# RGBD Salient Object Detection: A Benchmark and Algorithms

Houwen Peng[1], Bing Li[1], Weihua Xiong[1], Weiming Hu[1], and Rongrong Ji[2]

[1] Institute of Automation, Chinese Academy of Sciences, China
[2] Department of Cognitive Science, Xiamen University, China
http://sites.google.com/site/rgbdsaliency

**Abstract.** Although depth information plays an important role in the human vision system, it is not yet well-explored in existing visual saliency computational models. In this work, we first introduce a large scale RGBD image dataset to address the problem of data deficiency in current research of RGBD salient object detection. To make sure that most existing RGB saliency models can still be adequate in RGBD scenarios, we continue to provide a simple fusion framework that combines existing RGB-produced saliency with new depth-induced saliency, the former one is estimated from existing RGB models while the latter one is based on the proposed multi-contextual contrast model. Moreover, a specialized multi-stage RGBD model is also proposed which takes account of both depth and appearance cues derived from low-level feature contrast, mid-level region grouping and high-level priors enhancement. Extensive experiments show the effectiveness and superiority of our model which can accurately locate the salient objects from RGBD images, and also assign consistent saliency values for the target objects.

## 1 Introduction

Visual saliency has been a fundamental problem in neuroscience, psychology, and vision perception for a long time. It refers to the measurement of low-level stimuli that grab human attention in the early stage of visual processing [17]. We witness that the computation of saliency is originally a task of predicting where people look at an image, and recently has been extended to object-level saliency detection that involves separating the most conspicuous object from the background. This work focuses on the object-level saliency modeling, which benefits various applications including object detection and recognition [36], content based image retrieval [41][39], object aware image thumbnailing [31][14], etc.

Recently, detecting salient objects from RGBD images attracts lots of interest due to the birth of a new generation of sensing technologies, such as the *Microsoft Kinect* [1]. Although a small number of prior works aim to explore the role of depth in saliency analysis [25][13] and leverage depth to facilitate the saliency estimation [12][33], they are still at the initial stage of exploration and share common limitations: (1) Current studies on RGBD salient object detection are lack of a benchmark dataset that covers sufficient images with corresponding

D. Fleet et al. (Eds.): ECCV 2014, Part III, LNCS 8691, pp. 92–109, 2014.

accurate depth data and unified evaluation metrics. (2) The effective strategy that makes existing RGB-based saliency computation models work well in RGBD scenarios is not well explored. (3) Depth cues always work as an independent image channel for saliency detection in existing RGBD models [13][25], which inevitably ignores the strong complementarities between appearance and depth correspondence cues.

To address these problems, we first build up a large scale RGBD salient object benchmark with unified evaluation metrics, aiming at avoiding overfitting and biases. The benchmark contains 1,000 natural RGBD images captured by *Microsoft Kinect* together with the corresponding human-marked ground truth. To the best of our knowledge, it is the first large scale RGBD benchmark specially dedicated to the task of salient object detection. Second, to hold existing RGB-based saliency models still adequate in RGBD scenarios, we introduce a simple fusion strategy which extends RGB-based saliency models by incorporating depth-induced saliency. Specifically, the depth-induced saliency is produced by the proposed multi-contextual contrast method which computes depth rarity of a segmented patch from its local, global and background contexts. Finally, by considering low-level feature contrast, mid-level region grouping and high-level object-ware priors, we propose a novel multi-stage RGBD saliency estimation algorithm which combines depth information and appearance cues in a coupled manner. Experimental results on the benchmark show that our method can successfully identify salient content from RGBD images, which are difficult for existing visual saliency methods.

## 2 Related Work

**2D Saliency:** For saliency detection on 2D RGB image, most existing algorithms can be roughly divided into two categories, *i.e.*, local and global. Local approaches detect salient objects by measuring the rarity of a particular image region with respect to its neighborhoods. Itti *et al.* [17] first propose an influential saliency computational model, which performs center-surrounding differences on feature maps to obtain the local maxima of stimuli [24]. Harel *et al.* [15] define a graph on image and adopt random walks to compute saliency. To highlight the whole salient object, multi-scale contrast [29][44][27] and multi-cues integration [20,21] techniques are used. Due to lacking of global relations and structure, local contrast methods are sensitive to high frequency content or noises.

Global methods estimate saliency of a region based on its holistic rarity from an image. In [2], the authors define saliency by computing color difference from the mean image color on pixel level. Yet, this definition only accounts for first order average color and easily results in degraded performance on cluttered scenes. Goferman *et al.* [14] propose an improved method that highlights salient objects with their contexts in terms of low-level clues and global relationships. Cheng *et al.* [10] design a global contrast model that computes dissimilarities between 3D color histogram bins of all image regions. Perazzi *et al.* [34] formulate saliency estimation as two Gaussian filters performing on region uniqueness and distribution respectively. Other global models such as appearance reconstruction [28] and the fully connected MRF [18] are recently proposed to identify salient objects uniformly. Although global methods present superior results in some cases, they face challenges when an image contains similar foreground and background.

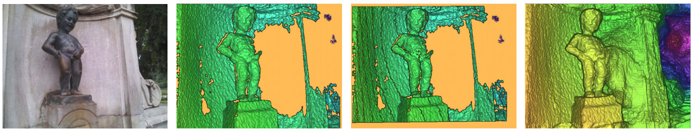
In addition, high-level priors are also incorporated into recent proposed methods to enhance the detection. Wei *et al.* [42] turn to background priors to guide the saliency detection, while Yang *et al.* [45] and Jiang *et al.* [19] integrate the background cues into the designed manifold ranking model and absorbing Markov chain, respectively. Shen and Wu [40] unify the high-level center, color and semantic priors into a low-rank matrix recovery framework. The prior from general object detector [4] is also considered in recent works [9][22][18].

**3D Saliency:** Contrary to the significant progress in 2D saliency research, the work leveraging depth information for saliency analysis is a bit limited. Niu *et al.* [33] exploit binocular images to estimate a disparity map and only use depth data to identify salient objects. So the performance is highly dependent on the quality of disparity map estimation which is another classical and challenging computer vision problem. Later, Lang *et al.* [25] conduct a comparative study of eye fixation prediction, rather than salient object detection, in 2D and 3D scenes after collecting a pool of 600 2D-vs-3D image pairs. Most recently, two related works [13][12] focus the task of detecting salient regions (other than salient objects) from RGBD images: Desingh *et al.* [13] verify that depth really matters on a small datasets with 80 images and propose to fuse saliency maps, produced by appearance and depth cues independently, through non-linear support vector regression. Ciptadi *et al.* [12] demonstrate the effectiveness of 3D layout and shape features from depth images in computing more informative saliency maps.

Compared with previous works, this paper has three fundamental differences: (1) Our RGBD salient object detection benchmark contains 1,000 images with accurate depth data captured from various scenarios, while the existing 3D datasets, *i.e.*, SSB [33], GIT [12] and NTU [13], are comparatively much smaller and simpler as shown in Table 1. (2) Rather than directly combining depth-induced saliency with color-produced saliency via simple fusion strategies [25][13], the proposed RGBD saliency model simultaneously takes account of depth and appearance information from multiple layers. (3) Last but not the least, a detailed quantitative analysis is given out about under what circumstance depth is indeed helpful, which is not explored in previous studies.

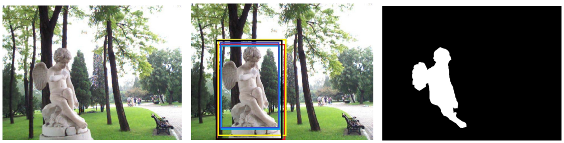
**Table 1.** Comparison of the benchmark and existing 3D datasets in terms of dataset size, number of objects contained within the images, type of scene and object, center bias, depth data, and publicity

| 3D Salient Object Detection Datasets | | | | | | | |
|---|---|---|---|---|---|---|---|
| Name | Size | Object No. | Scene Types | Object Types | Center Bias | Depth | Publicly Available |
| SSB [33] | 1000 | one (mostly) | >10 | >400 | Yes | No | Yes |
| NTU [13] | 33 | – – | – – | – – | – – | Yes | No |
| GIT [12] | 80 | multiple | <5 | <20 | No | Yes | Yes |
| Ours | 1000 | one (mostly) | >10 | >400 | Yes | Yes | Yes |



(a) RGB image (b) Raw Depth (c) Calibrated D (d) Smoothed D

**Fig. 1.** Depth image calibration and filling



(a) Input image (b) Labelled image (c) Ground truth

**Fig. 2.** A sample of image annotation. The image (b) is consistently labeled by five participants and included into our benchmark. (c) shows the final annotated salient object.

## 3 RGBD Salient Object Benchmark

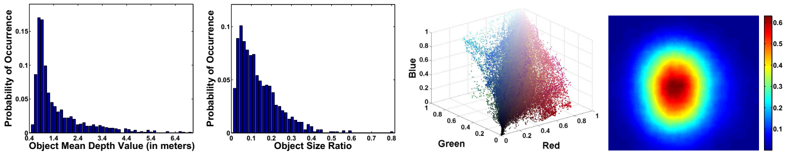
### 3.1 Dataset Construction

To remedy the data deficiency in current works and stimulate the research of detecting salient objects from RGBD images, we capture 5,000 images and their corresponding depth maps in diverse scenes. After preprocessing and annotation, we pick up 1,000 out of them to compose the final benchmark.

**Hardware Setup:** The reference depth map of our dataset is constructed using a standard *Microsoft Kinect*. The original *Kinect* device is not portable enough since it requires a mains power adapter with 110V or 220V AC. To solve the issue, we replace the adapter with a lithium battery (4400mAh 12V DC) that can power the *Kinect* for 4 hours of operation. To avoid camera shake and blur when capturing data, we strap the *Kinect* to a sturdy tripod. The output data of the *Kinect* is recorded by a connected laptop synchronously.

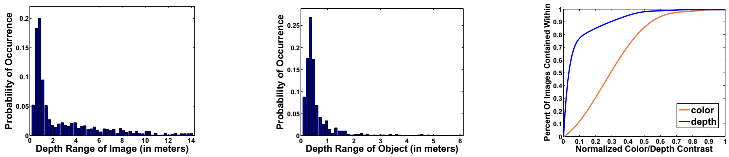
**Data Capture:** We visit a series of indoor and outdoor locations, *e.g.*, offices, supermarkets, campuses, streets and so on, and use the remoulded *Kinect* device to capture images of those scenes. Specifically, outdoor scenes are always captured in cloudy days or sunny dusks to avoid direct sunshine which may impair the precision of the infrared depth camera. To reduce imbalance due to human's preference, each scene is captured by a pair of collectors, and each object is photographed from at least four directions with different depth ranging from 0.5 to 10 meters.

**Data Preprocessing:** Because the color and infrared depth cameras on *Kinect* are a few centimeters apart horizontally, the captured color and depth images are not aligned as shown in Fig. 1(a) and (b). Thus, we calibrate each pair of color and depth images using the correction toolkit provided by *Microsoft Kinect SDK* to obtain more precise matching results (see Fig. 1(c)). It is worth noting that some regions in the aligned depth map are missing (see Fig. 1(c)) since they cannot be reached by the infrared laser projector. To obtain a filled and smoothed depth image, we adopt a colorization scheme [26] to repair the calibrated depth map. The processed depth image is shown in Fig. 1(d).



(a) Depth dist. (b) Size ratio dist. (c) Color dist. (d) Location dist.

**Fig. 3.** Bias statistics over object depth, color, size and location



(a) Image depth range dist. (b) Object depth range dist. (c) Contrast statistics

**Fig. 4.** Complexity statistics of the benchmark

**Salient Object Annotation:** After collecting 5,000 natural images and their depth maps, we first manually selected 2,000 images, each of which contains one or more distinctive foreground objects. Then, for each selected image, five participants are asked to draw a rectangle according to their first glance at the most attention-grabbing object. Since different people may have different opinions on what a salient object is in the same image, we exclude those images with low labeling consistency [29] and choose the top 1,000 satisfactory images. Finally, two participants use *Adobe Photoshop* to segment the salient object manually from each image. Fig. 2 shows a typical example.

### 3.2 Dataset Statistics and Analysis

We present the statistical characteristics of our dataset and show that it is suitable for evaluating different salient object detection algorithms.

**Diversity:** The resulting dataset contains more than 400 kinds of common objects captured in 11 types of scenes under different illumination conditions. The indoor scenes include offices, apartments, supermarkets, museums, etc., while the outdoor locations cover parks, campuses, streets, etc. Most images contain single salient object, while the others include multiple objects. Each object only appears once in the dataset after manually selection.

**Bias:** Besides high diversity, low bias is another important characteristic for a benchmark dataset. Fig. 3 shows the color, depth, size and location distributions of salient objects across all images in the dataset. We can see that the color and depth of patches in salient objects distribute across a widespread range in RGB and D(depth) space (see Fig. 3(a) and (c)). The size ratio between a salient object and its corresponding image varies from 0.16% to 80%, and most objects occupy less than half area of the images. The locations of salient objects correlate strongly with a centered Gaussian distribution as show in Fig. 3(d). It is caused by the fact that human naturally frame an object of interest near the center of

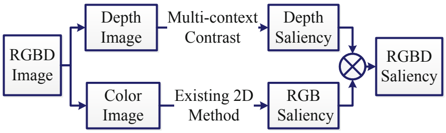


**Fig. 5.** The extension framework: depth saliency is produced by the multi-contextual contrast method, while RGB saliency is estimated by any existing 2D saliency methods

the image when taking pictures. We can also find such type of center bias in other public datasets such as MSRA [29] and SSB [12] (see Table 1).

**Difficulty:** To avoid that the salient objects can be easily extracted by a simple thresholding on depth maps, both the objects and depth images in the benchmark dataset share variable depth ranges as show in Fig. 4(a) and (b). We calculate the color and depth contrast between the foreground objects and background within a single image. Fig. 4(c) shows the cumulative histograms of the normalized contrast. It tells us that almost all the images in our benchmark have relatively low contrast between the background and salient objects. For example, 90% images have depth contrast within a distance of 0.3, while nearly 50% images have color contrast within this distance. The benchmark with low contrast inevitably brings up more challenges for detecting salient objects.

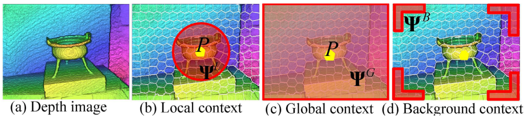
### 3.3 Evaluation Metrics

We introduce two types of measures to evaluate algorithm performance on the benchmark. The first one is the gold standard: Precision-Recall (PR) curve and $F$-measure. Precision corresponds to the percentage of salient pixels correctly assigned to all the pixels of extracted regions, 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 from 0 to 255 that determines if a pixel is on the salient object. $F$-measure indicates the weighted harmonic of precision and recall: $F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{(\beta^2 \times Precision + Recall)}$, where $\beta^2$ is set to be 0.3 to stress precision more than recall. The $F$-measure is computed with an adaptive saliency threshold that is defined as twice the mean saliency of the image [2].

The second is Receiver Operating Characteristic curve (ROC) and the Area Under the ROC Curve (AUC). By thresholding over the saliency maps and plotting true positive rate vs. false positive rate, an ROC curve is acquired. The AUC score is calculated as the area underneath the ROC. Perfect performance corresponds to an AUC score of 1 while a score of 0.5 indicates chance level.

## 4 Extending 2D Saliency Models for RGBD Images

To make existing RGB salient object detection models still adequate in RGBD scenarios, this section proposes a framework that extends 2D saliency models for RGBD images.



(a) Depth image (b) Local context (c) Global context (d) Background context

**Fig. 6.** Visual illustration for the definitions of contextual sets. The estimated patch is marked as Yellow, while its contextual patches are marked as Red.

### 4.1 Extension Framework

The proposed extension framework is a fusion process which includes three major steps as shown in Fig. 5: (1) separate the input RGBD image into two independent components: a RGB image and a depth map, (2) calculate their own saliency maps and (3) fuse these two saliency maps into a final one through the standard pixel-wise multiplication [6]. Specifically, to produce depth-induced saliency maps, we propose a multi-contextual contrast-based saliency estimation method. RGB saliency maps are obtained by any existing saliency detection algorithm on RGB images.

### 4.2 Depth Saliency from Multi-contextual Contrast

From the observation that an object lying at a different depth level from the others will noticeably attract our attention, thus we define our depth saliency computation based on the contrast analysis. To take advantage of local center-surrounding relationship, global distinctiveness and background information of the depth image, we introduce three types of contextual contrast, *i.e.*, local, global and background. Specifically, we first divide the depth image into $N$ non-overlapping patches using SLIC algorithm [3]. For any patch $P$, we define its saliency $S(P)$ as the multiplication of these three types contextual contrast

$$S(P) = \prod_{k \in \{L,G,B\}} C(P, \varPsi^k), \tag{1}$$

where $C(\cdot)$ is a typical contrast computation function, $\varPsi^L = \{P_1^L, ..., P_{n_L}^L\}$ represents the local contextual set which consists of $n_L$ nearest neighbor patches to $P$, $\varPsi^G = \{P_1^G, ..., P_{n_G}^G\}$ indicates the global contextual set including all patches of the depth image except for $P$, and $\varPsi^B = \{P_1^B, ..., P_{n_B}^B\}$ represents the pseudo-background context which consists of $n_B$ patches from the four corners of the depth image as shown in Fig. 6. The definition of pseudo-background is inspired by the observation that the patches from corners of images are more likely to be background and contain lots of scene information which contributes to distinguish salient objects. Background context from image corners is more robust than that from image boundaries [42] in practice, because the salient objects have lower probabilities to touch corners of image.

Different from the traditional definition of contrast computation function that is the difference between $P$ and the patches in $\varPsi^k$ ($k \in \{L, G, B\}$) with respect to visual features [10][20][34], we exploit Shannon's self-information, *i.e.*, $-\log(p(\cdot))$, as the measure of visual saliency. Self-information is a plausible biological metric [7], implying that an image patch will contain more distinctive information when it occurs in feature space with less probability. Thus, $C(\cdot)$ is defined as

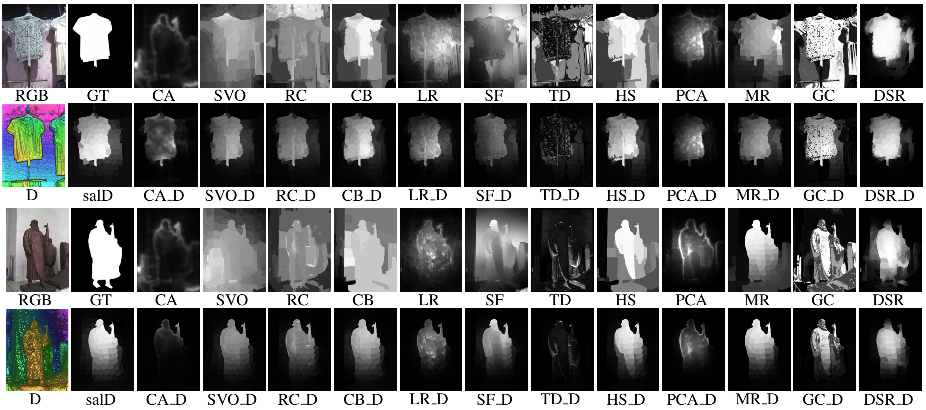


**Fig. 7.** Visual comparisons of saliency maps before and after fusing depth saliency. The odd rows are the results of existing 12 RGB saliency models, while the even rows show the maps after fusing depth saliency. Here, "salD" is the saliency map produced by our multi-contextual contrast method.

$$C(P, \Psi^k) = -\log(p(P|\Psi^k)). \tag{2}$$

Here, the conditional probability $p(P|\Psi^k)$ represents the underlying density of $P$ with the given context $\Psi^k$. Specially, let $\mathbf{d}$ represent the average depth of $P$, and $D^k = \{\mathbf{d}_1^k, \mathbf{d}_2^k, ..., \mathbf{d}_{n_k}^k\}$ be the depth value set corresponding to the average depth of all patches in $\Psi^k$, then we have

$$p(P|\Psi^k) = \hat{p}(\mathbf{d}|D^k). \tag{3}$$

To model the depth distribution of salient objects with mixture of depth values, we adopt Gaussian kernel density estimator [38] to compute the probability density of $\hat{p}(\mathbf{d}|D^k)$ in Eq.(3) as

$$\hat{p}(\mathbf{d}|D^k) = \frac{1}{n_k} \sum_{j=1}^{n_k} e^{-\frac{\|\mathbf{d}-\mathbf{d}_j^k\|^2}{2(\sigma_d^k)^2}}, \tag{4}$$

where $\sigma_d^k$ is the bandwidth of Gaussian kernel that controls the influence of depth difference. The kernel density of this type does not assume any specific underlying distribution and, theoretically, the estimation can converge to any density function with enough samples [38].

The above procedure works on each patch and results in patch-level saliency. The final pixel-wise depth-induced saliency map is obtained through assigning the saliency value of each patch to every pixel belonging to it.

Examples of depth-induced saliency map derived from multi-contextual contrast are shown in Fig. 7. We can see that when the depth level of foreground object is distinct from background, our method is able to roughly locate the salient regions and approximately highlight the object. Through fusing the depth saliency with RGB saliency produced by existing 2D saliency algorithms, we obtain more visually feasible results.

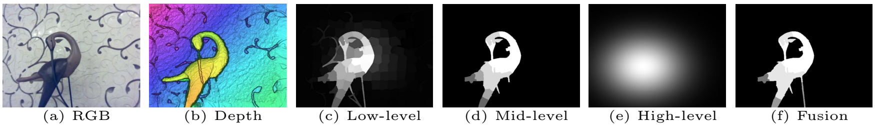

(a) RGB (b) Depth (c) Low-level (d) Mid-level (e) High-level (f) Fusion

**Fig. 8.** Saliency maps produced by the key three stages in our approach

## 5 A Novel Saliency Model for RGBD Images

Although the simple late fusion strategy achieves improvements, it still suffers from inconsistency in the homogeneous foreground regions and is lack of precision around object boundaries. It may be ascribed to treating the appearance and depth correspondence cues in an independent manner. To resolve the issue, we propose a novel and effective method leveraging both depth and appearance cues from multiple levels. Our approach consists of three stages. First, we extend the low-level multi-contextual contrast method proposed in the previous section to RGBD cases and produce an initial saliency map. Next, we exploit thresholding on the initial saliency map to yield saliency seeds which are diverse regions with high saliency values. Starting with any one of saliency seeds, region grouping is performed on a weighted graph by using Prim's algorithm [35] to select candidate regions which have high probabilities belonging to the foreground object. This procedure is repeated until all the seeds are traversed. A visual consistent saliency map is generated at the end of this stage. Finally, saliency maps generated by previous two stages are combined through a Bayesian fusion strategy [28]. Besides, a high-level object-aware prior is also integrated to boost the performance. We illustrate this process on an example image in Fig. 8.

### 5.1 Low-Level Feature Contrast

For RGBD saliency detection, the classical low-level center-surrounding contrast can still work as a fundamental and support principle. Thus we extend the multi-contextual contrast model presented in Sec. 4.2 from depth space to RGBD scenarios by simply altering the feature and the technique of density estimation.

**Feature Selection:** Based on the verified conclusions that color, size and spatial position are undoubted attributes for guiding visual attention [43], we define the feature representation of RGBD images as the stack of these low-level features and depth value. Formally, for any patch $P$, its feature vector is defined as $\mathbf{f} = [\mathbf{c}, \mathbf{l}, \mathbf{r}, \mathbf{d}]^T$, where $\mathbf{c}$ is the average CIELab color value of pixels in $P$ since CIELab can better approximate human color perception [5], $\mathbf{l}$ represents center location of pixels in $P$ on the image plane, $\mathbf{r}$ is the region size defined by the number of contained pixels, and $\mathbf{d}$ is the average depth value of pixels in $P$.

**Density Estimation:** With the above constructed features and multiple contexts, we estimate the probability $p(P|\Psi^k)$ in Eq. (3) by a weighted version of Gaussian kernel density technique formulated as

$$p(P|\Psi^k) = \hat{p}(\mathbf{f}|F^k) = \frac{1}{n_k}\sum_{j=1}^{n_k} \alpha_j^k e^{-\frac{\|\mathbf{c}-\mathbf{c}_j^k\|^2}{2(\sigma_c^k)^2}} e^{-\frac{\|\mathbf{l}-\mathbf{l}_j^k\|^2}{2(\sigma_l^k)^2}} e^{-\frac{\|\mathbf{d}-\mathbf{d}_j^k\|^2}{2(\sigma_d^k)^2}}, \tag{5}$$

where $F^k = \{\mathbf{f}_1^k, ..., \mathbf{f}_{n_k}^k\}$ is the feature representation corresponding to the surrounding patch set $\Psi^k$, and $\alpha_j^k = \mathbf{r}_j^k/\mathbf{r}$ is the weight coefficient defined by the size ratio between the target patch $P$ and the contextual patch $P_j^k$. Here, the use of different bandwidths for different features is desirable since the variances are inconsistent in each feature dimension. For example, the color usually has more variance than the location and therefore should be assigned wider range. Through computing the probability density for each patch and merging all together, we obtain the initial patch-level saliency map (see Fig. 8(c)).

### 5.2 Mid-Level Region Grouping

To completely extract salient objects with precise boundaries, we perform mid-level salient region grouping with the help of the initial results obtained from low-level stage. Following the general object detection methods [8][30], the proposed region grouping is based on Prim's algorithm which greedily computes the maximum spanning tree of a weighted graph.

**Graph Construction:** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \rho)$ be the weighted connective graph of the superpixel segmentation [3] of an RGBD image, where the vertices $\mathcal{V}$ are the patches and the edges $(P, Q) \in \mathcal{E}$ connect the neighbor patches $P$ and $Q$. The weight function $\rho : \mathcal{E} \to [0, 1]$ assigns weights $\rho(P, Q){=}\rho_{P,Q}$ to edges. Following [30], we model the weight $\rho$ with a logistic function:

$$\rho_{P,Q}{=}\sigma(\mathbf{w}^T \mathbf{x}_{P,Q} + b), \qquad \sigma(x) = (1 + \exp(-x))^{-1}, \tag{6}$$

where $\mathbf{x}_{P,Q}$ is a feature vector containing efficient features that measure the similarity and compactness of patches $P$ and $Q$. The computation of the weight parameter $\mathbf{w}$ and bias $b$ is resort to learning on the training data. Towards this end, we first assign a patch to a foreground object if over 80% of the number of pixels in the patch belongs to the object. Then we mine for pairs of patches which are involved in the same object and label them as the positive samples ($y_{P,Q} = 1$), otherwise, as the negative samples ($y_{P,Q} = 0$). The estimation of optimal parameters are computed by maximizing the log likelihood:

$$\{\mathbf{w}^*, b\} = \arg\max_{\mathbf{w},b} \sum_{\forall(P,Q)\in\mathcal{E}} y_{P,Q} \log \rho_{P,Q} + (1 - y_{P,Q}) \log(1 - \rho_{P,Q}). \tag{7}$$

The features defined in $\mathbf{x}$ consist: (1) Color similarity $x_c$: color consistency is a important cue for objectness. With the CIELab color values $\mathbf{c}_P$ and $\mathbf{c}_Q$ of patches $P$ and $Q$, we define the color similarity $x_c \in [0, 1]$ of patches as $x_c{=} \exp(-\|\mathbf{c}_P{-}\mathbf{c}_Q\|^2)$. (2) Surface normal consistency $x_d$: patches in the same plane have high probability belonging to one semantic region. To determine whether patches come from one plane, we compare their surface normals following the procedure proposed in [16]. The surface normal $\mathbf{s}_P$ of patch $P$ is calculated as the cross product of two principle tangential vectors to local surface at the center location of $P$. Similar to color similarity, surface normal consistency is computed by $x_d = \exp(-\|\mathbf{s}_P - \mathbf{s}_Q\|^2)$. This cue is favor of combining patches from the same plane. (3) Border overlapping $x_b$: patches share common borders are likely to be grouped into one same region. Let $l_P$ and $l_Q$ be

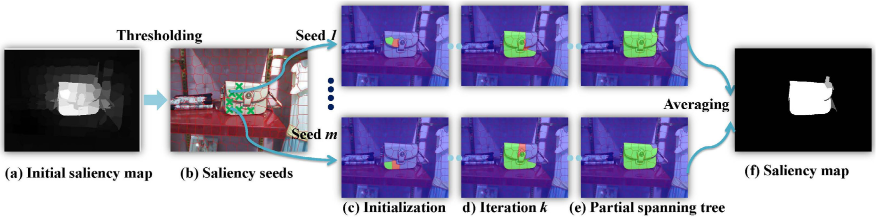


**Fig. 9.** Illustration of mid-level salient region grouping. The image patches existing in the partial spanning tree are marked as Green, while the patches to be added into the tree are marked as Red.

the perimeters of patches $P$ and $Q$, then the border overlapping is defined as $x_b = (l_P \cap l_Q)/\min(l_P, l_Q)$ which represents the maximum ratio between their common border and each of their perimeters. This feature favors the merging of patches with longer overlapping border. Through learning on the training data, the weight $\mathbf{w}^*$ for these features is obtained.

**Salient Region Grouping:** Given the constructed graph, we perform region grouping using Prim's algorithm which generates a partial spanning tree with high sum of edge weights starting from a salient seed. The key procedure of our region grouping for an RGBD image is summarized as follows (see Fig. 9):

1. Generate a salient seed set $\{s_1, ..., s_m\}$, which is consisted by $m$ image patches with high saliency values, through thresholding on the initial saliency map.
2. For each seed $s_i$ in the set, repeat following procedure:
   - Initialize a spanning tree $T_1^{(i)}$ with the seed $s_i$.
   - Perform an iterative tree-grouping procedure based on Prim's algorithm which greedily selects the connected edge $(P, Q) \in \mathcal{E}$ with the maximum weight $\rho_{P,Q}$ and adds into the spanning tree $T_k^{(i)}$
   - Output the partial spanning tree $T^{(i)}$ when Prim's algorithm meets the terminal condition.
3. Generate a saliency map by computing the frequency of each patch appeared in all the spanning trees $\{T^{(1)}, ..., T^{(m)}\}$.

Specifically, to generate the salient seed set in Step 1, we experimentally set a threshold $T$ to partition the patches embedded in the initial saliency map into a high-confidence group and a low-confidence group according to their saliency values. The members in the high-confidence group server as the seminal patches and constitute the salient seed set. In the iteration of Prim's algorithm, an effective stopping criterion is necessary in order to yield desirable partial spanning trees that firmly cover the nodes within the object, rather than the full nodes on the graph. To the end, we design a termination function which includes two terms: (1) The probability $1 - \Omega_{P,Q}$ that the candidate edge $(P, Q)$ does not connect patches of the same object. (2) The difference of saliency value $S(P) - S(Q)$, which is estimated in the low-level computation stage, reflects the low-level saliency contrast. In practice, the termination function defined as the mean of these two terms works well:

$$f_{P,Q} = (1 - \Omega_{P,Q} + S(P) - S(Q))/2. \tag{8}$$

With this function, Prim's algorithm checks the terminal condition $f_{P,Q} > f_0$ at each iteration to decide whether the edge $(P, Q)$ is added to the tree. Here, $f_0$ is a parameter set at the initialization step of the algorithm. An illustration example of the whole procedure of mid-level region grouping is shown in Fig. 9.

### 5.3 High-Level Priors Enhancement

In the final stage, we fuse the saliency maps produced by previous two stages, and further incorporate high-level priors to boost the performance. To combine the saliency maps from low-level and mid-level, we adopt the Bayesian integration method proposed in [28] which sums two posterior probabilities computed by one saliency map serves as the prior while the other works as the likelihood. Different from the high-level priors adopted in previous work [23][40][44][28], we propose an object-aware prior which take account of both location and size of salient objects. The prior is formulated as a Gaussian model:

$$G(a) = \exp[-(\frac{(x_a - \mu_x)^2}{2\sigma_x^2} + \frac{(y_a - \mu_y)^2}{2\sigma_y^2} + \frac{(z_a - \mu_z)^2}{2\sigma_z^2})], \tag{9}$$

where $(x_a, y_a, z_a)$ are the coordinates of pixel $a$ in the normalized image plan $X - Y$ and the depth range $Z$, $(\mu_x, \mu_y, \mu_z)$ are the coordinates of object center derived from the average values of patches within the salient seed set. Considering the impact of object size, we set the variance $(\sigma_x^2, \sigma_y^2, \sigma_z^2)$ to be $(2o_x, 2o_y, 2o_z)$, where $o_x$ is the range of all saliency seeds on $X$-coordinate, while $o_y$ and $o_z$ are ranges on $Y$ and $Z$ coordinates respectively. Finally, the pixel-wise saliency map induced by the object-aware prior is integrated with the resulting map of Bayesian fusion by simple multiplication, as shown in Fig. 8, which generates the final result of our RGBD saliency model.

## 6 Experiments and Comparisons

### 6.1 Experimental Setup

Depth Model: In the implementation of the multi-contextual contrast model for depth images, we first use the SLIC superpixels [3] to partition the input depth image into $N = 200$ non-overlapping patches. For each patch $P$, we select $n_L = 32$ spatial nearest neighbor patches on the image plane as its local context $\Psi^L$, while all patches of the depth image except for $P$ as the global context $\Psi^G$. To get the pseudo-background context $\Psi^B$, we pick out $n_B$= 36 boundary patches that are closest to the four corners of the image. The bandwidths of Gaussian kernel are empirically set as $(\sigma_d^L)^2 = 0.1$, $(\sigma_d^G)^2 = 0.05$, and $(\sigma_d^B)^2 = 0.25$.

RGBD Model: For low-level feature contrast computation in the RGBD model, we set $N$ and $n_k (k \in L, G, B)$ to the same values as in the depth model. For the multivariable extension of kernel density estimation, we adopt the applicable Sheather-Jones plug-in approach [38] which minimizes an estimate of mean integrated squared error to estimate the kernel bandwidth $\sigma_\xi^k$ $(\xi \in \{\mathrm{c}, \mathrm{l}, \mathrm{d}\})$. To obtain the parameters $\mathbf{w}$ and $b$, we first divide the benchmark into two equal

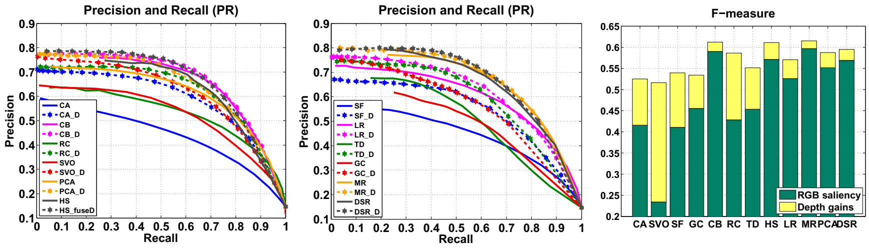
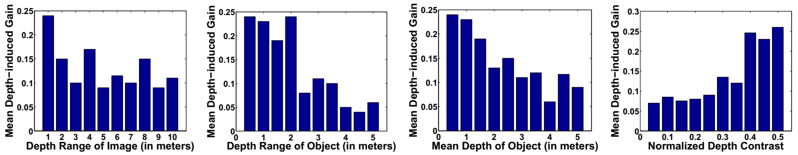


**Fig. 10.** The quantitative comparisons of the performance of existing saliency detection methods before and after fusing depth saliency by the proposed extension framework



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

**Fig. 11.** Depth-induced gains analysis with respect to (a) depth range of images (DRI), (b) depth range of objects (DRO), (c) average depth of objects (ADO), and (d) normalized depth contrast between foreground and background (DC)

subsets and then choose one subsect for training and the other for testing. The optimal $\mathbf{w}^*$ and $b^*$ are computed through 5-fold cross validation on the training subset according to Eq. (7), which is solved by gradient ascent. The threshold $T$ and terminal parameter $f_0$ are empirically set to 0.8 and 0.45 respectively.

### 6.2 Experimental Results and Comparisons

In this section, we perform two sets of experiments: (1) evaluations of 12 prevailing RGB saliency methods before and after the fusion of depth saliency, (2) comparisons between our RGBD model and existing 3D models, and analysis of performance contributions of each individual component in the RGBD model.

**Evaluations of 2D Models:** We first compare the performances of existing RGB saliency models before and after fusing depth saliency produced by our multi-contextual contrast model. We select 12 state-of-the-art RGB saliency detection approaches, including the top 4 models ranked in the survey[6]: SVO[9], CA[14], CB[20], and RC[10], and 8 recently developed prominent methods: SF[34], LR[40], HS[44], MR[45], PCA[32], TD[37], GC[11], and DSR[28]. Fig. 10 presents the experimental results, in which the postfix '_D' denotes the method after fusing the depth saliency. We can see that both the PR curves and F-measure values of all the RGB salient object detection algorithms are improved by extra depth-produced saliency. It indicates that (1) the additional depth information is beneficial for salient object detection, (2) our multi-contextual contrast method performed on depth images are effective, (3) existing 2D saliency models are able to hold availability in RGBD scenarios through the proposed extension framework. Furthermore, let's review the examples shown in Fig 7 and investigate the

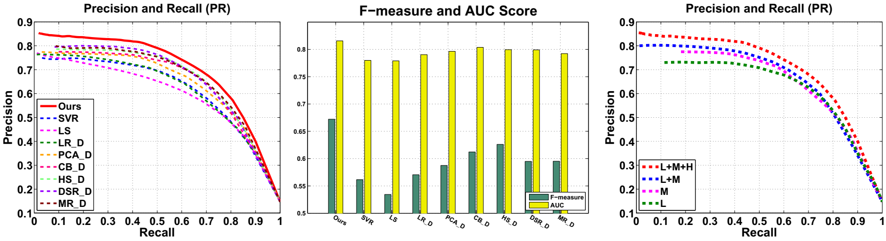


**Fig. 12.** Left and Middle: Quantitative comparisons of our approach with 8 competing RGBD methods. Right: Evaluation of performance contributions of each individual component in the RGBD model. The uppercase 'L', 'M', and 'H' represent the results of low-, mid- and high-level saliency respectively.

underlying reason why depth brings in improved performance. It is obviously that most existing saliency detection methods will fail when the object has similar appearance to the background. However, the introduction of depth contrast helps to extract the salient objects successfully.

We further analyze the situations in which depth is more helpful. To the end, we quantify the depth-induced gains for existing RGB saliency computational models against four aspects: depth range of images (DRI), depth range of objects (DRO), average depth of objects (ADO), and depth contrast between foreground and background (DC). The statistical diagrams are shown in Fig. 11. It is observed that (1) DRI nearly has no direct effect on depth effectiveness, but objects with lower depth ranges are more possible to be identified by depth saliency (see Fig. 11(a, b)). (2) If objects lie at close depth levels, *i.e.,* near to the camera, depth-induced gains are relatively high (Fig. 11(c)). (3) The higher depth contrast (DC) between foreground and background is, the more improvements from depth in identifying salient objects can be achieved (Fig. 11(d)). In conclusion, depth is more helpful when objects have relatively lower depth range, lie closer to the camera, or have high depth contrast with background.

**Comparisons and Analysis of RGBD Models:** We choose the top 6 2D saliency approaches after fusing depth saliency: DSR_D, MR_D, HS_D, CB_D, PCA_D and LR_D, and 2 recent proposed RGBD salient region detection methods: SVR [13] and LS [12] as baselines to compare their performances with our RGBD model. Fig. 12 shows the quantitative comparisons among these method on the constructed RGBD datasets in terms of PR curves, F-measure and AUC scores. It is observed that the proposed RGBD method is superior to baselines in terms of all the evaluation metrics. Interestingly, the method SVR, which uses non-linear support vector regression to fuse depth and RGB saliency, has lower performance compared to methods which uses the simple multiplication as fusion strategy. The underlying reasons are (1) SVR selects RC [10] to serve as the RGB method which is not the best one in 2D models, (2) SVR is designed for salient region detection, which is a little different from object-level saliency. Fig. 13 shows the qualitative comparisons on several RGBD images. We can see that the proposed method can accurately locate the salient objects, and also produce

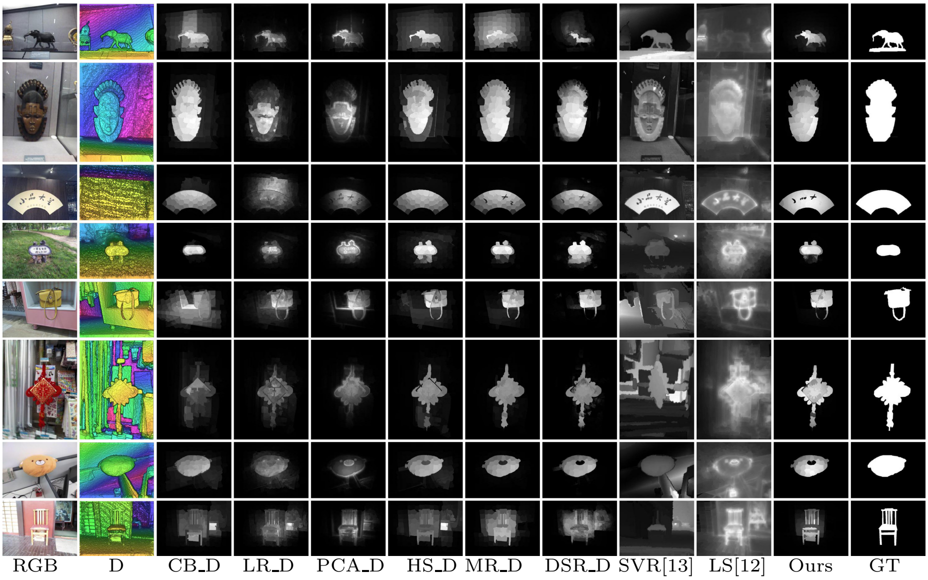


**Fig. 13.** Visual comparison of saliency maps

nearly equal saliency values of the pixels within the target objects. It confirms the effectiveness of our RGBD model which takes advantage of both depth and appearance cues from low-, mid-, and high-level.

We are also interested on the contributions of each component in our RGBD model. So we quantify the three key stages respectively. Fig. 12 (Right) illustrates the PR curves resulting from the accumulation of each component. It is seen that every stage in the RGBD model helps to improve the final performance. Particularly, the second stage brings lots of contributions on promoting the precision, this is mainly because the mid-level salient region grouping enhances the consistency and compactness of salient patches.

## 7 Conclusion

In this paper, we provide a comprehensive study on RGBD salient object detection including building up a benchmark, introducing an extension framework for 2D saliency models, and proposing a novel multi-stage RGBD model. Experiments verify that the depth-produced saliency can work as a helpful complement to existing color-based saliency models, especially when objects stay closer to the camera, have high depth contrast with background, or experience relatively low depth range. Compared with other competing 3D models, our proposed RGBD model achieves superior performance and produces more robust and visual favorable results. We believe that the constructed benchmark and our work are helpful to stimulate further research in the area.

**Acknowledgements.** This work is partly supported by the 973 Basic Research Program of China (No. 2014CB349303), the National 863 High-Tech R&D Program of China (No. 2012AA012503, 2012AA012504), the National Nature Science Foundation of China (No. 61370038, 61202327, 61303086), the Natural Science Foundation of Beijing (No. 4121003), the Fundamental Research Funds for the Central Universities (No. 2013121026), and the Guangdong Natural Science Foundation (No. S2012020011081).

## References

1. Microsoft Corp. Redmond WA. Kinect for Xbox 360
2. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR, pp. 1597–1604 (2009)
3. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE TPAMI 34(11), 2274–2282 (2012)
4. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR, pp. 73–80 (2010)
5. Borji, A., Itti, L.: Exploiting local and global patch rarities for saliency detection. In: CVPR, pp. 478–485 (2012)
6. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 414–429. Springer, Heidelberg (2012)
7. Bruce, N.D.B., Tsotsos, J.K.: Saliency based on information maximization. In: NIPS (2005)
8. Carreira, J., Sminchisescu, C.: Cpmc: Automatic object segmentation using constrained parametric min-cuts. PAMI 34(7), 1312–1328 (2012)
9. Chang, K.Y., Liu, T.L., Chen, H.T., Lai, S.H.: Fusing generic objectness and visual saliency for salient object detection. In: ICCV, pp. 914–921 (2011)
10. Cheng, M., Zhang, G., Mitra, N.J., Huang, X., Hu, S.: Global contrast based salient region detection. In: CVPR, pp. 409–416 (2011)
11. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: ICCV, pp. 1–8 (2013)
12. Ciptadi, A., Hermans, T., Rehg, J.M.: An in Depth View of Saliency. In: BMVC, pp. 1–11 (2013)
13. Desingh, K., Krishna, K.M., Jawahar, C.V., Rajan, D.: Depth really matters: Improving visual salient region detection with depth. In: BMVC, pp. 1–11 (2013)
14. Goferman, S., Manor, L.Z., Tal, A.: Context-aware saliency detection. In: CVPR, pp. 1915–1926 (2010)
15. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS, pp. 545–552 (2006)
16. Holz, D., Holzer, S., Rusu, R.B., Behnke, S.: Real-time plane segmentation using rgb-d cameras. In: RoboCup, pp. 306–317 (2011)
17. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI 20(11), 1254–1259 (1998)
18. Jia, Y., Han, M.: Category-independent object-level saliency detection. In: ICCV, pp. 1761–1768 (2013)
19. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.H.: Saliency detection via absorbing markov chain. In: ICCV, pp. 1665–1672 (2013)

20. Jiang, H., Wang, J., Yuan, Z., Liu, T., Zheng, N., Li, S.: Automatic salient object segmentation based on context and shape prior. In: BMVC, pp. 1–12 (2011)
21. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: CVPR, pp. 1–8 (2013)
22. Jiang, P., Ling, H., Yu, J., Peng, J.: Salient region detection by UFO: Uniqueness, focusness and objectness. In: ICCV, pp. 1976–1983 (2013)
23. Judd, T., Ehinger, K.A., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV, pp. 2106–2113 (2009)
24. Koch, C., Ullman, S.: Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology 4(4), 219–227 (1985)
25. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Influence of depth cues on visual saliency. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 101–115. Springer, Heidelberg (2012)
26. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. ACM Trans. Graph. 23(3), 689–694 (2004)
27. Li, X., Li, Y., Shen, C., Dick, A.R., van den Hengel, A.: Contextual hypergraph modeling for salient object detection. In: ICCV, pp. 3328–3335 (2013)
28. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction. In: ICCV, pp. 2976–2983 (2013)
29. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: CVPR, pp. 1–8 (2007)
30. Manen, S., Guillaumin, M., Gool, L.J.V.: Prime object proposals with randomized prim's algorithm. In: ICCV, pp. 2536–2543 (2013)
31. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: ICCV, pp. 2232–2239 (2009)
32. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: CVPR, pp. 1139–1146 (2013)
33. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR, pp. 454–461 (2012)
34. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR, pp. 733–740 (2012)
35. Prim, R.: Shortest connection networks and some generalizations. Bell System Tech. J., 1389–1401 (1957)
36. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: CVPR, pp. 37–44 (2004)
37. Scharfenberger, C., Wong, A., Fergani, K., Zelek, J.S., Clausi, D.A.: Statistical textural distinctiveness for salient region detection in natural images. In: CVPR, pp. 979–986 (2013)
38. Scott, D.W.: Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley (1992)
39. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: CVPR, pp. 3506–3513 (2012)
40. Shen, X., Wu, Y.: A unified approach to salient object detection via low rank matrix recovery. In: CVPR, pp. 2296–2303 (2012)
41. Wang, P., Wang, J., Zeng, G., Feng, J., Zha, H., Li, S.: Salient object detection for searched web images via global saliency. In: CVPR, pp. 1–8 (2012)
42. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 29–42. Springer, Heidelberg (2012)

43. Wolfe, J.M., Horowitz, T.S.: Opinion: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5(6), 495–501 (2004)
44. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR, pp. 1155–1162 (2013)
45. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR, pp. 3166–3173 (2013)