds of features are combined according to the feature integration theory (Treisman and Gelad 1980). As the early efforts, contrast in the pixel (patch) level has been widely studied, from multi-scale analysis, e.g., Liu and Gleicher (2006) and Liu et al. (2011), to effective contrast measures, e.g., discriminant center-surround hypothesis (Gao et al. 2007; Gao and Vasconcelos 2007) and feature statistics for computing the center-surround divergence (Klein and Frintrop 2011).

The region-based solution dominates recent efforts, because of its superior performance over the pixel-based solution. The representative works include regional contrast with multi-level segmentation (Jiang et al. 2011), optimized hierarchical saliency map combination (Yan et al. 2013), cost-sensitive SVM for measuring the separability of a center region w.r.t. its surroundings (Li et al. 2013), and so on. Beyond the center-surrounding contrast which is computed between the center and its local context, global contrast, initially studied in the patch-based solution (Margolin et al. 2013), computes the contrast of a region by comparing it with all other part in the image in various forms (Cheng et al. 2014; Perazzi et al. 2012). Global uniqueness, an alternative form of global contrast, is exploited, e.g., global color and texture uniqueness (Scharfenberger et al. 2013; Shi et al. 2013), low-rank matrix factorization (Shen et al. 2012; Zou et al. 2013; Peng et al. 2013) in which the low-rank matrix corresponds to the background regions while sparse noises are indications of salient regions. The soft image segmentation, obtained using a Gaussian Mixture Model, is adopted in Cheng et al. (2013) to address the boundary problem in hard image segmentation based salient object detection.

In addition to uniqueness and contrast, many other priors are also explored for saliency computation. Central prior, i.e., the salient object usually lies in the center of an image, is investigated in Jiang et al. (2011) and Wang et al. (2012a). Object prior, such as connectivity prior (Vicente et al. 2008), concavity context (Lu et al. 2011), and autocontext cue (Wang et al. 2011), backgroundness prior (Wei et al. 2012; Yang et al. 2013; Li et al. 2013; Jiang et al. 2013a), generic objectness prior (Chang et al. 2011; Jiang et al. 2013c; Jia and Han 2013), and background connectivity prior (Zou et al. 2013; Zhang and Sclaroff 2013; Zhu et al. 2014) are also studied for saliency computation. The object-sensitive descriptors introduced in this paper are related to the object prior, but our approach is more focused on forming a descriptor and then combining them as the input of a regressor to learn a saliency map instead of giving a score in an unsupervised manner.

Some salient object detection algorithms are developed beyond the center-surrounding framework. Submodular salient object detection (Jiang and Davis 2013d) connects salient object detection with the submodular facility location problem. The Bayesian framework (Rahtu et al. 2010; Xie et al. 2013), partial differential equation (Liu et al. 2014), and spectral analysis (Achanta et al. 2009), are introduced. Example-based approaches, searching for similar images of the input, are developed for salient object detection (Marchesotti et al. 2009; Wang et al. 2011). Besides in an RGB image, detecting salient objects from stereopsic image pairs (Niu et al. 2012) and RBGD images (Desingh 2013) offering the depth information, and from light fields (Li et al. 2014), are also studied.

2.2 Supervised

The supervised solutions learn a salient object detector from the training data, where the salient object is given as a binary map or with a bounding box. There are two basic supervision ways: early fusion and later fusion.

Early fusion directly learns the saliency map from the raw features. The conference version (Jiang et al. 2013b) of this paper belongs to this category, and is the first successful approach demonstrating the state-of-the-art performance for salient object detection. The early attempt (Mehrani et al. 2010), which extracts the non-contrast features (similar to the objectness feature introduced in this paper) and learns a regressor from the features to the saliency scores, provides neither a systematical study nor any insight and thus no promising result is achieved. In contrast, our approach shows that the contrast and backgroundness descriptors as well as multi-scale analysis take effects as well, and inspires many subsequent studies. For instance, a boosted decision tree is learnt to generate an initial saliency map (Kim et al. 2014) with which a color transform algorithm is adopted to generate the final saliency map. Using the CNN features (Li and Yu 2015) is shown to be able to enhance the performance of our approach. The followup works, such as Chen et al. (2015), Li et al. (2015), and Li and Yizhou (2016), extend supervised feature integration using deep learning, which jointly learns feature representation and feature integration for saliency computation.

Late fusion learns to combine saliency maps, which are computed in an unsupervised manner or obtained from other algorithms, to produce the final saliency map. The early work (Liu et al. 2011) adopts conditional random fields to optimally combine saliency maps as well as the spatial smoothness. Later a mixture of linear SVMs (Khuwuthyakorn et al. 2010) and a large margin framework (Lu et al. 2014) are instead exploited to learn the combination. Besides saliency maps, generic objectness is used to help compute the saliency map. For example, generic objectness (Alexe et al. 2012) is fused with saliency maps using a graphical model (Chang et al. 2011) in an unsupervised manner. Object proposals over windows are fused to generate the saliency map using the adaptive averaging scheme (Li et al. 2014). A saliency map is computed out of the results from classifying windows either as background (non-salient) or as objects (salient) (Moosmann et al. 2006), which can be viewed as a form of late fusion. The supervised approach is explored in the task of outputting a bounding box, to indicate the salient object, e.g., random forest is used to regress the box from the saliency map (Wang et al. 2012a).

3 Preliminaries

3.1 Salient Objection Detection

The concept *salient object* includes two perspectives: *object* and *visual saliency* (or *visual attention*). The object is generic and means all types of objects, rather than a specific object (e.g., face or human). Visual saliency is a process of first distributing uniformly the attention over the visual scene and then concentrating attention to a specific area of the visual scene, which is regarded to be visually salient.

Salient object detection is a process of identifying the most salient objects from an image. The detection result may be

(1) a bounding box indicating the salient objects, termed as salient object localization, (2) a soft map indicating the degree that each pixel belongs to the salient object, termed as saliency map prediction, and (3) a binary map, indicating if a pixel belongs to the salient object, termed as salient object cut. This paper is interested in saliency map prediction that is widely studied in the saliency detection research field, and also shows the superior salient object cut performance achieved based on our predicted saliency map.

3.2 Feature Integration

Feature integration theory is a theory of attention developed in 1980 by Anne Treisman and Garry Gelade (1980). It posits that different kinds of attention are responsible for binding various features into consciously experienced wholes. The computational model (Itti and Niebur 1998) represents the input image from the color, intensity and orientation channels, computes three conspicuity (saliency) maps using center-surround differences, and combines them heuristically together to form the final master saliency map. Most subsequent salient object detection algorithms follow the feature integration framework and the center-surround difference scheme, and then design other cues besides color, texture and orientation. Our approach is a supervised extension of feature integration, which is the first method to demonstrate the success of the supervised solution in salient object detection. In addition, the experimental results indicate that the center-surround difference or the contrast is not the most significant and our proposed image-specific backgroundness and objectness descriptors are more important.

4 Model

The pipeline of computing the saliency map for an image is presented in Fig. 1. It consists of (1) multi-level segmentation, which is generated in the graph-based approach (Felzenszwalb and Huttenlocher 2004), (2) saliency computation in each level, which is computed using the proposed discriminative regional feature integration approach, and (3) multi-level saliency fusion using a linear combinator.

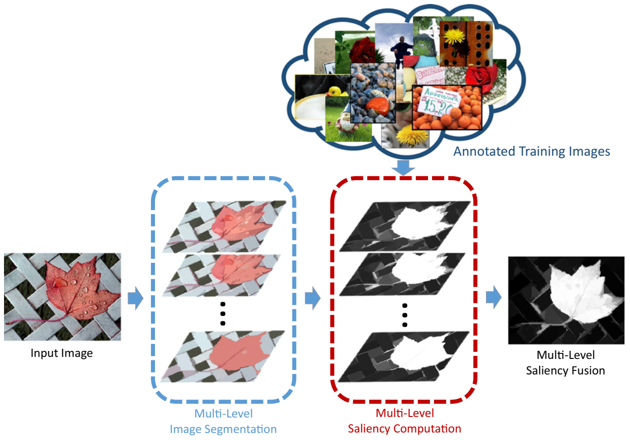


Fig. 1 The pipeline of multi-level discriminative regional feature integration (DRFI). It consists of (1) multi-level segmentation, (2) saliency computation in each level, and (3) multi-level saliency fusion

4.1 Discriminative Regional Feature Integration

Given an image I, we segment it into a set of K superpixels $\{R_1, R_2, \ldots, R_K\}$ and each region R is represented by a saliency feature vector $\mathbf{x} = [\mathbf{x}_c^T\ \mathbf{x}_b^T\ \mathbf{x}_o^T]^T$, consisting of three kinds of feature vectors, a contrast descriptor $\mathbf{x}_c$, a backgroundness descriptor $\mathbf{x}_b$, and a objectness descriptor $\mathbf{x}_o$. Our approach adopts a random forest regressor, $a = f(\mathbf{x})$, to map the descriptor to a saliency value a, assigned to the corresponding region R and thus each pixel in the region R, resulting a saliency map $\mathbf{A}$ for the image I.

4.2 Multi-Level Fusion

A single image segmentation might not produce reliable superpixels for computing the saliency map in the testing phrase. For example, the superpixel might be across the boundary of the salient object and is composed of both salient object and background pixels; the superpixel might be too small and does not contain enough characters to judge if it belongs to a salient object. We propose to explore multi-level segmentations, $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}$, which are generated using the (unsupervised) graph-based image segmentation approach (Felzenszwalb and Huttenlocher 2004) with M groups of different parameters. $\mathcal{S}_m = \{R_1^m, R_2^m, \ldots, R_{K_m}^m\}$ is the mth-level segmentation. Let $\mathbf{A}_m$ be the saliency map computed from the mth segmentation $\mathcal{S}_m$ using discriminative regional feature integration.

We generate the final saliency map by fusing the M saliency maps together, $\mathbf{A} = g(\mathbf{A}_1, \ldots, \mathbf{A}_M)$, where $g(\cdot)$ is a combinator function. We use a linear combinator, $\mathbf{A} = \sum_{m=1}^{M} w_m \mathbf{A}_m$, which already performs well in our experiment.

5 Training

The training data consists of N pairs of $(I, \mathbf{A}^*)$, where $\mathbf{A}^*$ is the ground truth of the salient object, indicating which pixels belong to the salient object.

5.1 Model Learning

There are two training problems: learn the random forest regressor for discriminative regional feature integration, and the linear combinator function for multi-level fusion. The linear combinator is relatively easy and learnt in a standard way: The weights are estimated using a least square estimator, i.e., minimizing the sum of the losses ($\sum_{n=1}^{N} \|\mathbf{A}_n^* - \sum_{m=1}^{M} w_m \mathbf{A}_{nm}\|_F^2$), where $\{\mathbf{A}_n^*\}_{n=1}^{N}$ are the ground truth saliency maps of the N training images.

The random forest regressor requires the region-level ground-truth: $\{(R_k, a_k)\}_{k=1}^{K}$. Once the ground-truth is constructed and each region R_k is described by a saliency feature vector $\mathbf{x}_k$, we adopt the standard learning algorithm to train a random forest regressor. It is known that constructing a reliable ground truth is critical to learn a high-quality random regressor. We propose an effective approach for reliable region-level training sample construction.

5.2 Reliable Region-Level Training Set Construction

The initial way of constructing the region-level ground truth is to build multi-level segmentation by directly using the graph-based image segmentation approach (Felzenszwalb and Huttenlocher 2004), which is the same with the testing stage. We observe that many resulting superpixels are not pure, across the boundary of the salient object and the background and thus not reliable for training the random forest regressor. There would not be enough training samples left if discarding all those impure superpixels.

We propose to adopt a learning approach to construct multi-level segmentation. We follow the graph-based image segmentation framework, and replace the pairwise similarity over two adjacent regions (R_i, R_j) with a score $s(a_i, a_j)$, indicating if they both belong to the salient object or the background. We learn the score $s(a_i, a_j)$ using a boosted decision tree classifier from a set of positive pairs, $\{(R_i, R_j)|a_i = a_j\}$ and negative pairs $\{(R_i, R_j)|a_i \neq a_j\}$. The positive and negative pairs are constructed from an over-segmentation $\mathcal{S}_0$, containing a large number of superpixels created using (Felzenszwalb and Huttenlocher 2004), so that almost all the superpixels are not across the object and the background. The adjacent superpixel pairs with the same saliency label form the positive pairs, and those with different saliency labels form the negative pairs. The three descriptors of each superpixel (presented in Sect. 6) as well as additional the contrast between the two superpixels and geometry features (the details are given in the supplementary material) describing the boundary of the two superpixels are used to represent a pair of superpixels for learning a boosted decision tree.

The multi-level segmentations $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}$ are sequentially constructed. $\mathcal{S}_1$ is constructed from $\mathcal{S}_0$: build an affinity graph by connecting the adjacent segments in $\mathcal{S}_0$ and setting the weight over each edge (R_i, R_j) to be the score $s(a_i, a_j)$, and conduct the graph-based segmentation algorithm (Felzenszwalb and Huttenlocher 2004) to group similar segments together. During the grouping process, there is a parameter k specifying the allowed minimum size of the generated regions (Felzenszwalb and Huttenlocher 2004). We sequentially do the process to generate $\{\mathcal{S}_2, \ldots, \mathcal{S}_M\}$ with increasingly larger $\{k_2, \ldots, k_M\}$[1]. To avoid too small number of superpixels in one level, we discard $\mathcal{S}_i$ if $\frac{|\mathcal{S}_0|}{|\mathcal{S}_i|} > 0.6$, where $|\cdot|$ denotes the number of superpixels.

Finally, we remove unconfident regions. A region is regarded to be unconfident if the number of pixels belonging to the salient object and the background are both smaller than 80% of the total number of pixels in the region. The saliency score of the remaining superpixels is set as 1 or 0 accordingly. In experiments we find that few regions of all the training examples, around 6%, are unconfident.

6 Features

In this section, we present three regional saliency features: regional contrast descriptor $\mathbf{x}_c$, image-specific backgroundness descriptor $\mathbf{x}_b$, and regional objectness descriptor $\mathbf{x}_o$, resulting in a 93-dimensional ($29 + 29 + 35$) feature vector for each region. The implementation[2] shows more details about computing them.

6.1 Contrast Descriptor

The regional contrast descriptor is defined globally: compare the region with all the other regions in the image. We describe the region using the color and texture features. The color features consist of the average values and a histogram with 256 bins in the RGB, HSV, and L*a*b* color spaces. The texture features consist of 15 absolute responses from the LM filters, a max-response histogram with 15 bins from the LM filters, and a histogram with 256 bins from the LBP feature. There are totally 9 groups of features: $\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{r}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_5\}$. The details are depicted in Table 1.

The contrast descriptor of the region R in the mth-level segmentation $\mathcal{S}_m$ is computed as

$$\mathbf{x}_c = \sum_{R' \in \mathbf{S}_m, R' \neq R} \alpha(R') W(\mathbf{p}, \mathbf{p}') D(\mathbf{v}, \mathbf{v}'). \tag{1}$$

Here $\mathbf{p}$ ($\mathbf{p}'$) is the mean position of the region R (R'). $\mathbf{v}$ ($\mathbf{v}'$) is the feature (the concatenation of 9 groups of features) of the region R (R'). $\alpha(R')$ is the normalized area of the region R', introduced to account for the irregular shape of a

[1] They are empirically set to range from 5 to 1800 with increasingly larger gaps. Check our code at https://github.com/playerkk/drfi_matlab for more details.

[2] http://supermoe.cs.umass.edu/~hzjiang/drfi/

Table 1 Color and texture features describing the visual characteristics of a region which are used to compute the regional feature vector

	Color and texture features		Differences of features		Contrast	Backgroundness
	Features	Dim	Definition	Dim		
$\mathbf{a}_1$	Average RGB values	3	$d(\mathbf{a}_1^{R_i}, \mathbf{a}_1^S)$	3	$c_1 - c_3$	$b_1 - b_3$
$\mathbf{h}_1$	RGB histogram	256	$\chi^2(\mathbf{h}_1^{R_i}, \mathbf{h}_1^S)$	1	c_4	b_4
$\mathbf{a}_2$	Average HSV values	3	$d(\mathbf{a}_2^{R_i}, \mathbf{a}_2^S)$	3	$c_5 - c_7$	$b_5 - b_7$
$\mathbf{h}_2$	HSV histogram	256	$\chi^2(\mathbf{h}_2^{R_i}, \mathbf{h}_2^S)$	1	c_8	b_8
$\mathbf{a}_3$	Average L*a*b* values	3	$d(\mathbf{a}_3^{R_i}, \mathbf{a}_3^S)$	3	$c_9 - c_{11}$	$b_9 - b_{11}$
$\mathbf{h}_3$	L*a*b* histogram	256	$\chi^2(\mathbf{h}_3^{R_i}, \mathbf{h}_3^S)$	1	c_{12}	b_{12}
$\mathbf{r}$	Absolute response of LM filters	15	$d(\mathbf{r}^{R_i}, \mathbf{r}^S)$	15	$c_{13} - c_{27}$	$b_{13} - b_{27}$
$\mathbf{h}_4$	Max response histogram of the LM filters	15	$\chi^2(\mathbf{h}_4^{R_i}, \mathbf{h}_4^S)$	1	c_{28}	b_{28}
$\mathbf{h}_5$	Histogram of the LBP feature	256	$\chi^2(\mathbf{h}_4^{R_i}, \mathbf{h}_5^S)$	1	c_{29}	b_{29}

$d(\mathbf{a}_1, \mathbf{a}_2) = [|a_1 - a'_1| \ \cdots \ |a_t - a'_t|]^T$ where t is the number of elements in the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$. $\chi^2(\mathbf{h}_1, \mathbf{h}_2) = \sum_{i=1}^b \frac{2(h_{1i}-h_{2i})^2}{h_{1i}+h_{2i}}$ with b being the number of histogram bins. The last two columns denote the symbols for the contrast and image-specific backgroundness descriptors. In the definition, S corresponds to R_j for the regional contrast descriptor and B for the image-specific backgroundness descriptor, respectively

region. $W(\mathbf{p}, \mathbf{p}')$ is the spatial weight between regions R and R' and computed as $\exp(-\frac{1}{2\sigma^2}\|\mathbf{p} - \mathbf{p}'\|_2^2)$, resulting in the larger contribution to the contrast from the nearer regions. $D(\mathbf{v}, \mathbf{v}')$ is the concatenation of the differences of 9 groups of features. The difference of the non-histogram features is defined as $d(\mathbf{a}, \mathbf{a}') = [|a_1 - a'_1| \ \cdots \ |a_t - a'_t|]^T$, where t is the dimension of the vector $\mathbf{a}$. The difference of the histogram features is defined as $\chi^2(\mathbf{h}, \mathbf{h}') = \sum_{i=1}^b \frac{2(h_i - h'_i)^2}{h_i + h'_i}$, where b is the number of histogram bins. $D(\mathbf{v}, \mathbf{v}')$ and accordingly the contrast descriptor are 29-dimensional vectors.

6.2 Image-Specific Backgroundness Descriptor

The study on the MSRA-B data set with 5000 images shows that 98% among the pixels in the 15-pixel wide narrow border region of the image, called pseudo-background, belongs to the background. On the other hand, in an image, the regions with the appearance similar to the pseudo-background are likely to belong to the background.

Motivated by these two observations, we introduce an image-dependent backgroundness descriptor for a region R by comparing it with the pseudo-background region B, which is computed as

$$\mathbf{x}_b = D(\mathbf{v}, \mathbf{v}'), \qquad (2)$$

where $\mathbf{v}$ ($\mathbf{v}'$) is the feature of the region R (the pseudo-background region B), and the feature and the definition $D(\cdot, \cdot)$ are the same to those described in Sect. 6.1. The backgroundness descriptor is a 29-dimensional vector, and the details are given in Table 1.

6.3 Objectness Descriptor

The objectness descriptor[3] aims to characterize the common property of the salient object and the background in various images, and consists of geometric and appearance features. The geometric features, summarized in Table 2, a set of features describing the size and the position of a region, are useful to describe the spatial and geometric prior of the salient object and the background because the salient object, for example, usually lies in the center of the image while the background usually scatters over the entire image. The appearance features are about the variances of colors and textures in a region, which is helpful because, for example, the background usually has homogeneous color or similar texture pattern. We obtain a 35-dimensional regional property descriptor. The details are given in Table 2.

7 Experiments

In this section, we empirically analyze our proposed approach and present the comparisons with state-of-the-art methods on benchmark data sets.

7.1 Data Sets

MSRA-B[4] This data set (Liu et al. 2011) includes 5000 images, originally containing labeled rectangles from nine

[3] Objectness is a feature vector in this paper and different from the concept objectness in Alexe et al. (2012) where objectness is used to quantify how likely it is for an image window to contain an object of any class.

[4] http://research.microsoft.com/en-us/um/people/jiansun/

Table 2 The regional objectness descriptor

Description	Notation	Dim
Average normalized x coordinates	o_1	1
Average normalized y coordinates	o_2	1
10th percentile of the normalized x coord.	o_3	1
10th percentile of the normalized y coord.	o_4	1
90th percentile of the normalized x coord.	o_5	1
90th percentile of the normalized y coord.	o_6	1
Normalized perimeter	o_7	1
Aspect ratio of the bounding box	o_8	1
Variances of the RGB values	$o_9 - o_{11}$	3
Variances of the L*a*b* values	$o_{12} - o_{14}$	3
Variances of the HSV values	$o_{15} - o_{17}$	3
Variance of the response of the LM filters	$o_{18} - o_{32}$	15
Variance of the LBP feature	o_{33}	1
Normalized area	o_{34}	1
Normalized area of the neighbor regions	o_{35}	1

The abbreviation coord. indicates coordinates

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 segment the salient object (contour) within the user-drawn rectangle to obtain binary masks. The ASD data set (Achanta et al. 2009) is a subset (with binary masks) of MSRA-B, and thus we no longer make evaluations on it.

iCoSeg[5] This is a publicly available co-segmentation data set (Batra et al. 2010), including 38 groups of totally 643 images. Each image is along with a pixel-wise groundtruth annotation, which may contain one or multiple salient objects. In this paper, we use it to evaluate the performance of salient object detection.

SED[6] This data set (Alpert et al. 2007) contains two subsets: SED1 that contains 100 images containing only one salient object and SED2 that contains 100 images containing exactly two salient objects. Pixel-wise groundtruth annotations for the salient objects in both SED1 and SED2 are provided. We only make evaluations on SED2. Similar to the larger MSRA-B dataset, only a single salient object exists in each image in SED1, where state-of-the-art performance was reported in our conference version (Jiang et al. 2013b).

ECSSD[7] To overcome the weakness of existing data set such as ASD, in which background structures are primarily simple and smooth, a new data set denoted as Extended Complex Scene Saliency Dataset (ECSSD) is proposed recently in Yan et al. (2013). It contains 1000 images with diversified patterns in both foreground and background, where many semantically meaningful but structurally complex images are available. Binary masks for salient objects are produced by 5 subjects.

DUT-OMRON[8] This dataset is introduced to evaluate salient object detection algorithms on images with more than a single salient object and relatively complex background. It contains 5168 high quality natural images, where each image is resized to have a maximum side length of 400 pixels. Annotations are available in forms of both bounding boxes, and pixel-wise binary object masks which we use for the evaluation in our experiments.

HKU-IS[9] The data set (Li and Yu 2015) contains 4447 images with pixelwise annotation of salient objects. It is divided into three parts, 2500 images for training, 500 images for validation, and the remaining 1447 images for testing. We evaluate the performance over the 1447 testing images.

Training set We randomly sample 3000 images from the MSRA-B data set to train our model. Five-fold cross validation is run to select the parameters. The remaining 2000 images are used for testing. 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 can help test the adaptability of the model trained from one data set to other different data sets and avoid the model overfitted to a specific one.

7.2 Evaluation Metrics

We have evaluated the performance using the measures used in Borji et al. (2015) based on the overlapping area between groundtruth annotation and saliency prediction, including the PR (precision-recall) curve, the ROC (receiver operating characteristic) curve, the AUC (Area Under ROC Curve) score, and the mean absolute error (MAE) 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. For a grayscale saliency map, where the pixel values are in the range [0, 255], we vary the threshold from 0 to 255 to obtain a series of salient object segmentations. The PR curve is created by computing the precision and recall values at each threshold. In the following, we present the comparison in terms of the PR curve and the AUC score. Other results can be found from the supplementary material.

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

[6] http://www.wisdom.weizmann.ac.il/∼vision/Seg_Evaluation_DB/

[7] http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency

[8] http://ice.dlut.edu.cn/lu/dut-omron/homepage.htm

[9] https://sites.google.com/site/ligb86/mdfsaliency/

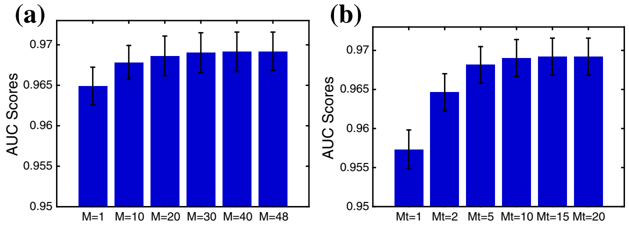


Fig. 2 Empirical analysis of multi-level segmentation for training and testing in terms of the AUC scores based on five-fold cross-validation of the training set. **a** AUC scores versus #(levels of segmentation) used in training, where #(levels of segmenation) is 15 in the testing phase. **b** AUC scores versus #(levels of segmentation) used in testing, where #(levels of segmentation) is 48 in the training phase

7.3 Empirical Analysis

When training a random forest, several parameters, e.g., the number of decision trees and the number of features for node splitting, should be determined. We studied how they affect the five-fold cross-validation performance over the training set. The results suggest that the random forest regressor with 200 trees[10] trained with randomly sampled 15 features[11] for node splitting reaches a balance between the efficiency and the effectiveness. In the following, we present detailed studies for other factors.

7.3.1 Multi-Level Segmentation

We present the empirical results, which is obtained by running five-fold cross-validation on the training data, to show how multi-level segmentation affects the performance.

Figure 2a shows the five-fold validation performance over the training set when different number of segmentations are used to train the random forest regressor with 200 decision trees. As expected, the larger number of segmentations, thus more training data, yield better validation performance. In our experiment, we choose 48-level segmentations, resulting in around 1.7 million training region samples, to train the random forest regressor.

Figure 2b shows the five-fold validation performance over the training set when different number of segmentations are used in the testing stage. It can be seen that the AUC score increases when more levels of segmentations are adopted. The reason is that more layers increase the chance that some regions cover the most (even entire) part of an object. Considering both validation accuracy and time cost, we choose 15-level segmentations for testing in our experiments.

7.3.2 Training with Unsupervised and Supervised Segmentation

We report the results for the two cases: unsupervised segmentation is used to generate training examples and supervised segmentation is used to generate training examples. The performance over the validation dataset and the performances over the testing datasets are shown in Table 3. We can see that the performance for salient object detection with supervised segmentation performs slightly better in terms of AUC and the performance for salient object cut is much better in terms of F_β. The experimental results validate that the random forest regressor learnt from the noisy superpixels are less reliable than the random forest regressor learnt from the cleaned superpixels using supervised segmentation, which is consistent to the analysis in Sect. 5.2.

7.3.3 Feature Importance

Coarse Importance The saliency features consist of three kinds: the contrast descriptor, the image-specific backgroundness descriptor, and the objectness descriptor. We do the ablation study to show the contribution from each kind of descriptor.

The AUC scores of saliency maps are given in Fig. 3. We have the following overall observations: (1) The contrast descriptor is the least important one[12]; (2) The objectness descriptor plays the most important role; (3) The backgroundness descriptor is in-between. In addtion, rather than considering all the 93 features, we adopt the top 60 features, which occupy around 90% of the energy of total features, for training. Surprisingly, this feature vector performs as well as the entire feature descriptor, even slightly better on DUT-OMRON. Another observation is that the performances without contrast and without backgroundness are different. This indicates that the characteristics of the contrast and backgroundness descriptors are different and there is almost no redundancy between them though their computations are similar: one is defined over a pair of superpixel and the other is defined over a superpixel and the image border region.

We illustrate the most important feature for each kind of descriptor in Fig. 4, where we do not adopt the multi-level fusion enhancement. We can see that the most powerful backgroundness feature provides far less reliable information for salient object detection. By integrating all the weak information, much better saliency maps are achieved.

Fine-Grained Importance We empirically study which dimensions in the 93-dimensional feature vector are the most important for the saliency computation. Figure 5 shows the

[10] We tested the performance of random forests with 5, 10, 50, 100, 150, 200, and 300 trees.

[11] We tested the performance for 3, 5, 10, 15 and 20 features.

[12] Note that this observation holds for salient object detection. It might not hold for other saliency detection task, e.g., eye-fixation prediction, in which the contrast feature is perhaps more important.

Table 3 Performance comparison with unsupervised (unsup.) and supervised (sup.) segmentation adopted in the training stage

	Validation	iCoSeg	ECSSD	DUT-OMRON	SED2	HKU-IS	DUT-OMRON*
AUC							
Unsup.	0.965 ± 0.002	0.952 ± 0.085	0.863 ± 0.260	0.918 ± 0.109	0.910 ± 0.169	0.945 ± 0.064	0.805 ± 0.165
Sup.	0.969 ± 0.002	0.966 ± 0.071	0.868 ± 0.265	0.926 ± 0.116	0.923 ± 0.187	0.952 ± 0.064	0.811 ± 0.183
F_β							
Unsup.	0.730 ± 0.006	0.698 ± 0.116	0.610 ± 0.131	0.538 ± 0.109	0.701 ± 0.135	0.636 ± 0.138	0.353 ± 0.092
Sup.	0.769 ± 0.005	0.769 ± 0.265	0.708 ± 0.315	0.618 ± 0.353	0.758 ± 0.225	0.875 ± 0.181	0.445 ± 0.370

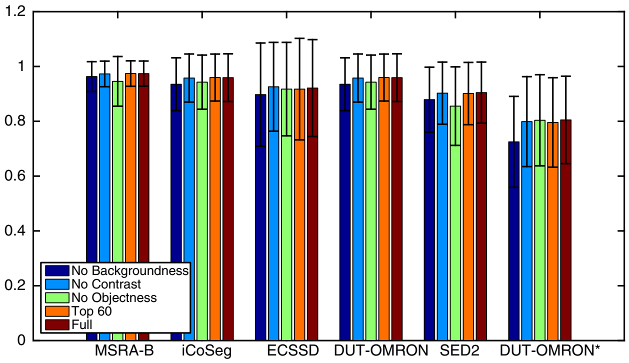


Fig. 3 Feature importance across different data sets. For each data set, we report the AUC scores of saliency maps by removing each kind of descriptor to see the performance drop. Additionally, we also demonstrate the performance exploiting only the top 60 features shown in Fig. 5 (Color figure online)

top-ranked 60 dimensions produced during the training of the random forest regressor.

Out of the top 60 dimensions, 35 dimensions come from the objectness descriptor, which indicates that the objectness descriptor is the most critical one in our feature set, consistent with the above analysis. The highly ranked geometric features o_5, o_3 and o_6, corresponding to the compositional bias of salient objects, are able to record the spatial prior. The importance of variance features o_{12} and o_{26} is related with the background properties. As opposed to our initial guess, the contrast descriptor is the least important.

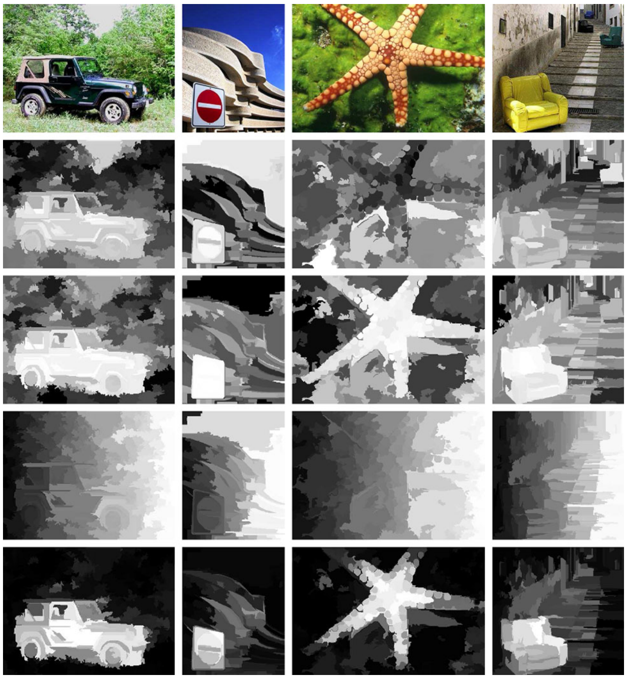


Fig. 4 Illustration of the most important features. From *top* to *bottom*: input images, the most important contrast feature (c_{12}), the most important image-specific backgroundness feature (b_{12}), the most important objectness feature (o_5), and the saliency map of our approach (DRFIs) produced on a single-level segmentation. The *whiter area* indicates the larger response value. It can be seen that while a single feature provides far less reliable information for salient object detection, integrating all the weak information leads to much better saliency maps

7.4 Performance Comparison

We report both quantitative and qualitative comparisons of our approach with state-of-the-art methods. To save the space, we only consider the top four models ranked in the survey (Borji et al. 2012): SVO (Chang et al. 2011), CA (Goferman et al. 2010), CB (Jiang et al. 2011), and RC (Cheng et al. 2014) and recently-developed methods: SF (Perazzi et al. 2012), LRK (Shen et al. 2012), HS (Yan et al. 2013), GMR (Yang et al. 2013), PCA (Margolin et al. 2013), MC (Jiang et al. 2013a), DSR (Li et al. 2013), RBD (Zhu et al. 2014), which are not covered in Borji et al. (2012). Note that we compare our approach with the extended version of RC. In total, we make comparisons with 12 approaches that are all unsupervised. In Sect. 8.5, we will present the results of the algorithms that follow our approach using deep learning. We also report the performance of our DRFI approach with a single level segmentation (DRFIs).

Quantitative Comparison The quantitative comparisons in terms of the AUC score and the PR curve over the seven datasets are shown in Table 4 and Fig. 6, respectively.

As can be seen, our approach (DRFI) consistently outperforms others on all benchmark data sets with large margins in terms of the AUC scores and the PR curve. Specifically, it

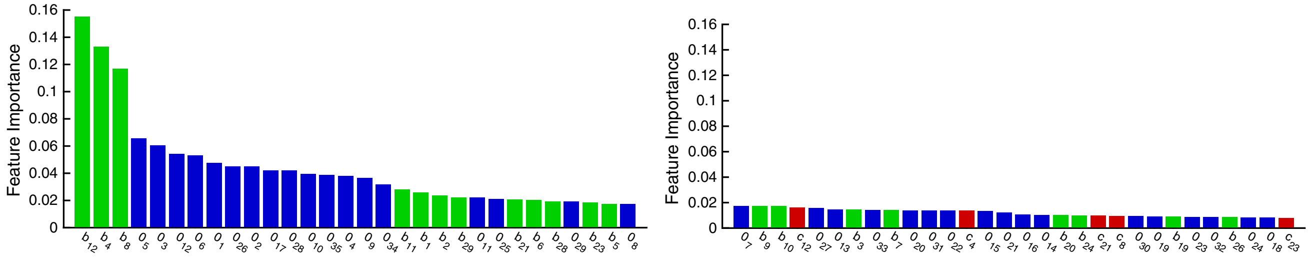


Fig. 5 The most important 60 regional saliency features occupying around 90% of the energy of total features. There are 5 contrast features, 20 image-specific backgroundness features, and 35 objectness features. From *left* to *right*: the first and second 30 important features, respectively. See Tables 1 and 2 for the description of the features

Table 4 Comparison in terms of AUC (larger is better) and the standard deviation for salient object detection

	MSRA-B	iCoSeg	ECSSD	DUT-OMRON	SED2	HKU-IS	DUT-OMRON*
SVO (Chang et al. 2011)	0.915 ± 0.123	0.862 ± 0.183	0.806 ± 0.249	0.873 ± 0.150	0.862 ± 0.168	0.894 ± 0.112	**0.794 ± 0.213**
CA (Goferman et al. 2010)	0.861 ± 0.141	0.831 ± 0.190	0.741 ± 0.240	0.814 ± 0.188	0.861 ± 0.176	0.833 ± 0.159	0.758 ± 0.224
CB (Jiang et al. 2011)	0.935 ± 0.092	0.872 ± 0.153	0.824 ± 0.248	0.833 ± 0.186	0.843 ± 0.208	0.867 ± 0.142	0.620 ± 0.231
RC (Cheng et al. 2014)	0.939 ± 0.091	0.88*3 ± 0.133	0.834 ± 0.248	0.858 ± 0.158	0.844 ± 0.183	0.912 ± 0.101	0.673 ± 0.203
SF (Perazzi et al. 2012)	0.917 ± 0.116	0.908 ± 0.151	0.800 ± 0.234	0.822 ± 0.183	0.882 ± 0.193	0.862 ± 0.141	0.721 ± 0.240
LRK (Shen et al. 2012)	0.934 ± 0.089	0.904 ± 0.128	0.815 ± 0.241	0.866 ± 0.148	0.886 ± 0.161	0.881 ± 0.112	0.758 ± 0.203
HS (Yan et al. 2013)	0.941 ± 0.088	0.904 ± 0.174	0.837 ± 0.249	0.869 ± 0.159	0.844 ± 0.234	0.908 ± 0.104	0.731 ± 0.222
GMR (Yang et al. 2013)	0.950 ± 0.088	0.913 ± 0.139	0.843 ± 0.254	0.855 ± 0.192	0.815 ± 0.242	0.899 ± 0.123	0.622 ± 0.266
PCA (Margolin et al. 2013)	0.941 ± 0.081	0.889 ± 0.133	0.817 ± 0.242	0.886 ± 0.134	**0.903 ± 0.155**	0.900 ± 0.098	0.771 ± 0.198
MC (Jiang et al. 2013a)	**0.954 ± 0.066**	0.908 ± 0.110	**0.849 ± 0.254**	0.888 ± 0.143	0.876 ± 0.166	**0.925 ± 0.080**	0.705 ± 0.200
DSR (Li et al. 2013)	0.950 ± 0.076	0.915 ± 0.112	0.844 ± 0.249	0.893 ± 0.132	0.886 ± 0.151	**0.928 ± 0.083**	0.769 ± 0.196
RBD (Zhu et al. 2014)	0.947 ± 0.077	**0.939 ± 0.090**	0.833 ± 0.246	**0.893 ± 0.141**	0.869 ± 0.185	0.915 ± 0.098	0.774 ± 0.204
DRFIs	**0.952 ± 0.062**	**0.939 ± 0.082**	**0.848 ± 0.252**	**0.902 ± 0.119**	**0.892 ± 0.176**	0.910 ± 0.089	**0.793 ± 0.175**
DRFI	**0.972 ± 0.052**	**0.966 ± 0.071**	**0.868 ± 0.265**	**0.926 ± 0.116**	**0.923 ± 0.187**	**0.952 ± 0.064**	**0.811 ± 0.183**

The best three results are highlighted in bold

improves by 1.89, 2.88, 2.24, 3.70, 2.21, and 2.59% over the best-performing state-of-the-art algorithm according to the AUC scores on MSRA-B, iCoSeg, ECSSD, DUT-OMRON, SED2, and HKU-IS, respectively.

Qualitative Comparison We also provide the qualitative comparisons of different methods in Fig. 7. As can be seen, our approach (shown in Fig. 7 (m) (n)) 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 entire salient object. It is also worth pointing out that our approach performs well when the object touches the image border, e.g., the last and last fourth rows in Fig. 7, even though it violates the pseudo-background assumption.

7.5 Robustness Analysis

The image-specific backgroundness and objectness descriptors, especially the geometric properties, play important roles in our approach, as studied in Sect. 7.3.3. In some images, the pseudo-background assumption may not always hold. Additionally, the distributions of salient objects, which are explored in the objectness descriptor, may be different from our training set. It is required to test whether our approach can still perform well on these challenging cases. To this end, we select 635 images from the DUT-OMRON dataset (we call it DUT-OMRON*), where salient objects touch the image border and are far from the image center. Quantitative comparisons with the state-of-the-art approaches are presented in Table 4 and Fig. 6. The details of the challenging dataset are available in our project website http://supermoe.cs.umass.edu/~hzjiang/drfi and the supplementary material.

Not surprisingly, the performances of all approaches decline. Our approach DRFI still significantly outperforms other methods in terms of the PR curve and the AUC score. In specific, DRFI is better than the second-best method by around 2.14%. This justifies the robustness of the discriminative feature integration.

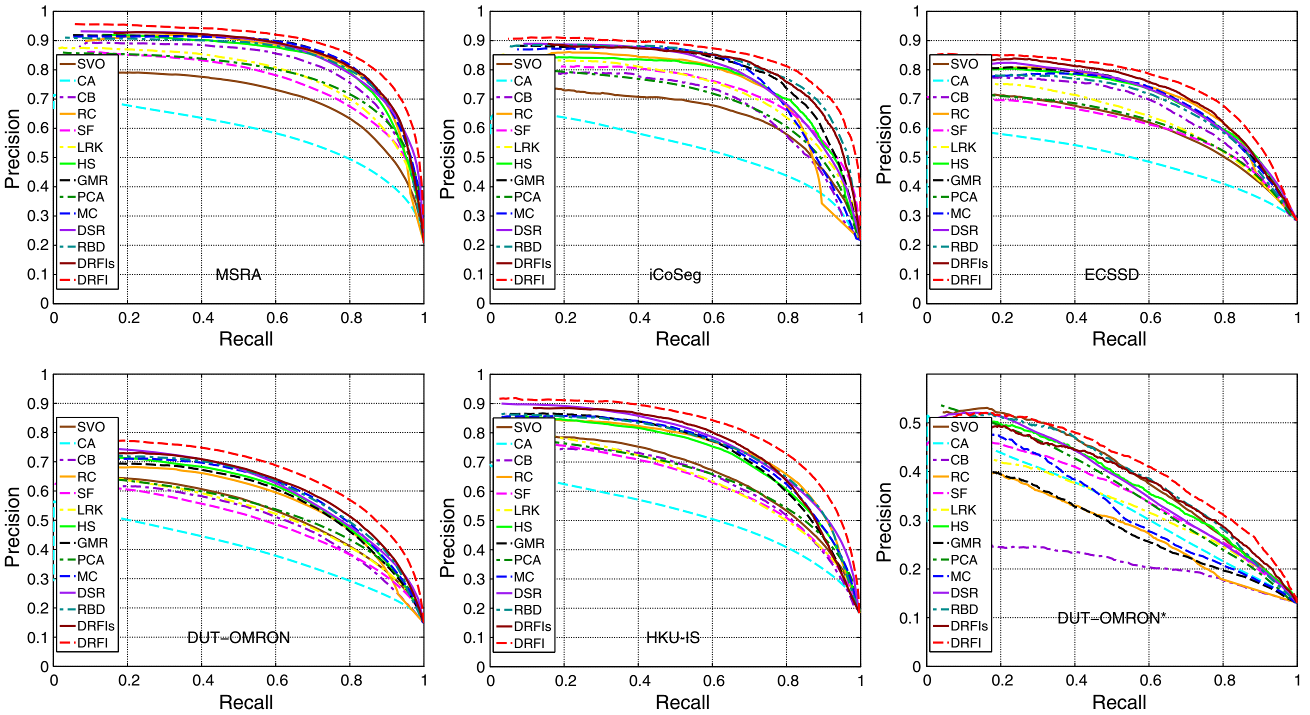


Fig. 6 Quantitative comparisons of saliency maps produced by different approaches on different data sets in terms of the PR curves (Color figure online)

7.6 Efficiency

Since the computation on each level of the multiple segmentations is independent, we utilize the multi-thread (8 threads in our experiment) technique to accelerate the computation. Table 5 summarizes the running time of different approaches, tested on the MSRA-B data set with a typical 400 × 300 image using a PC with an Intel i5 CPU of 2.50 GHz and 8 GB memory. As we can see, our approach can run as fast as most existing approaches. If used as a pre-processing step for an application, e.g., picture collage, our approach will not harm the user experiences.

For training, it takes around 24 h with around 1.7 million training samples. As training each decision tree is also independent to each other, the parallel computing technique can be exploited for acceleration.

7.7 Failure Examples

Our approach, like other algorithms may fail on extremely cluttered scenes. Figure 8 shows three failure examples. In the first failure example (a), high saliency values are assigned to the texture areas in the background as they are distinct in terms of either contrast or background features. In the second example (b), the salient object has similar color with the background and occupies a large portion of the image, making it challenging to generate good detection result. In the third case (c), our approach highlights the flag, which is indeed salient, but not labeled as salient in the the ground truth labeling. The reason that our approach does not highlight the statue (the ground truth) is that the statue violates the pseudo-background assumption and occupies a large portion as well.

We also quantitatively analyze the average AUC and MAE scores of our approach against the normalized size of salient objects on all the six testing datasets. It can be observed from Fig. 9 that the performance is relatively steady when the object is small (less than 50% of the image) but quickly decreases when the dominant salient object becomes larger. One possible reason might be that when the graph-based segmentation algorithm fails to group pixels of a large salient object together and thus the random forest makes incorrect estimations.

7.8 Salient Object Cut

We present the salient object cut results using the SaliencyCut algorithm (Cheng et al. 2014) over the saliency maps generated from our approach and other competitive algorithms. The evaluation criterion is the average F_β score,

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}. \quad (3)$$

As suggested in Cheng et al. (2014), we set $\beta^2 = 0.3$ to give the precision score larger weight.

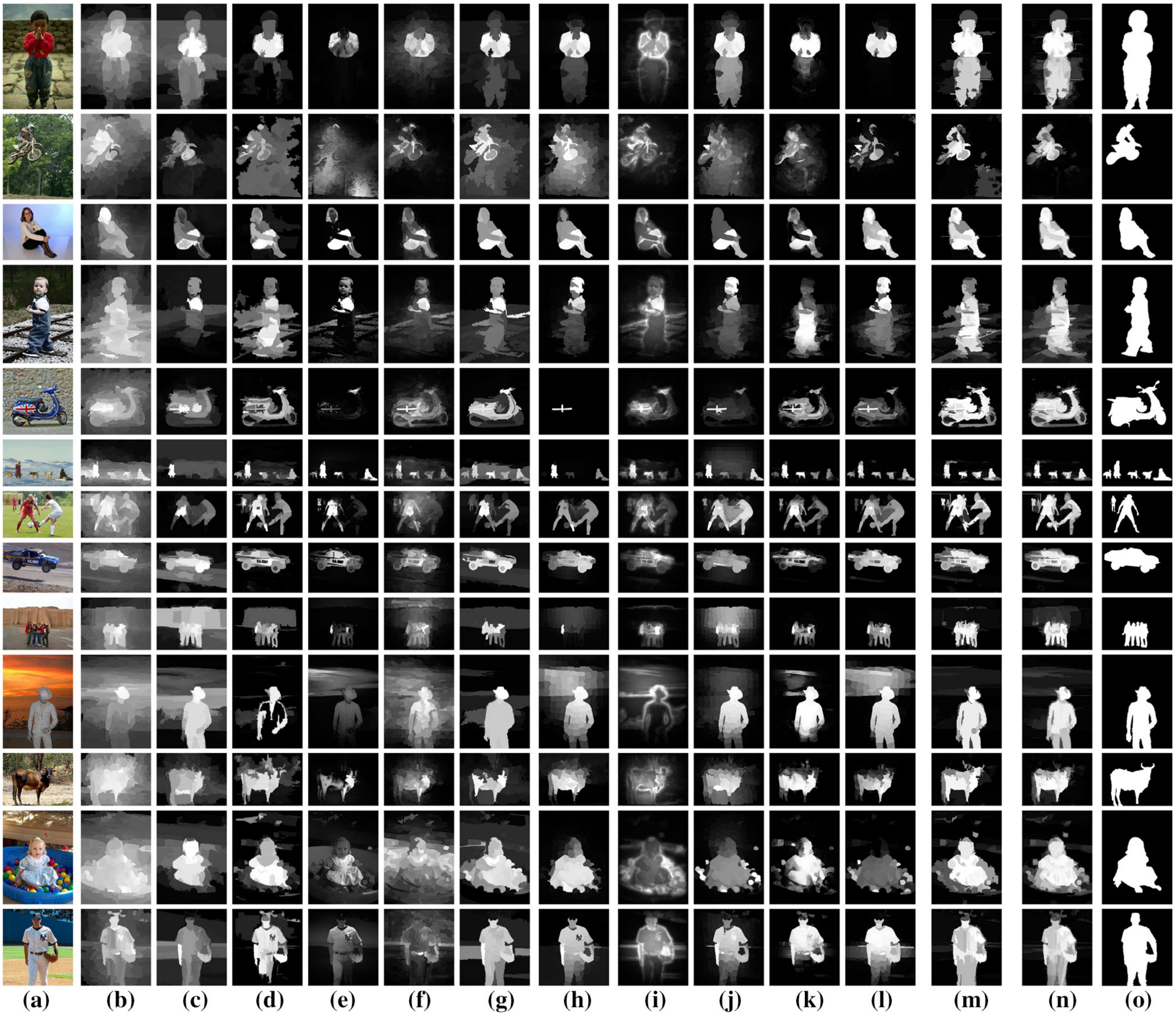


Fig. 7 Visual comparison of the saliency maps. These examples are sampled from the challenging images, such as with salient objects touching the image border, with complex structures, with multiple salient objects, etc. Our method (DRFI) consistently generates better saliency maps. **a** input, **b** SVO, **c** CB, **d** RC, **e** SF, **f** LRK, **g** HS, **h** GMR, **i** PCA, **j** MC, **k** DSR, **l** RBD, **m** DRFIs, **n** DRFI, **o** GT

Table 5 Comparison of running time

Method	SVO	CA	CB	RC	SF	LRK	HS
Time (s)	56.5	52.3	1.40	0.138	0.210	11.5	0.365
Code	M + C	M + C	M + C	C	C	M + C	EXE
Method	GMR	PCA	MC	DSR	RBD	DRFIs	DRFI[a]
Time (s)	1.16	2.07	0.129	4.19	0.267	0.183	0.418
Code	M + C	M + C	M + C	M + C	M	C	C

M indicates the code is written in MATLAB and EXE is corresponding to the executable. ([a] 8 threads are used for acceleration.)

Quantitative Comparison The average F_β scores are presented in Table 6. It can be observed that our proposed DRFI approach is ranked as the best one for four out of seven benchmark datasets. It performs slightly worse, as the second best one, on the challenging ECSSD, HKU-IS, and DUT-OMRON* datasets. One possible reason might be that images

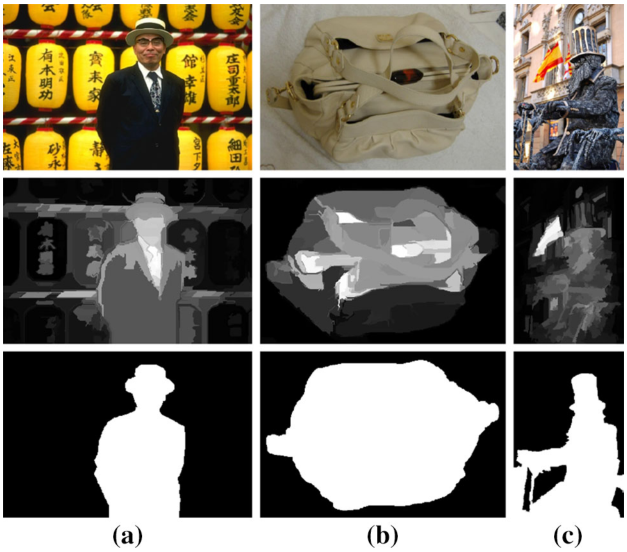


Fig. 8 Failure cases of our approach. The *first*, *second*, and *third rows* show input images, the results of our approach, and the ground-truth saliency maps

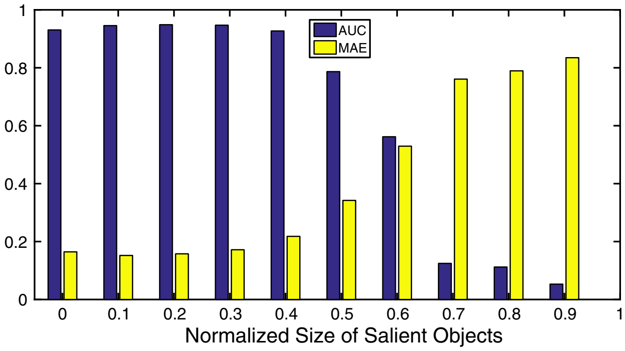


Fig. 9 The performance of our approach versus the normalized size of salient objects all the six testing datasets (Color figure online)

in these three datasets are more structurally complex than others that the initializations to the SaliencyCut algorithm plays less important roles during the iterative refinement of the SaliencyCut algorithm.

Qualitative Comparison Some qualitative comparisons of different salient object cut results are also demonstrated in Fig. 10. In the first two rows, the segmentation results of DRFI are nearly identical to others, where the main differences are on the boundary of the salient objects. In the third row, DRFI performs much better than others, suggesting it provides better initialization to the SaliencyCut algorithm. The last row shows a failure example, where two faces are semantically salient. Due to lack of semantic knowledge, all four approaches fail on this image. One possible solution is to add some features, e.g., face detection results, into regional saliency features.

8 Discussions

8.1 Object Detection

Salient object detection differs from object detection that localizes the position of the object usually in the form of a box at least in two aspects. On the one hand, salient object detection aims to detect a generic object, not a specific object whose class is pre-given. In contrast, object detection usually aims to identify where an object, whose class belongs to a known concept, e.g., dog or face, is. On the other hand, instead of detecting all objects, salient object detection needs to select the objects that attract the most attention.

The objectness descriptor in our approach is different from object proposal (usually conducted as the preprocess of object detection), in which the concept objectness is used to quantify how likely it is for an image window to contain an object of any class (Alexe et al. 2012), and our approach creates a feature vector to characterize a region from the geometric and appearance aspects.

8.2 Figure–Ground Segregation

Figure–ground segregation is the process by which the visual system organizes a visual scene into figures and their backgrounds. Gestalt psychologists were the first to recognize the importance of this problem (Koffka 1935; Rubin 1958). Much of the research on figure–ground segregation has been concerned with identifying the properties that determine which regions will appear as figures. In contrast, salient object detection has a clear definition of the figure: the figure is an object. The study in Kimchi and Peterson (2008) about the relation between figure–ground segmentation and attention (saliency) demonstrates that figure–ground segregation can occur without focal attention.

There are some studies in computer vision, e.g., unsupervised figure–ground segmentation (Yu and Shi 2003), and supervised figure–ground segmentation (Ren et al. 2006; Küttel and Ferrari 2012) to which could be turned from semantic segmentation by simply merging all the object labels into one class and the rest into the background. Salient object detection could be viewed as a kind of semantic segmentation. However, instead of segmenting out all objects, it only concerns about the salient ones. In other words, salient object detection in some sense could be viewed as supervised *salient*-figure and ground segmentation. As a minor difference, the object classes in the datasets used in the research area of salient object detection are not limited and more various while the object classes in supervised figure–ground segmentation or semantic segmentation are usually limited in a small number categories, and in this sense, salient object detection is more challenging.

Table 6 Comparison in terms of the F_β scores (larger is better) and the standard deviation for salient object cut

	MSRA-B	iCoSeg	ECSSD	DUT-OMRON	SED2	HKU-IS	DUT-OMRON*
SVO (Chang et al. 2011)	0.789 ± 0.257	0.661 ± 0.330	0.692 ± 0.299	0.570 ± 0.353	0.695 ± 0.267	**0.866 ± 0.213**	**0.464 ± 0.361**
CA (Goferman et al. 2010)	0.706 ± 0.306	0.624 ± 0.322	0.540 ± 0.336	0.464 ± 0.354	0.497 ± 0.326	0.818 ± 0.267	0.398 ± 0.356
CB (Jiang et al. 2011)	0.816 ± 0.233	0.673 ± 0.327	0.702 ± 0.289	0.548 ± 0.372	0.625 ± 0.352	0.785 ± 0.295	0.286 ± 0.333
RC (Cheng et al. 2014)	**0.831 ± 0.224**	0.718 ± 0.285	**0.717 ± 0.283**	0.577 ± 0.373	0.624 ± 0.335	0.825 ± 0.243	0.349 ± 0.357
SF (Perazzi et al. 2012)	0.702 ± 0.316	0.618 ± 0.331	0.321 ± 0.369	0.334 ± 0.384	0.447 ± 0.368	0.492 ± 0.386	0.256 ± 0.362
LRK (Shen et al. 2012)	0.796 ± 0.248	0.700 ± 0.302	0.644 ± 0.324	0.559 ± 0.364	**0.724 ± 0.245**	**0.879 ± 0.205**	0.409 ± 0.375
HS (Yan et al. 2013)	0.814 ± 0.240	0.701 ± 0.317	**0.712 ± 0.286**	0.569 ± 0.370	0.676 ± 0.298	0.826 ± 0.248	0.404 ± 0.376
GMR (Yang et al. 2013)	0.808 ± 0.240	0.697 ± 0.296	0.669 ± 0.316	0.545 ± 0.383	0.613 ± 0.353	0.788 ± 0.270	0.322 ± 0.367
PCA (Margolin et al. 2013)	0.792 ± 0.252	0.686 ± 0.271	0.651 ± 0.309	0.577 ± 0.350	0.666 ± 0.310	0.857 ± 0.212	**0.445 ± 0.362**
MC (Jiang et al. 2013a)	0.815 ± 0.235	0.701 ± 0.288	0.688 ± 0.316	0.571 ± 0.377	0.604 ± 0.354	0.842 ± 0.229	0.374 ± 0.370
DSR (Li et al. 2013)	0.803 ± 0.247	0.696 ± 0.289	0.629 ± 0.319	0.545 ± 0.376	0.587 ± 0.342	0.778 ± 0.272	0.366 ± 0.371
RBD (Zhu et al. 2014)	0.813 ± 0.251	**0.736 ± 0.279**	0.678 ± 0.310	**0.600 ± 0.363**	0.721 ± 0.265	0.782 ± 0.276	0.443 ± 0.378
DRFIs	**0.843 ± 0.211**	**0.769 ± 0.260**	**0.716 ± 0.304**	**0.622 ± 0.350**	**0.756 ± 0.238**	0.804 ± 0.253	0.444 ± 0.370
DRFI	**0.852 ± 0.200**	**0.769 ± 0.265**	0.708 ± 0.315	**0.618 ± 0.353**	**0.758 ± 0.225**	**0.875 ± 0.181**	**0.445 ± 0.370**

The best three results are highlighted in bold

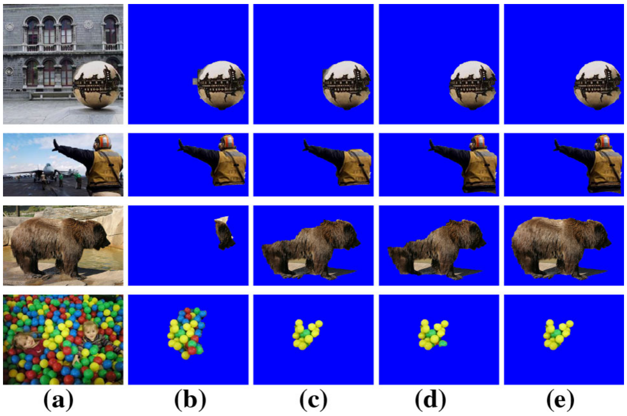


Fig. 10 Salient object cut results with the SaliencyCut algorithm initialized with saliency maps produced by different approaches. **a** input, **b** SVO, **c** RBD, **d** DRFIs, **e** DRFI

In terms of the solution, our approach is inspired by the success of random forest applied to geometric context segmentation (Hoiem et al. 2005). Random forest also shows the powerfulness in other semantic segmentation problems, e.g., object class segmentation (Schroff et al. 2008). The key message of our paper is that saliency besides the object class can also be learnt by adopting a discriminative regressor to combine multiple features. In fact, salient object detection could be solved using the same technique adopted for semantic segmentation, with carefully designed features as done in our paper, or automatically learnt features using deep learning which will be discussed later.

8.3 Eye-Fixation Prediction

Eye-fixation prediction aims to predict the *points* that people look at (free-viewing of natural scenes usually for 3-5 seconds). In contrast, salient object detection aims to predict the *object* that people are most interested in, and is a high-level semantic attention prediction problem. Though there exist differences between eye-fixation prediction and salient object detection, their solutions become similar now. The solutions to eye-fixation also contain unsupervised and supervised categories (Borji and Itti 2013). For instance, the learning-based solution to eye fixation prediction (Lu et al. 2012), integrating multiple features too, is similar to our solution. Recently, deep learning is also exploited for eye-fixation prediction (Liu et al. 2015). Previous research has shown that eyes are drawn to informative and salient areas in a scene, i.e., eye-fixation points lie in salient objects/regions. The analysis (Elazary and Itti 2008) over the LabelMe annotation data (Russell et al. 2008) demonstrates that human observers tend to annotate more salient objects first, and it is concluded that salient objects are interesting. More discussions about eye-fixation prediction and salient object detection could be found from the survey (Borji et al. 2014).

8.4 Multi-Level Fusion and Manifold Ranking

Multi-level fusion and manifold ranking (or Markov random field) are standard schemes to impose the smoothness. Manifold ranking or Markov random field has been studied in Yang et al. (2013) and Jiang et al. (2013a) for salient object detection. The multi-level scheme in our approach enjoys a benefit: training a supervised regressor is relatively easier while the manifold ranking approach, if trained under supervision, requires a complex learning technology, like structured learning.

8.5 Extension with Deep Learning

Our approach includes two steps: hand-crafted feature extraction and random forest regression. There are two possible improvements: (1) Can the representation be learnt instead of hand-crafted? (2) Can representation learning benefit from the specific task (saliency detection)? Inspired by the recent deep learning approach, in which representation learning and end-to-end optimization are two main points, there are several deep learning based works, e.g., Chen et al. (2015), Li and Yu (2015), Li et al. (2015), and Li and Yizhou (2016), handle the two aspects.

Multiscale deep features (MDF) (Li and Yu 2015) follows our discriminative feature integration framework and extracts the deep feature for the superpixels from a deep neural network that is learnt using regions from a set of labeled saliency maps. Unlike our approach and MDF that separate the feature extraction stage and the saliency computation stage, deep image saliency computing (DISC) (Chen et al. 2015), DeepSaliency (DS) (Li et al. 2015), and deep contrast learning (DCL) (Li and Yizhou 2016) jointly learn feature representation for a set of sampled pixels rather than superpixels as done in (MDF) (Li and Yu 2015), and predict the saliency map based on the fully convolutional neural network (FCN) (Long et al. 2015). In particular, DISC (Chen et al. 2015) stacks two FCNs: the first one takes the raw image as the input and the second one takes the post-processed output of the first FCN, a coarse saliency map, and the raw image as the input, yielding a fine saliency map. DS (Li et al. 2015) adopts the multi-task framework and jointly handles the segmentation task and the saliency detection task by sharing a sub-network. DCL (Li and Yizhou 2016) additionally exploits the super-pixel segmentation and other cues to refine the saliency map from FCN.

The comparison in terms of AUC is reported in Table 7. Overall speaking, deep learning-based algorithms perform better than our approach. This is because of deep-learnt features which are better than hand-crafted features and joint optimization. In addition, one can see that our approach performs similarly to (slightly worse than) deep learning-based approaches (MDF, DS, and DCL) over five out of the seven datasets while in other two datasets, ECSSD and DUT-OMRON*, deep learning-based algorithms significantly outperform our approach. Recall that the ECSSD dataset is more semantical. The reasons of the superior performances over ECSSD and DUT-OMRON* for deep learning algorithms include: deep learning has already been shown to capable to learn the semantic representation, and in particular, the four deep learning algorithms use the model pretrained over ImageNet. In contrast, the features in our approach are designed specifically for saliency and less capa-

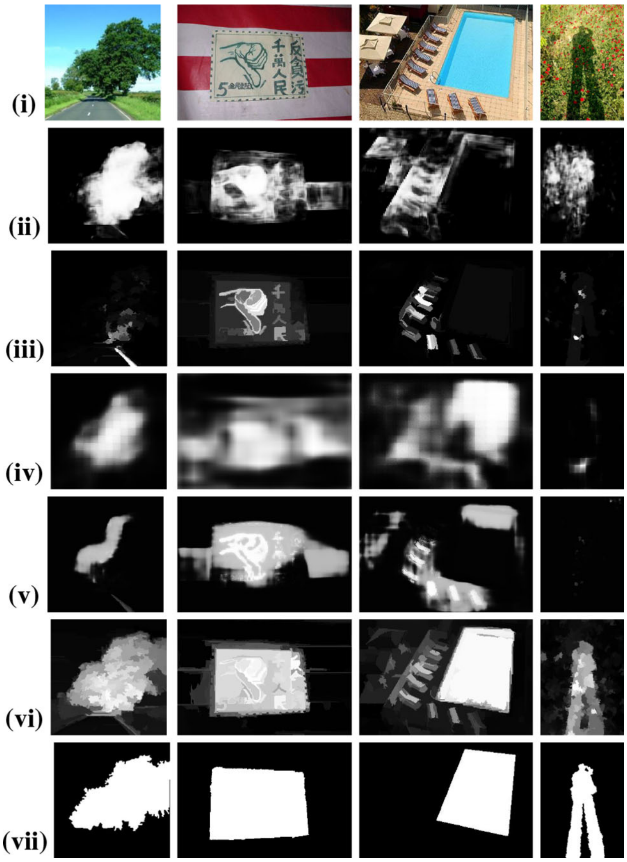


Fig. 11 Example saliency detection results for which our approach performs better than deep learning algorithms: **i** Input, **ii** DISC (Chen et al. 2015), **iii** MDF (Li and Yu 2015), **iv** DS (Li et al. 2015), **v** DCL (Li and Yizhou 2016), **vi** Our approach, **vii** Ground-truth

Table 7 Comparison between our approach and the followup approaches using deep learning in terms of AUC (larger is better) and the standard deviation for salient object detection

	MSRA-B	iCoSeg	ECSSD	DUT-OMRON	SED2	HKU-IS	DUT-OMRON*
DISC (Chen et al. 2015)	**0.984** ± **0.039**	0.937 ± 0.087	0.937 ± 0.083	0.917 ± 0.129	0.918 ± 0.151	0.947 ± 0.077	0.789 ± 0.205
MDF (Li and Yu 2015)	0.983 ± 0.049	**0.970** ± **0.055**	**0.949** ± **0.073**	0.921 ± 0.125	**0.959** ± **0.125**	**0.974** ± **0.064**	**0.861** ± **0.177**
DS (Li et al. 2015)	**0.987** ± **0.034**	0.965 ± 0.054	**0.972** ± **0.045**	**0.946** ± **0.098**	**0.963** ± **0.108**	**0.982** ± **0.034**	**0.865** ± **0.170**
DCL (Li and Yizhou 2016)	**0.983** ± **0.048**	**0.977** ± **0.049**	**0.969** ± **0.062**	**0.934** ± **0.119**	**0.955** ± **0.121**	**0.981** ± **0.043**	**0.876** ± **0.172**
DRFI	0.972 ± 0.052	**0.966** ± **0.071**	0.868 ± 0.265	**0.926** ± **0.116**	0.923 ± 0.187	0.952 ± 0.064	0.811 ± 0.183

The best three results are highlighted in bold

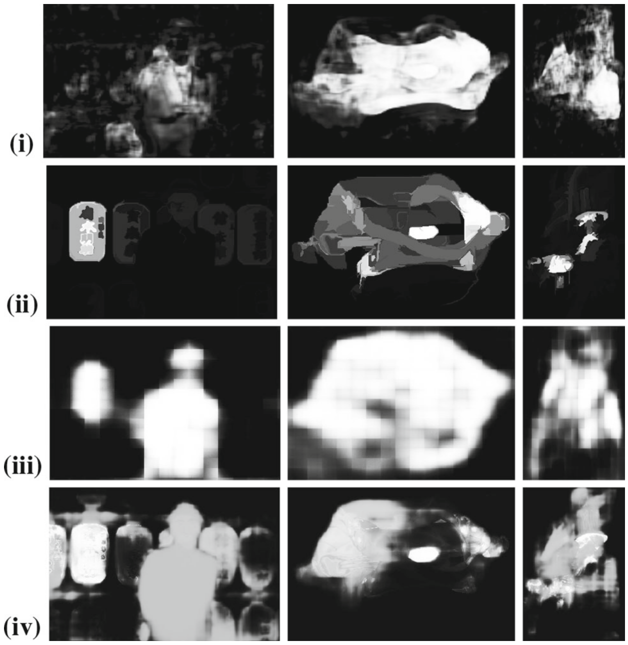


Fig. 12 The results of deep learning algorithms for the three images in Fig. 8: **i** DISC (Chen et al. 2015), **ii** MDF (Li and Yu 2015), **iii** DS (Li et al. 2015), **iv** DCL (Li and Yizhou 2016)

ble for capturing the semantics, and our approach does not exploit other datasets. The significant improvement over the most challenging dataset, DUT-OMRON*, might come from the same reasons.

We noticed that deep learning algorithms fail in some cases if the salient objects in testing images do not appear in the training images (the last example in Fig. 11) or are not regarded as salient objects in training images (the first example in Fig. 11), while our approach can still achieve good performance because our approach is less over-fitted. Several examples are shown in Fig. 11. We also observed some failure cases in our approach where deep learning algorithms also do not show good performances. Figure 12 shows the results of deep learning algorithms for the three failure examples of our approach given in Fig. 8. We can see that DS (Li et al. 2015) performs fairly well and other three algorithms perform poorly.

To summary, the main contribution of our work in comparison to deep learning-based algorithms lies in the introduction of the supervised feature integration framework for saliency computation, which suggests a research path in salient object detection and has already inspired the followup deep learning-based solutions.

9 Conclusion

This paper presents a discriminative regional feature integration approach, which is the first successful supervised approach to salient object detection, and already inspires subsequent supervised algorithms, e.g., exploring deep convolutional neural networks. The experimental results indicate the objectness and image-specific backgroundness descriptors are more significant than the contrast features (the center-surround differences) which are widely used before and thought as the basis of most salient object detection algorithms.

Acknowledgements This work was supported in part by the National Basic Research Program of China under Grant No. 2015CB351703 and 2012CB316400, and the National Natural Science Foundation of China under Grant No. 91120006 and NSFC (No. 61572264), Huawei Innovation Research Program (HIRP), and CAST young talents plan.

References

Achanta, R., Hemami, S. S., Estrada, F. J., & Süsstrunk S. (2009). Frequency-tuned salient region detection. In *CVPR*.

Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(11), 2189–2202.

Alpert, S., Galun, M., Basri, R., & Brandt, A. (2007). Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*.

Batra, D., Kowdle, A., Parikh, D., Luo, J., & Chen, T. (2010). iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *IEEE CVPR* (pp. 3169–3176).

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2014). Salient object detection: A survey. *CoRR*, arXiv:1411.5878.

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5706–5722.

Borji, A., & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. In *CVPR* (pp. 478–485).

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. *ECCV*, *2*, 414–429.

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

Chen, T., Lin, L., Liu, L., Luo, X., & Li, X. (2015). DISC: Deep image saliency computing via progressive representation learning. *CoRR*, arXiv:1511.04192.

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H. S., & Hu, S.-M. (2014). Global contrast based salient region detection. In *IEEE TPAMI*.

Cheng, M.-M., Warrell, J., Lin, W.-Y., Zheng, S., Vineet, V., & Crook, N. (2013). Efficient salient region detection with soft image abstraction. In *ICCV* (pp. 1529–1536).

Desingh, K., Krishna, K. M., Rajan, D., & Jawahar, C. V. (2013). Depth really matters: Improving visual salient region detection with depth. In *BMVC*.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3), 3.1–3.15.

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

Gao, D., Mahadevan, V., Vasconcelos, N. (2007). The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*.

Gao, D., & Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *ICCV* (pp. 1–6).

Goferman, S., Tal, A., & Zelnik-Manor, L. (2010). Puzzle-like collage. *Computer Graphics Forum*, *29*(2), 459–468.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *CVPR* (pp. 2376–2383).

Hoiem, D., Efros, A. A., & Hebert, M. (2005). Geometric context from a single image. In *ICCV* (pp. 654–661).

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE TIP*.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*.

Jia, Y., & Han, M. (2013). Category-independent object-level saliency detection. In *ICCV*.

Jiang, B., Zhang, L., Lu, H., Yang, C., & Yang, M.-H. (2013). Saliency detection via absorbing markov chain. In *ICCV*.

Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., & Li, S. (2011). Automatic salient object segmentation based on context and shape prior. In *BMVC*.

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., & Li, S. (2013). Salient object detection: A discriminative regional feature integration approach. In *IEEE CVPR* (pp. 2083–2090).

Jiang, P., Ling, H., Jingyi, Y. & Peng, J. (2013). Salient region detection by UFO: Uniqueness, focusness and objectness. In *ICCV*.

Jiang, Z., & Davis, L. S. (2013). Submodular salient region detection. In *CVPR* (pp. 2043–2050).

Kanan, C., & Cottrell, G. W. (2010). Robust classification of objects, faces, and flowers using natural image statistics. In *CVPR* (pp. 2472–2479).

Khuwuthyakorn, P., Robles-Kelly, A., & Zhou, J. (2010). Object of interest detection by saliency learning. In *ECCV*.

Kim, J., Han, D., Tai, Y.-W., Kim, J. (2014). Salient region detection via high-dimensional color transform. In *CVPR*.

Kimchi, R., & Peterson, M. A. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, *19*(7), 660–668.

Klein, D. A., & Frintrop, S. (2011). Center-surround divergence of feature statistics for salient object detection. In *ICCV*.

Koffka, K. (1935). *Principles of Gestalt Psychology*. Brace: Harcourt.

Küttel, D., & Ferrari, V. (2012). Figure-ground segmentation by transferring window masks. In *CVPR* (pp. 558–565).

Li, G., & Yu, Y. (2015). Visual saliency based on multiscale deep features. In *CVPR* (pp. 5455–5463).

Li, G., & Yizhou, Y. (2016). Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, N., Ye, J., Ji, Y., Ling, H., & Yu, J. (2014). Saliency detection on light fields. In *CVPR*.

Li, X., Li, Y., Shen, C., Dick, A. R., & van den Hengel, A. (2013). Contextual hypergraph modeling for salient object detection. In *ICCV* (pp. 3328–3335).

Li, X., Zhao, L., Wei, L., Yang, M., Wu, F., Zhuang, Y., Ling, H., & Wang, J. (2015). Deepsaliency: Multi-task deep neural network model for salient object detection. *CoRR*, arXiv:1510.05484.

Li, X., Lu, H., Zhang, L., Ruan, X., & Yang, M.-H. (2013). Saliency detection via dense and sparse reconstruction. In *ICCV*.

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *CVPR*.

Liu, F., & Gleicher, M. (2006). Region enhanced scale-invariant saliency detection. In *ICME* (pp. 1477–1480).

Liu, N., Han, J., Zhang, D., Wen, S., & Liu, T. (2015). Predicting eye fixations using convolutional neural networks. In *CVPR* (pp. 362–370).

Liu, R., Cao, J., Zhong, G., Lin, Z., Shan, S., & Su, Z. (2014). Adaptive partial differential equation learning for visual saliency detection. In *CVPR*.

Liu, T., Sun, J., Zheng, N.-N., Tang, X., & Shum, H.-Y. (2007). Learning to detect a salient object. *CVPR* (pp. 1–8).

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2011). Learning to detect a salient object. *IEEE TPAMI*, *33*(2), 353–367.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR* (pp. 3431–3440).

Lu, S., Mahadevan, V., & Vasconcelos, N. (2014). Learning optimal seeds for diffusion-based salient object detection. In *CVPR*.

Lu, Y., Zhang, W., Jin, C., & Xue, X. (2012). Learning attention map from images. In *CVPR* (pp. 1067–1074).

Lu, Y., Zhang, W., Lu, H., & Xue, X. (2011) Salient object detection using concavity context. In *ICCV* (pp. 233–240).

Ma, Y.-F., & Zhang, H.-J. (2003). Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*.

Marchesotti, L., Cifarelli, C., & Csurka, G. (2009). A framework for visual saliency detection with applications to image thumbnailing. In *ICCV* (pp. 2232–2239).

Margolin, R., Tal, A., & Zelnik-Manor, L. (2013). What makes a patch distinct? In *CVPR*.

Mehrani, P., & Veksler, O. (2010). Saliency segmentation based on learning and graph cut refinement. In *BMVC*.

Moosmann, F., Larlus, D., & Jurie, F. (2006). Learning saliency maps for object categorization. In *EECVW*.

Niu, Y., Geng, Y., Li, X., & Liu, F. (2012) Leveraging stereopsis for saliency analysis. In *CVPR* (pp. 454–461).

Peng, H., Li, B., Ji, R., Hu, W., Xiong, W., & Lang, C. (2013). Salient object detection via low-rank and structured sparse matrix decomposition. In *AAAI*.

Perazzi, F., Krähenbühl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *CVPR* (pp. 733–740).

Rahtu, E., Kannala, J., Salo, M., & Heikkilä, J. (2010). Segmenting salient objects from images and videos. *ECCV*, *5*, 366–379.

Ren, X., Fowlkes, C. C., & Malik, J. (2006). Figure/ground assignment in natural images. *ECCV, Part II* (pp. 614–627).

Rubin, E. (1958). Figure and ground. In *Readings in Perception* (pp. 194–203). Princeton, NJ: Van Nostrand.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1–3), 157–173.

Scharfenberger, C., Wong, A., Fergani, K., Zelek, J. S., & Clausi, D. A. (2013). Statistical textural distinctiveness for salient region detection in natural images. In *CVPR* (pp. 979–986).

Schroff, F., Criminisi, A., & Zisserman, A. (2008). Object class segmentation using random forests. In *BMVC* (pp. 1–10).

Shen, X., & Wu, Y. (2012). A unified approach to salient object detection via low rank matrix recovery. In *CVPR*.

Shi, K., Wang, K., Lu, J., & Lin, L. (2013). Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In *CVPR* (pp. 2115–2122).

Treisman, A., & Gelad, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Vicente, S., Kolmogorov, V., & Rother, C. (2008). Graph cut based image segmentation with connectivity priors. In *CVPR*.

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

Wang, J., Quan, L., Sun, J., Tang, X., & Shum, H.-Y. (2006). Picture collage. *CVPR*, *1*, 347–354.

Wang, L., Xue, J., Zheng, N., & Hua, G. (2011). Automatic salient object extraction with contextual cue. In *ICCV*.

Wang, M., Konrad, J., Ishwar, P., Jing, K., & Rowley, H. A. (2011). Image saliency: From intrinsic to extrinsic context. In *CVPR* (pp. 417–424).

Wang, P., Wang, J., Zeng, G., Feng, J., Zha, H., & Li, S. (2012). Salient object detection for searched web images via global saliency. In *CVPR* (pp. 3194–3201).

Wang, P., Zhang, D., Wang, J., Wu, Z., Hua, X.-S. & Li, S. (2012). Color filter for image search. In *ACM Multimedia*.

Wang, P., Zhang, D., Zeng, G., & Wang, J. (2012). Contextual dominant color name extraction for web image search. In *ICME Workshops* (pp. 319–324).

Wei, Y., Wen, F., Zhu, W., & Sun, J. (2012). Geodesic saliency using background priors. *ECCV*, *3*, 29–42.

Xie, Y., Huchuan, L., & Yang, M.-H. (2013). Bayesian saliency via low and mid level cues. *IEEE TIP*, *22*(5), 1689–1698.

Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. In *CVPR* (pp. 1155–1162).

Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *CVPR*.

Yu, S. X., & Shi, J. (2003). Object-specific figure-ground segregation. In *CVPR* (pp. 39–45).

Zhang, J., & Sclaroff, S. (2013). Saliency detection: A boolean map approach. In *ICCV* (pp. 153–160).

Zhu, W., Liang, S., Wei, Y., & Sun, J. (2014). Saliency optimization from robust background detection. In *CVPR*.

Zou, W., Kpalma, K., Liu, Z., Ronsin, J., et al. (2013). Segmentation driven low-rank matrix recovery for saliency detection. In *BMVC* (pp. 1–13).