Depth-Aware Salient Object Detection Using Anisotropic Center-Surround Difference

Ran Ju, Yang Liu, Tongwei Ren, Ling Ge, Gangshan Wu

State Key Laboratory For Novel Software Technology, Nanjing University, China

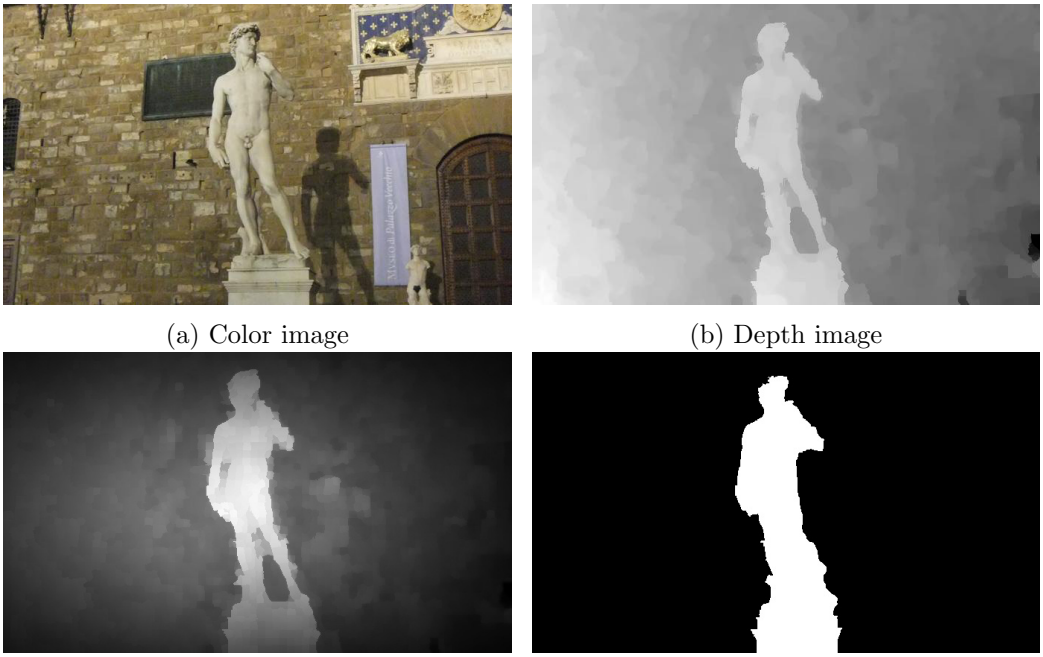
Abstract

Most previous works on salient object detection concentrate on 2D images. In this paper, we propose to explore the power of depth cue for predicting salient regions. Our basic assumption is that a salient object tends to stand out from its surroundings in 3D space. To measure the object-to-surrounding contrast, we propose a novel depth feature which works on a single depth map. Besides, we integrate the 3D spatial prior into our method for saliency refinement. By sparse sampling and representing the image using superpixels, our method works very fast, whose complexity is linear to the image resolution. To segment the salient object, we also develop a saliency based method using adaptive thresholding and GrabCut. The proposed method is evaluated on two large datasets designed for depth-aware salient object detection. The results compared with several state-of-the-art 2D and depth-aware methods show that our method has the most satisfactory overall performance.

Keywords: Salient object detection, depth map, center-surround difference

1. INTRODUCTION

An inherent and powerful ability of human eye is visual attention, which quickly captures the most conspicuous regions from a scene, and passes them to high level visual cortexes. The attention selection reduces the complexity of visual analysis and thus makes human visual system considerably efficient in complex environments. Computational saliency models, which follow the attention mechanism of human eye, occupy an important place in image processing and computer vision society. By saliency analysis, vision tasks are concentrated on a few regions of interests instead of entire images, which benefits many applications in both efficiency and effectiveness, e.g. image



(a) Color image (b) Depth image

(c) Saliency map (d) Salient object

Figure 1: Salient object detection. (a) Original color image. (b) Depth image of (a). Pixels appear brighter are nearer, vice versa. (c) Saliency map generated by the proposed method. Higher brightness indicates the location is more conspicuous. (d) Salient object mask produced by our method. White indicates object and black indicates background.

classification [1, 2], object segmentation [3, 4, 5, 6], image retargeting [7, 8, 9, 10], adaptive image compression [11, 12], content-based image retrieval [13, 14] and quality assessment [15, 16].

There are generally two categories of visual saliency models: fixation prediction [17] and salient object detection [18]. The former model aims at predicting the gaze points, i.e. where people look. The prediction results usually highlight a few spots people are most likely to pay attention to. In contrast, the goal of salient object detection is to detect the entire object that appears most distinctive, which could be more useful for high level processing. In this paper, we focus on the latter model, salient object detection. We give an example in Figure 1. Given a color image with its depth map (first row), our goal is to produce a high resolution saliency map, and extract the object of the greatest interest (second row), which is Statue of David in this example.

Most previous works on salient object detection focus on 2D color images.

However, human perceives the world with not only color but also abundant 3D spatial information. While the saliency researches on 2D color images have been studied a lot and remarkable achievements have been made, little attention is paid to the effect of scene depth for saliency analysis. One reason is that depth data is hard to capture as the devices are expensive in the past. Fortunately, more and more cheap and convenient 3D devices have been developed in recent years, which makes it easier to obtain sufficient depth data to support the research on depth-aware salient object detection. The depth data, no matter captured by stereo [19], ToF (Time-of-Flight) [20] or structured light [21], is usually recorded as a depth map like that shown in Figure 1. (b). Recently, a few tentative researches have shown that depth could be powerful in saliency analysis. For example, Lang [22] et al. integrated depth prior into existing 2D methods and achieved a 6% to 7% increase in predictive power. Niu et al. [23] combined depth contrast and stereo photography knowledge for salient object detection. The result achieved the best evaluation score in their stereo saliency dataset. Inspired from all the above, we further explore the characteristics of depth information for salient object detection according to the following considerations [24]:

Good detection. A good salient object detection method should both miss real salient regions and falsely detect background as salient regions in a low probability. To this end, we propose to measure pixel-wise saliency according to a basic observation: a salient object usually outstands from its surroundings or background from a global view. This assumption is common for natural scenes and daily life, since generally objects in clutters or occluded by background are unlikely to be the most attention-grabbing. To get the contrast between a pixel and its surroundings, we search for background samples in several directions, and measure the center-surround difference as its saliency value. The background samples are estimated using the depth prior. We also set a few weighting parameters to highlight the contrasts in certain directions. Besides, as is well known and usually used in previous works, salient regions are usually nearer to viewers in 3D space and tend to locate at the center of an image. We employ the 3D spatial prior in a probabilistic manner to improve the detection results.

High resolution. Salient object detection methods should produce full resolution saliency maps and retain clear object boundaries. To this end we perform saliency computation in the superpixel granularity, which could both retain the object boundaries and avoid the noises of depth images.

Computational efficiency. Salient object detection is usually regarded

as an early process for vision tasks. High efficiency is required to allow large scale, highly complex computation of following process. Owing to the simple but effective feature we used, our method works very fast and the computational complexity is linear to the image resolution.

For evaluation we compare our method with 8 state-of-the-art 2D methods and 5 representative depth-aware methods. We first built a benchmark including 2000 stereo images with computer generated depth maps and manually labeled groundtruth, which is the largest at present to our knowledge. For a comprehensive evaluation we also compare our method with the others on another large dataset [25] including 1000 RGBD images and groundtruth. In both datasets our method shows superior overall performance to the others. We have also developed an object segmentation method using adaptive thresholding and GrabCut [26]. The segmentation method produces highly accurate binary maps, which shows that our method is competent for salient object detection task.

The contribution of this paper can be briefly stated as follows.

- We proposed a new depth feature for salient object detection, which combines both depth based background estimation and learning based direction contrast weighting.
- We employed the 3D spatial prior to improve the initial saliency map using Bayes method.
- We proposed a new object segmentation method using adaptive thresholding and GrabCut.
- We built a 2000 images dataset for evaluation, which is the largest at present to our knowledge. We also compared the related methods on another large dataset [25] for a comprehensive evaluation.

The rest of this paper is organized as following. In Sect. 2 we give a review of related works on both 2D and depth-aware salient object detection. Then we introduce the proposed saliency model in Sect. 3 and the object segmentation method in Sect. 4. Next, we demonstrate the experimental results and give some discussions of the results in Sect. 5. At last, we give a brief conclusion of our work in Sect. 6.

2. RELATED WORK

Computational saliency models originate from last 80s [27, 28, 29, 30], which is represented by Itti's work [17]. Two research waves were generated from then: attention prediction and salient object detection. The former model concerns predicting where people look. Itti et al. combined cognitive psychology, neuroscience and computer vision to model the problem as a bottom-up, task-independent, feature integration process. After that people began to study human visual attention with the help of eye tracking devices [31, 32, 33]. The fruitful researches on this topic can be found in a few recently published surveys [34, 35].

In this paper we focus on salient object detection, which aims at detecting the most conspicuous object from an image. Achanta et al. [18] first define the task as a binary segmentation problem. They propose to obtain pixel-wise saliency by simply measuring the difference between a pixel's color and the image's mean color. Cheng et al. [36] propose a more sophisticated method using global contrast and iterative GrabCut. These methods are essentially searching for regions with high color distinctness. Some methods [37, 38] also search for regions with high texture distinctness. A different thought is to find background regions instead of salient objects, or use the background priors for salient object detection [39, 40, 41, 42]. These methods usually represent an image as a tree or graph using appearance, spatial information and boundary priors, and then perform salient region selection using path planning or center-surround scheme. Recently, owing to the emergence of several large datasets, some learning based methods [43, 44] try to detect which patches of an image are more likely to be a part of an object. Besides, a few methods try to explore some other cues for salient object detection, like scale [45] and color space [46].

The above methods mostly consider only 2D color information, which is different from human who perceives the world in 3D space. Early in the 2000 Ouerhani et al. [47] have investigated the effect of depth on saliency analysis and found that depth cue can be powerful in predicting human gazes. Jeong et al. [48] directly take depth as a complement cue for saliency detection. In recent years some researchers try to study the effect of depth in saliency analysis with the help of modern depth capturing devices. For example, Lang et al. [22]. built a 3D eye fixation dataset using Kinect [49] to study the power of depth in attention prediction. They integrated depth into 2D methods as a probabilistic posterior and found that the predictive power could be increased

by 6% to 7%. These methods, however, only consider absolute depth while neglecting local structure information. Niu et al. [23] proposed to employ global contrast and photography knowledge to detect salient objects. They built a stereo saliency dataset where the images are taken by stereo cameras and collected from Internet. The results in their dataset showed a significant improvement in both precision and recall measure. However, it searches for regions with highly distinct depth, which may easily miss flat regions inside objects. Peng et al. [25] built an RGBD dataset using Kinect and combined depth and existing 2D models for improvement. For the depth cue, they proposed a multi-contextual feature combining local, global and background contrast to measure pixel-wise saliency. The feature performs a fixed, passive measurement of depth contrast. Differently, our method actively and selectively searches for background samples to measure saliency, which is better adapted for different depth structures and more efficient.

3. DEPTH-AWARE SALIENCY MODEL

Our method arises from a basic observation: a salient object tends to outstand from its surroundings in 3D space. This can be easily interpreted since any object has a definite volume, and a salient object should be different from surroundings in 3D space. Consequently, a salient object usually shows obvious contrast with its surroundings or background in 3D space. To measure the contrast, we developed a depth feature called anisotropic center-surround difference. Furthermore, we employ the spatial prior to improve the results. Our method is applied on a single depth image, which can be captured by stereo, RGBD or other manners.

3.1. Anisotropic Center-Surround Difference

According to our basic assumption, the pixel-wise saliency is measured as the center to background difference:

$$\mathcal{S}_d(p) = \sum_{q \in \mathcal{N}(p)} P_B(q)(I(p) - I(q)) \tag{1}$$

where $\mathcal{S}_d(p)$ is the pixel-wise saliency and $I(p)$ is the image attributes, e.g. RGB or depth value, in pixel p. $\mathcal{N}(p)$ is a neighboring area of p which supplies background candidates. $P_B(p)$ is a background selection function,

which is supposed to be:

$$P_B(q) = \begin{cases} 1, q \in B \\ 0, q \in O \end{cases} \tag{2}$$

where $q \in B$ and $q \in O$ indicate q is a pixel belonging to background and salient object respectively. The function is unknown yet since it is right the solution for salient object detection. Fortunately, we can estimate P_B according to some priors, e.g. regarding image boundary as background [39]. In this paper, we take a more powerful prior by estimating $P_B(q)$ using the conditional probability of background given depth:

$$P_B(q) = P(q \in B | d(q)) = \frac{P(q \in B, d(q))}{P(d(q))} \tag{3}$$

where $q \in B$ indicates that pixel q belongs to background. $d(q)$ is the depth value of pixel q. One can also use other methods to estimate $P_B(q)$, e.g. color prior, color-depth joint distribution, location etc. Since generally salient object shows obvious difference to background in some attributes, Eq. (1) highlights salient object and suppresses background.

It should be noted that the proposed saliency measurement differs from existing contrast based methods [36, 23, 25] in two aspects. First, they fixedly select a few samples to measure center-surround contrast. Different from them, we actively search for surrounding samples using the conditional probability of background. Second, they perform indiscriminate summation of different samples. In contrast, we weight the center-surround difference using the background probability to emphasize inter-class contrast and suppress intra-class difference.

We now consider how to improve the computational efficiency of Eq. (1) since the complexity is $O(N)$ for each pixel and $O(N^2)$ for entire image, where N is the number of image pixels. Fortunately, it can be easily found that the computation is highly redundant. First, according to our basic assumption, we can measure the saliency for each pixel in a local area instead of entire image, since the pixels too distant seem to be effectless. Second, depth image is featured with large areas of nearly constant or smooth values, which enables us to make a sparse sampling around the central pixel for acceleration. Suppose the surrounding area of p is divided into infinite small sections. The depth distribution between neighboring sections are highly similar. Therefore, we select only a few radiuses emitting from p within

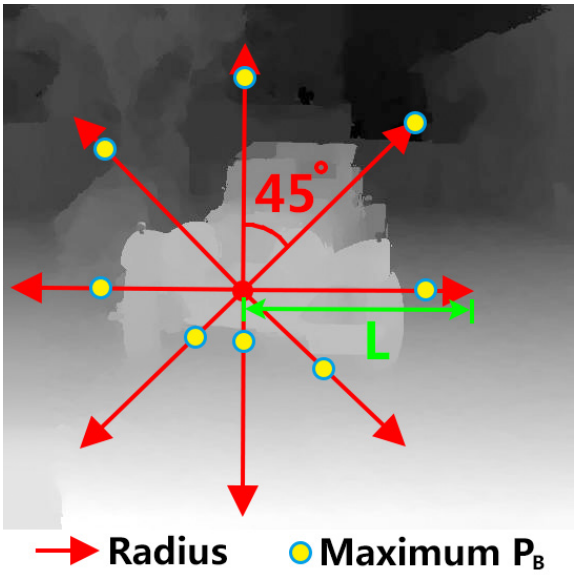


Figure 2: Example of the anisotropic center-surround difference operator. Red arrow line indicates the scan radiuses $\mathcal{R}_i(p)$. Yellow circles are the pixels with maximum background probabilities.

length L as background candidates, which is shown in Fig. 2. Now the pixel-wise saliency can be written as:

$$\mathcal{S}_d(p) = \sum_{q \in \mathcal{R}(p)} P_B(q)(I(p) - I(q)) \tag{4}$$

where $\mathcal{R}(p) = \{\mathcal{R}_1(p), \mathcal{R}_2(p), ...\mathcal{R}_K(p)\}$ indicates K radiuses emitting from p, each of them is limited to length L. In our experiments we set L as a half of the image's diagonal length. We further reduce the sampling by selecting the pixel with the maximum probability to be background along each radius. Finally, given $I(p) = d(p)$ for a depth image, the saliency is computed as an anisotropic center-surround difference (ACSD):

$$\begin{aligned} \mathcal{S}_d(p) &= \sum_q P_B(q)(d(p) - d(q)), \\ q &= \text{argmax}\, P_B(\mathcal{R}_i(p)), i \in [1, K] \end{aligned} \tag{5}$$

Now the computational complexity for each pixel is $O(KL)$. We will show how to further reduce the complexity using superpixels in Sect. 3.4.

An example of ACSD with 8 scan radiuses is shown in Fig. 2. In each radius, the pixel with maximum background probability P_B is shown in yellow circle. The center point gets a high saliency score as it appears outstanding in all the radiuses. What we concerns more is the ground that extends from far to near. Obviously the distant background gets very low ACSD value and thus looks inapparent. The nearer part of the ground, which is located at the bottom of the image, has a definitely high depth value. However, it shows less conspicuity because it gets high ACSD values only in the upper three directions. In the horizontal and lower directions it is suppressed effectively.

3.2. Weighted ACSD

Generally the center-surround differences may contribute different saliency scores in different directions. For example, humans are more likely to notice the depth difference in horizontal direction than in vertical direction because human eyes are arranged horizontally. To simulate the property, we improve the original ACSD measure using a weighting function:

$$\begin{aligned} \mathcal{S}_w(p) = \sum_q w_i P_B(q)(d(p) - d(q)), \\ q = \text{argmax}\, P_B(\mathcal{R}_i(p)), i \in [1, K] \end{aligned} \tag{6}$$

where w_i is a weighting parameter to control the power of contribution in radius i, as illustrated in Fig. 3. We learn the parameter $\boldsymbol{w} = (w_1, w_2, ..., w_K)^T$ from a training dataset with groundtruth salient object masks. The weighted ACSD in Eq. (6) can be rewritten as:

$$\mathcal{S}_w(p) = \sum_{i \in [1,K]} w_i D(p, i) \tag{7}$$

$$\begin{aligned} D(p, i) =& P_B(q)(d(p) - d(q)), \\ & q = argmax P_B(\mathcal{R}_i(p)) \end{aligned} \tag{8}$$

Then the desired $\boldsymbol{w}$ is supposed to minimize the error between $\mathcal{S}_w(p)$ and groundtruth $G(p)$:

$$\begin{aligned} \boldsymbol{w} = \text{argmin} \sum_p \|\boldsymbol{w^T D(p)} - G(p)\|_2 \\ s.t. \quad w_i >= 0, i \in [1, K] \end{aligned} \tag{9}$$

where $\boldsymbol{D(p)} = (D(p, 1), D(p, 2), ..., D(p, K))^T$ and $\boldsymbol{w}$ is forced to be non-negative. The minimization problem is solved using the fast non-negative least square method [50]. For details please refer to our experiments.

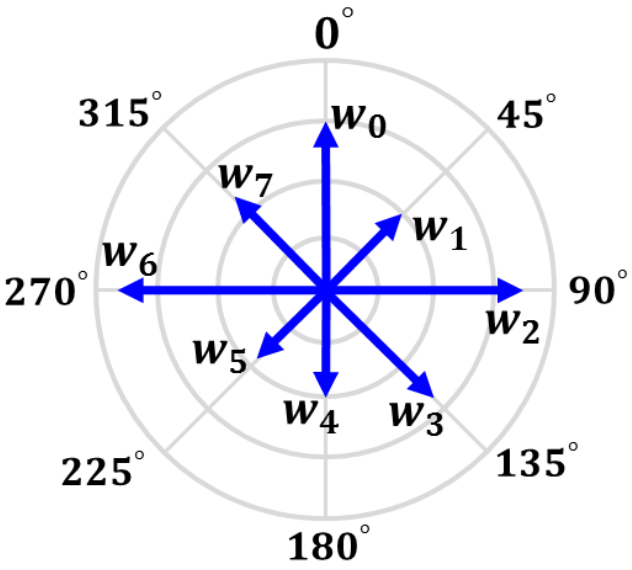


Figure 3: An example of $\boldsymbol{w}$ with 8 radiuses. The length of each blue arrow line indicates the weight in each direction.

3.3. Saliency Refinement With 3D Spatial Prior

Spatial priors are well known and frequently used for saliency analysis [48, 32, 22, 51, 4]. For example, in 3D space nearer regions appear more salient than distant ones. Besides, people like to place the object of interest near the center of a photograph. The spatial priors are integrated into our method in a probabilistic manner. Given the 3D location $(p_{x,y}, d_p)$ of a pixel p, where $p_{x,y}$ represents its 2D image coordinate, the salient object posterior probability can be estimated as:

$$\mathcal{S}_p(p) = P(p \in O | p_{x,y}, d_p) \tag{10}$$

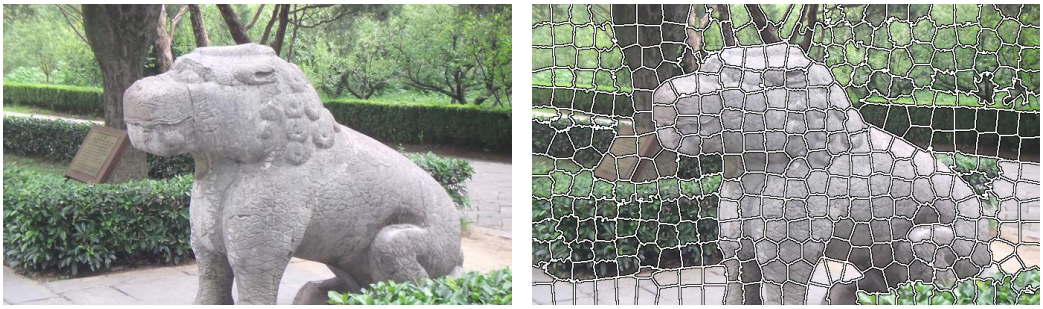
where $p \in O$ indicates pixel p belongs to the salient object. Suppose $p_{x,y}$ is independent of d_p, the posterior probability based saliency can be written as:

$$\mathcal{S}_p(p) = \frac{P(p_{x,y}|p \in O)P(d_p|p \in O)P(p \in O)}{P(p_{x,y})P(d(p))} \tag{11}$$

Then we refine the weighted ACSD saliency by multiplying $\mathcal{S}_p$:

$$\mathcal{S}(p) = \mathcal{S}_w(p)\mathcal{S}_p(p) \tag{12}$$

where $\mathcal{S}(p)$ is the final saliency score.



(a) Color image (b) Superpixels

Figure 4: Superpixel segmentation. The ACSD score is calculated at the centroid of each superpixel.

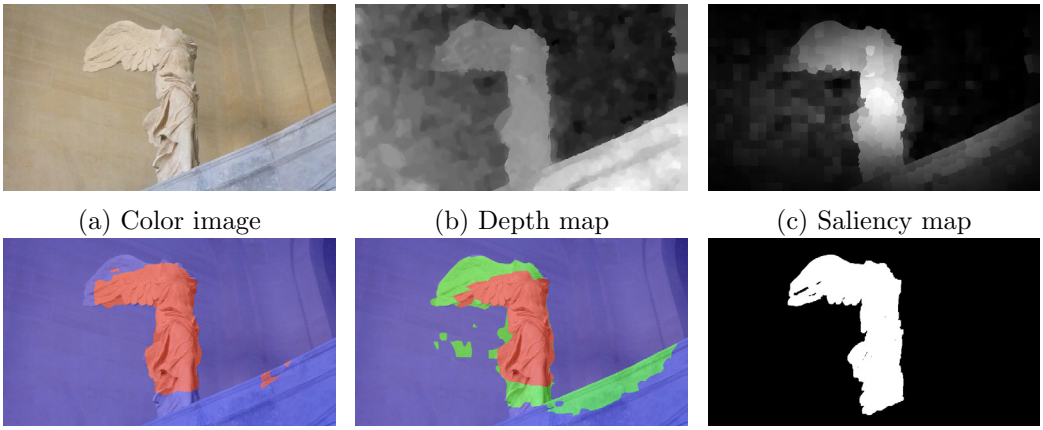
3.4. Implementation details

To deal with noises and errors in depth images, we perform saliency detection on the granularity of superpixels, which is generated using SLIC [52] on the color images. We show an example in Fig. 4. The superpixels are regular regions which preserve local structures well. Therefore, an average operation would restrain the influence of noises and errors effectively while keeping the saliency precision well. For each superpixel, we compute its saliency according to Eq. (12) at the centroid. The depth value of the centroid is calculated as the mean depth over the superpixel.

The number of superpixels has a considerable influence to the performance, since large superpixels tend to miss the details of salient objects while small superpixels introduce more depth noises and take more computation time. To this problem we set the number of superpixels as the length of diagonal to fit in with the image size. After computing the saliency value for each superpixel, the final saliency is rescaled to $[0, 255]$ and assigned to each pixel.

3.5. Complexity Analysis

SLIC [52] is $O(N)$ complex and generate $O(X)$ superpixls where X equals to the diagonal length. Saliency computation is also $O(X)$ complex as the scan length is limited to L. And thus the saliency computation is $O(X^2)$ complex. Suppose the aspect ratio of the image is r, we can get $X^2 = \frac{r^2+1}{r}N$. Therefore, we conclude that our method works within an $O(N)$ complexity.



(a) Color image (b) Depth map (c) Saliency map

(d) Adaptive thresholding (e) Tri-map (f) Binary mask

Figure 5: Salient object segmentation.

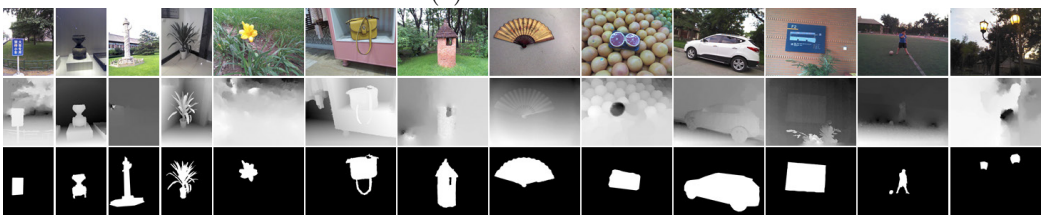
4. SALIENT OBJECT SEGMENTATION USING ADAPTIVE THRESHOLDING AND GRABCUT

Given the saliency map, the next end is to cut the salient object out. The result is given in the form of a binarized salient object mask. In some previous works the segmentation is done by fixed thresholding [18], where the threshold is supposed to give an optimum trade-off between precision and recall. However, due to the inaccuracy and irregularity of saliency maps, the segmentation result would be poor by simply fixed thresholding. Besides, in a large collection of images it is difficult to select a fixed threshold that fits every image well. To solve the problem, we perform the segmentation similar to [53] and [36] but differed in seeds generation. We first calculate an adaptive threshold T_b using Otsu [54] for each image on its saliency map. For a good saliency map, the salient object and background usually distribute on separate sides in gray levels. Consequently, the adaptive thresholding gives a coarse estimation of object and background. Then we generate a tri-map according to T_b to handle the errors in saliency maps:

$$L(p) = \begin{cases} O, & \mathcal{S}(p) > T_b^+ \\ B, & \mathcal{S}(p) \leq T_b^- \\ U, & otherwise \end{cases} \tag{13}$$

where $L(p)$ is the label of pixel p. O and B are object and background seeds respectively. U indicates the labeling is unsure. The threshold T_b^+ and T_b^-

(a) NJUDS2000



(b) RGBD1000

Figure 6: Examples of the two datasets for evaluation. (a) NJUDS2000. (b) RGBD1000. First row: color image. Second row: depth map. Third row: manually labeled groundtruth.

are used to adjust the size of the unsure region, which can be set manually. In our experiments, we set the two thresholds adaptively according to T_b. At last, we feed the tri-map to the GrabCut [26] framework to produce a high quality salient object mask.

We give an example of our salient object segmentation process in Fig. 5. In both depth and saliency maps there are errors. However, we can get a coarse estimation by adaptive thresholding since the saliency map highlights most salient regions and suppresses most background. Then a tri-map is generated and fed to GrabCut as seeds. Owing to the high quality seeds labeling, an accurate binary segmentation is produced.

5. EXPERIMENTS AND ANALYSIS

5.1. Datasets and Experimental Settings

We evaluate our method on two datasets designed for depth-aware salient object detection: NJUDS2000 and RGBD1000 [25]. The former one includes 2000 stereo images, which is an extension of the previous version of this work [55]. We extend the dataset to increase the diversity of objects and scenes and to cover more complex and challenging cases. The images are collected from Internet, 3D movies and photographs taken by a Fuji W3 stereo

camera. We recover the depth maps using Sun's optical flow method [56]. Then we invite four volunteers to label the groundtruth salient object masks according to the procedure introduced in [43]. To our knowledge it is the biggest dataset for depth-aware salient object detection at present. The second dataset is built by Peng et al. [25], which includes 1000 RGBD images captured by Kinect, along with manually labeled groundtruth masks. We give a few examples of the two datasets in Fig. 6. It should be noted that in both datasets some depth maps are quite misleading (The 6th and 9th column in Fig. 6 (a), and the 3rd and last column in Fig. 6 (b)), which is a limitation of current depth perception techniques. Fortunately, for most images in the datasets the depth information is sufficiently accurate for salient object detection.

We compare our method with 8 state-of-the-art 2D methods: FT [18], RC [36], PCA [37], CA [38], AMC [41], MR [40], UL [44], HS [45], and 5 representative depth-aware saliency methods namely DM (original depth map, which has been employed in [48]), CV [57], SS [23], DP [22], SD [25]. Our method is implemented in C++. For the other methods, we use the authors' source codes for evaluation except for CV, SS and DP. We realize the three algorithms as we failed to get the authors' implementations. For saliency map comparison we employ the widely used precision-recall curve. Specifically, we obtain a binary image from each saliency map using a gradually increasing threshold d_t from 0 to 255, and then compare with the groundtruth salient object mask to get a precision and recall:

$$Precision = \frac{|\{p_i|d(p_i) \geq d_t\} \cap \{p_g\}|}{|\{p_i|d(p_i) \geq d_t\}|} \tag{14}$$

$$Recall = \frac{|\{p_i|d(p_i) \geq d_t\} \cap \{p_g\}|}{|\{p_g\}|} \tag{15}$$

where $\{p_i|d(p_i) \geq d_t\}$ indicates the set that binarized from a saliency map using threshold d_t. $\{p_g\}$ is the set of pixels belonging to groundtruth salient object. The precision-recall curve is plotted by connecting the P-R scores for all thresholds. We use Cheng's evaluation code [58] to get the P-R curves of different methods. For segmentation evaluation we use precision, recall and F-measure, where F-measure is defined as:

$$F_\beta = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{16}$$

Table 1: 16 direction weights trained in 100 randomly selected depth images.

Direction	0°	22.5°	45°	67.5°	90°	112.5°	135°	157.5°
Weight	0.0657	0.0587	0.0162	0.0922	0.1900	0.0789	0.0503	0.0243
Direction	180°	202.5°	225°	247.5°	270°	292.5°	315°	337.5°
Weight	0.0375	0.0884	0.0000	0.0967	0.2111	0.0000	0.0549	0.0671

where β^2 is set to 0.3 according to [36] to favor precision more than recall. Throughout our experiments the evaluation is performed on a machine with a 2GHz Intel Xeon E5-2620 CPU and 32GB memory. In the running time evaluation, all parallel executions are disabled for the sake of fairness.

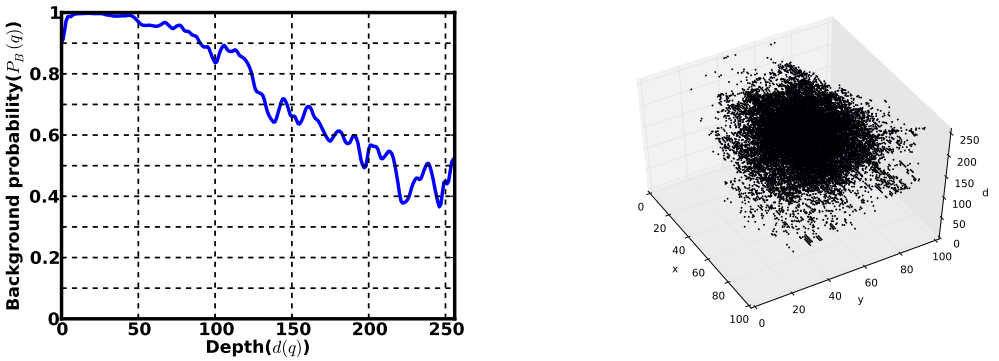
Our C++ source code and evaluation dataset including original stereo images, depth maps and manually labeled groundtruth masks are available at http://mcg.nju.edu.cn/member/jur/index.html.

5.2. Parameter Settings

We set the number of scan radiuses as 16 to make a trade-off between efficiency and accuracy. To learn the weight for each direction, we randomly select 100 depth images with groundtruth as training set. We first calculate $\boldsymbol{D(p)}$ in all directions. $\boldsymbol{G(p)}$ is obtained from the groundtruth images. Then we solve for the optimal $\boldsymbol{w}$ in Eq. 9 using [50]. A typical result is given in Table 1. The result shows that generally contrasts in horizontal directions are stronger than those in vertical directions, and thus contribute more in salient object detection.

We calculate the background prior $P_B(q)$ and the spatial posterior probability $\mathcal{S}_p(p)$ on the same training set as above. Each image is resized to 100×100 for coordinates normalization. Then we compute the two probabilities according to Eq. 3 and Eq. 11. The background prior is illustrated in Fig. 7 (a). We can find that the smaller depth, the more likely a pixel turns to be background. It should be noted that when $d(q) \leq 50$ the probabilities are very close to 1, which gives very strong predictions of background. The 3D spatial distribution of salient object are shown in Fig. 7 (b). It can be observed that salient object tends to locate at the center of an image, and nearer to viewers in depth.

The two thresholds T_b^+ and T_b^- for salient object segmentation are set according to the following rule. Suppose a saliency map is divided into two sets by T_b, where the higher part O_T indicates salient object and the lower part B_T stands for background. We set T_b^+ to make the threshold range



(a) Background probability given depth. (b) 3D spatial distribution of salient object.

Figure 7: Background and 3D spatial priors.

$(T_b, T_b^+]$ recall 60% pixels of O_T. Similarly, for T_b^- we set the range $(T_b^-, T_b]$ to cover 30% pixels of B_T. The unsure region also contributes to color modeling but will be relabeled after GrabCut [26].

5.3. Saliency Map Results and Discussion

We show a few saliency maps generated by different methods in Fig. 8. Due to space limitation we only show the results of some representative and most recent 2D and depth-aware methods. From the figure we can find that it is difficult to extract the entire salient objects using only color information due to the interference of background, textures and shading. For example, in the first row, the fighters are cluttered in the blue sky with white clouds in the color image, which makes the 2D methods quite confusing. And in the second row, each component of the car has a different color. This leads to an incomplete detection of salient objects using only color information. In certain cases, even the salient object has an uniform color itself, it may be under different illuminations such as in the 5th row. This also results in false negatives. However, in the depth images it is clear to see the salient region since the depth information is free from texture and illumination, which makes it easier to detect entire salient objects.

We further give the precision-recall curves in Fig. 9. The evaluations on RGBD1000 [25] and our NJUDS2000 dataset are shown in the left and right column respectively. In the first row we show the comparison with state-of-the-art 2D methods. The P-R curves of depth-aware methods are shown in the second row. From the statistical evaluation we can find our method

Figure 8: Saliency maps generated by different methods.

outperforms all the other competitors on NJUDS2000 and is top ranked on RGBD1000. For the RGBD1000 dataset, when the recall increases, SD, AMC and MR have less false positives than our method. This is mainly caused by misleading depth maps and flat images. The former case makes depth-aware methods completely confusing, which usually happens in outdoor scenes as Kinect can't work well under strong sunlight. A few examples are given in the 3rd and last column of Fig. 6 (b). The latter case supplies little depth information of the salient object, such as a poster pasted on a wall, or like those shown in 8th and 11th column of Fig. 6 (b). The two cases could be regarded as the limitation of depth-aware methods. However, our method still produces comparable results with the other top ranked methods, which shows that our method can explore the power of depth more effectively. On NJUDS2000 our method greatly improves the performance over the other methods because most stereo images supply sufficient depth information for display. The 2D methods are limited in detecting salient object with different colors and shading, especially in complex scenes. The depth-aware methods DM and DP assume nearer regions are more salient, which fails to handle near background. CV detects corners and edges with high depth curvatures and thus is incapable for salient object detection. SS and SD also take relative depth difference into consideration. However, they tend to improve detection precision at the cost of generating many false negatives, as shown in Fig. 8.

Running time comparison. We give the running times of different meth-

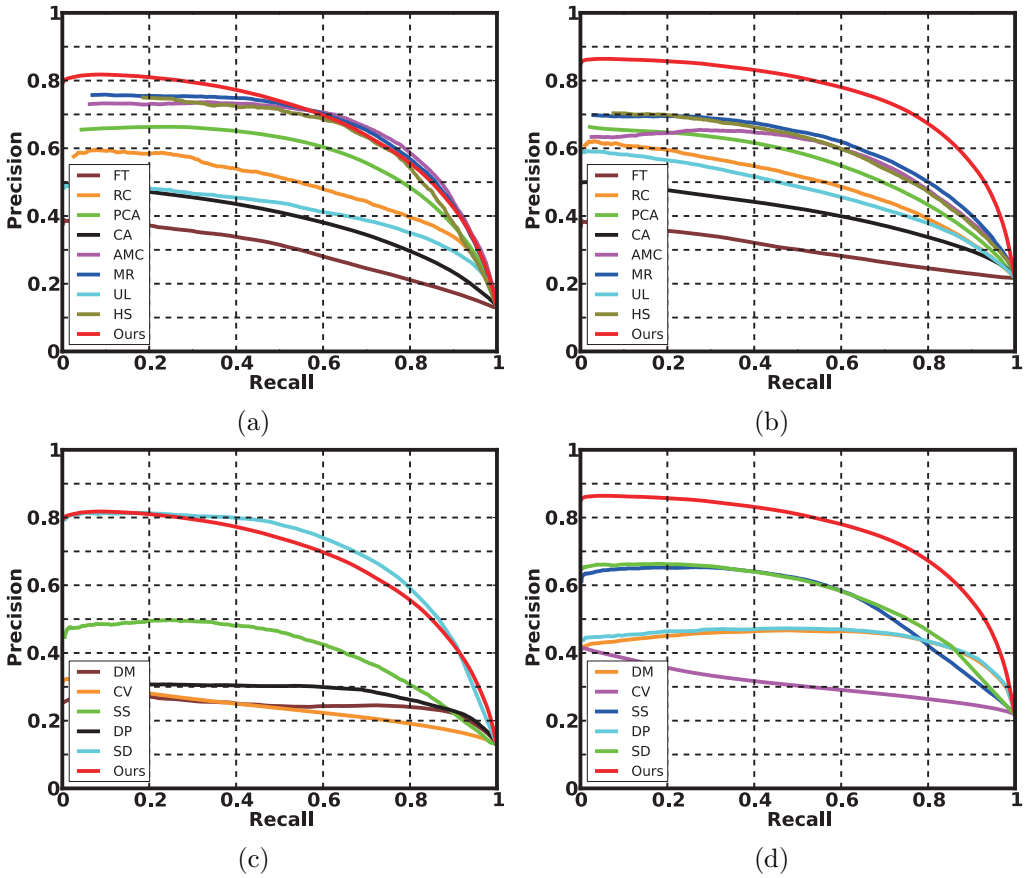


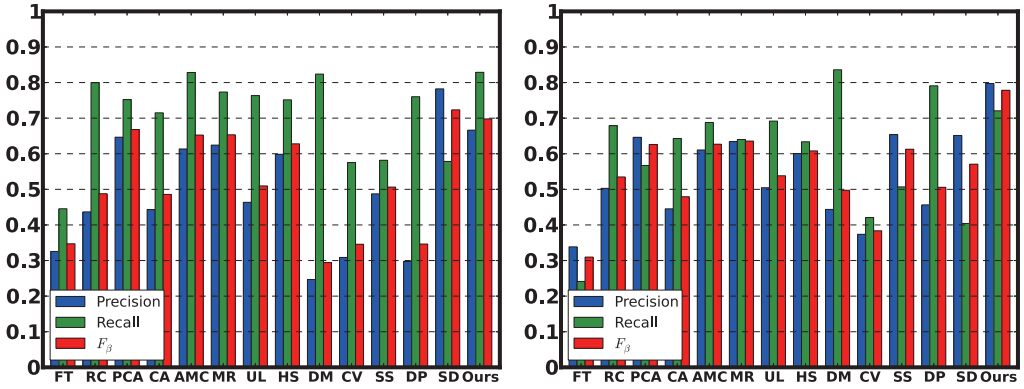
Figure 9: Precision-recall curves of different methods. (a) Comparison with 2D methods on RGBD1000. (b) Comparison with 2D methods on NJUDS2000. (c) Comparison with depth-aware methods on RGBD1000. (d) Comparison with depth-aware methods on NJUDS2000.

ods in Table. 2. The result is averaged over all images of the two datasets. The average image resolution for NJUDS2000 is 500×726, and 528×592 for RGBD1000. Our method takes $0.672s$ in average. Specifically, the superpixel segmentation takes 0.597s and saliency computation takes 0.075s. DP, CV, FT, RC, AMC and our method all run very fast and take less than one second in average. Among these methods, DP, CV, FT perform poorly in the two datasets though they work very fast. Considering of both detection results and running time, our method gives the most satisfactory performance overall.

Table 2: Average running time comparison.

Method	FT [18]	RC [36]	PCA [37]	CA [38]	AMC [41]	MR [40]	UL [44]
Time(s)	0.585	0.727	29.614	108.6	0.951	4.839	449.14
Code	Matlab	C++	Matlab	Matlab	Matlab	Matlab	Matlab
Method	HS [45]	DM	CV [57]	SS [23]	DP [22]	SD [25]	Ours
Time(s)	2.007	-*	0.133	12.091	0.039	7.537	0.672
Code	C++	-*	Matlab	C++	Matlab	Matlab	C++

* The running time of DM is not given since it directly takes the depth maps as saliency results without any processing.

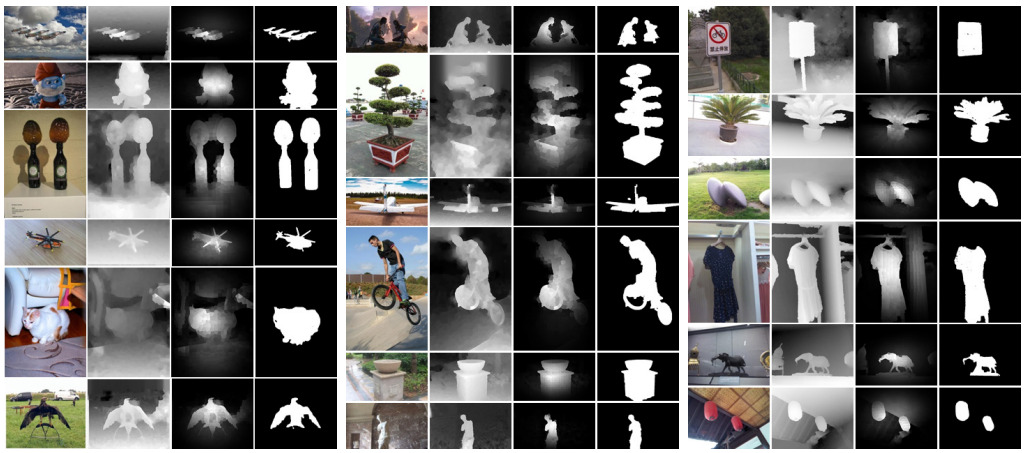


(a) (b)

Figure 10: Statistical comparison of segmentation results. (a) Results comparison on RGBD1000. (b) Results comparison on NJUDS2000.

5.4. Segmentation Results and Discussion

Our segmentation method is implemented in C++ and the average running time is 1.164s. The statistical comparison of segmentation results are shown in Fig 10. It can be found that the statistics are approximately agreed with the P-R curve evaluation in Fig 9, since generally better saliency map leads to more accurate object mask. On both datasets, DP and DM are featured for their abnormal high recalls and low precisions. This is mainly caused by that they simply assume near regions are salient, which erroneously highlight lots of near background regions. On the contrary, SD is featured for its high precision and low recall, as it generates many false negatives in the saliency map. In contrast, our result shows well-balanced performance in precision, recall and F-measure. Specifically, on NJUDS2000 and RGBD1000 our method ranks the 1st and 2nd respectively in precision, 3rd and 1st in



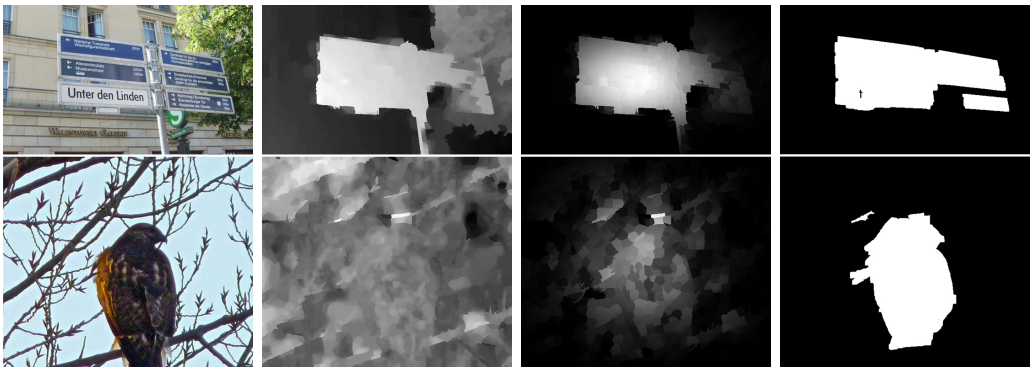
(a) Results on NJUDS2000 (b) Results on NJUDS2000 (c) Results on RGBD1000

Figure 11: Salient object segmentation results. In each group, from left to right: original color image, depth map, saliency map and salient object mask generated by our method.

recall, and 1st and 2nd in F-measure.

We give a few salient object segmentation results of our method in Fig 11. It could be found that our method works well not only in simple scenes with a single salient object, but also in complex cases. In the 1st row of (a) and (b) the background is cluttered. In the 5th row of (a) the background has a similar region to the salient object. There are also salient regions with disconnected components as shown in the 3rd row of (a) and 1st row of (b). Besides, salient objects may have different shading or components with different colors, as shown in the last row of (b), 1st row and 3rd row of (c). Even for the depth maps with serious noises as shown in the 2nd row of (b), our method shows desirable tolerance to errors.

Limitations. We have explored the power of depth for salient object detection and demonstrated its effectiveness by experiments. However, it could fail to work when depth is quite inaccurate, as we have shown in Fig 6. We give a few failure examples generated by our method in Fig 12. In these cases, the depth maps are accompanied with serious errors, and thus lead to unsatisfactory saliency and segmentation results. Fortunately, with the advance of depth capturing techniques, the problem will be hopefully solved in the near future.



(a) Color image (b) Depth map (c) Saliency map (d) Segmentation

Figure 12: Failure cases of depth cue. For depth maps with serious noises and errors, depth based methods will produce suboptimal saliency and segmentation results.

6. CONCLUSION

We proposed a salient object detection method that works on depth images. The method is based on a simple but effective assumption that salient objects tend to stand out from surroundings. We developed a new depth feature called anisotropic center-surround difference to measure the contrast between a region and its surroundings. Besides, the spatial priors are combined for saliency refinement. We also implemented a salient object segmentation method using adaptive thresholding and GrabCut. Our method is fast and works within a linear complexity. The experiments demonstrated that our method can be used for rapid and accurate salient object detection, and consequently could benefit many high level applications.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation of China under Grant No.61321491 and No.61202320, Research Project of Excellent State Key Laboratory (No.61223003), Natural Science Foundation of Jiangsu Province (No.BK2012304) and National Special Fund (No.2011ZX05035-004-004HZ).

References

[1] G. Sharma, F. Jurie, C. Schmid, Discriminative spatial saliency for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3506–3513.

[2] L.-K. Wong, K.-L. Low, Saliency-enhanced image aesthetics class prediction, in: Image Processing (ICIP), 2009 16th IEEE International Conference on, IEEE, 2009, pp. 997–1000.

[3] F. Meng, H. Li, G. Liu, K. N. Ngan, Object co-segmentation based on shortest path algorithm and saliency model, Multimedia, IEEE Transactions on 14 (5) (2012) 1429–1441.

[4] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, Z. Zhang, Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut, Multimedia, IEEE Transactions on 14 (4) (2012) 1275–1289.

[5] R. Ju, X. Xu, Y. Yang, G. Wu, Stereo grabcut: Interactive and consistent object extraction for stereo images, in: Advances in Multimedia Information Processing–PCM 2013, Springer, 2013, pp. 418–429.

[6] X. Xu, W. Geng, R. Ju, Y. Yang, T. Ren, G. Wu, Obsir: Object-based stereo image retrieval, in: Multimedia and Expo (ICME), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–6.

[7] O. Le Meur, X. Castellan, P. Le Callet, D. Barba, Efficient saliency-based repurposing method, in: Image Processing, 2006 IEEE International Conference on, IEEE, 2006, pp. 421–424.

[8] M. Grundmann, V. Kwatra, M. Han, I. Essa, Discontinuous seam-carving for video retargeting, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 569–576.

[9] T. Ren, Y. Liu, G. Wu, Rapid image retargeting based on curve-edge grid representation, in: Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE, 2010, pp. 869–872.

[10] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, M. Gross, Nonlinear disparity mapping for stereoscopic 3d, ACM Transactions on Graphics (TOG) 29 (4) (2010) 75.

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

[12] Z. Li, S. Qin, L. Itti, Visual attention guided bit allocation in video compression, Image and Vision Computing 29 (1) (2011) 1–14.

[13] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, S.-F. Chang, Mobile product search with bag of hash bits and boundary reranking, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3005–3012.

[14] Y. Yang, L. Yang, G. Wu, S. Li, A bag-of-objects retrieval model for web image search, in: Proceedings of the 20th ACM international conference on Multimedia, ACM, 2012, pp. 49–58.

[15] X. Feng, T. Liu, D. Yang, Y. Wang, Saliency based objective quality assessment of decoded video affected by packet losses, in: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, IEEE, 2008, pp. 2560–2563.

[16] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, P. Bonnet, A metric for no-reference video quality assessment for hd tv delivery based on saliency maps, in: Multimedia and Expo (ICME), 2011 IEEE International Conference on, IEEE, 2011, pp. 1–5.

[17] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.

[18] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 1597–1604.

[19] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International journal of computer vision 47 (1-3) (2002) 7–42.

[20] S. B. Gokturk, H. Yalcin, C. Bamji, A time-of-flight depth sensor-system description, issues and solutions, in: Computer Vision and Pattern

Recognition Workshop, 2004. CVPRW'04. Conference on, IEEE, 2004, pp. 35–35.

[21] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 1, IEEE, 2003, pp. I–195.

[22] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, S. Yan, Depth matters: Influence of depth cues on visual saliency, in: European Conference on Computer Vision, Springer, 2012, pp. 101–115.

[23] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 454–461.

[24] A. Borji, M. M. Cheng, H. Jiang, J. Li, Salient object detection: A survey, Submit to Pattern Analysis and Machine Intelligence, IEEE Transactions on.

[25] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, Rgbd salient object detection: a benchmark and algorithms, in: European Conference on Computer Vision (ECCV), 2014, pp. 92–109.

[26] C. Rother, V. Kolmogorov, A. Blake, Grabcut: Interactive foreground extraction using iterated graph cuts, in: ACM Transactions on Graphics (TOG), Vol. 23, ACM, 2004, pp. 309–314.

[27] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, in: Matters of intelligence, Springer, 1987, pp. 115–141.

[28] B. A. Olshausen, C. H. Anderson, D. C. Van Essen, A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information, The Journal of Neuroscience 13 (11) (1993) 4700–4719.

[29] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artificial intelligence 78 (1) (1995) 507–545.

[30] E. Niebur, C. Koch, Computational architectures for attention, The attentive brain (1998) 163–186.

[31] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, Pattern Analysis and Machine Intelligence, IEEE Transactions on 28 (5) (2006) 802–817.

[32] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Computer Vision, 2009 IEEE 12th international conference on, IEEE, 2009, pp. 2106–2113.

[33] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, T.-S. Chua, An eye fixation database for saliency detection in images, in: Computer Vision–ECCV 2010, Springer, 2010, pp. 30–43.

[34] A. Borji, D. N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study, Image Processing, IEEE Transactions on 22 (1) (2013) 55–69.

[35] A. Borji, L. Itti, State-of-the-art in visual attention modeling, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (1) (2013) 185–207.

[36] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, S.-M. Hu, Global contrast based salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 409–416.

[37] R. Margolin, A. Tal, L. Zelnik-Manor, What makes a patch distinct?, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013, pp. 1139–1146.

[38] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (10) (2012) 1915–1926.

[39] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 29–42.

[40] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 3166–3173.

[41] B. Jiang, L. Zhang, H. Lu, C. Yang, M.-H. Yang, Saliency detection via absorbing markov chain, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 1665–1672.

[42] Z. Liu, W. Zou, O. Le Meur, Saliency tree: a novel saliency detection framework., IEEE transactions on image processing: a publication of the IEEE Signal Processing Society 23 (5) (2014) 1937–1952.

[43] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2) (2011) 353–367.

[44] P. Siva, C. Russell, T. Xiang, L. Agapito, Looking beyond the image: Unsupervised learning for object saliency and detection, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 3238–3245.

[45] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 1155–1162.

[46] J. Kim, D. Han, Y.-W. Tai, J. Kim, Salient region detection via high-dimensional color transform, Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on.

[47] N. Ouerhani, H. Hugli, Computing visual attention from scene depth, in: International Conference on Pattern Recognition, Vol. 1, IEEE, 2000, pp. 375–378.

[48] S. Jeong, S.-W. Ban, M. Lee, Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment, Neural Networks 21 (10) (2008) 1420–1430.

[49] Kinect for xbox 360, Microsoft Corp. Redmond WA., 2010.

[50] R. Bro, S. De Jong, A fast non-negativity-constrained least squares algorithm, Journal of chemometrics 11 (5) (1997) 393–401.

[51] A. Borji, D. N. Sihite, L. Itti, Salient object detection: A benchmark, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 414–429.

[52] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2274–2282.

[53] Y. Fu, J. Cheng, Z. Li, H. Lu, Saliency cuts: An automatic approach to object segmentation, in: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE, 2008, pp. 1–4.

[54] N. Otsu, A threshold selection method from gray-level histograms, Automatica 11 (285-296) (1975) 23–27.

[55] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, Image Processing (ICIP), 2014 IEEE Conference on.

[56] D. Sun, S. Roth, M. J. Black, Secrets of optical flow estimation and their principles, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2432–2439.

[57] C. H. Lee, A. Varshney, D. W. Jacobs, Mesh saliency, in: ACM Transactions on Graphics, Vol. 24, ACM, 2005, pp. 659–666.

[58] M.-M. Cheng, Ming-ming cheng's open source projects, https://github.com/MingMingCheng/CmCode/.