A Three-pathway Psychobiological Framework of Salient Object Detection Using Stereoscopic Technology

Chunbiao Zhu
SECE, Shenzhen Graduate School
Peking University
Shenzhen, China
zhuchunbiao@pku.edu.cn

Ge Li*
SECE, Shenzhen Graduate School
Peking University
Shenzhen, China
*geli@ece.pku.edu.cn

Abstract

Saliency detection, finding the most important parts of an image, has become increasingly popular in computer vision. Existing proposal methods are mostly based on color information, which may not be effective for cluttered backgrounds. We propose a new algorithm leveraging stereopsis to generate optical flow which can obtain addition cue (depth cue) to get the final saliency map. The proposed framework consists of three pathways. The first pathway eliminates the background based on cellular automata. The second pathway gets the optical flow and color flow saliency map. The third pathway calculates a coarse saliency map. Finally, we fuse these three pathways to generate the final saliency map. Besides, we construct a new high-quality dataset with the complex scene to make computer challenge human vision. Experimental results on our dataset and another three popular datasets demonstrate that our method is superior to the existing methods in terms of robustness.

1. Introduction

Saliency detection is a process of getting the visual attention region precisely. The attention is the behavioral and cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, which responses how we actively process specific information in our environment.

Early works on computing saliency aim to locate the visual attention region. Recently the field has been extended to locate and refine the salient regions and objects. Served as a fundamental of various multimedia applications [2, 5, 5, 15], salient object detection is widely used in content-aware editing, image retrieval, object recognition, object segmentation, compression, image retargeting, etc.

In general, saliency detection algorithms mainly use top down or bottom-up approaches. Top-down approaches are task-driven and need supervised learning. While bottom-up approaches usually use low-level cues, such as color features, distance features and other heuristic saliency features. The most used features are heuristic saliency features and discriminative saliency features. Various measures based on heuristic saliency features have been proposed, including pixel-based or patch-based contrast, region-based contrast, pseudo-background, and similar images. Notwithstanding the demonstrated success, existing RGB-based algorithms [9, 16] may become ineffective when objects cannot be easily separated from the background (e.g., objects with similar colors to the background). In this case, additional cues are required as a complement to detect the objects from the image.

Recently, advances in the rapid deployment of stereoscopic equipment have motivated the adoption of structural features, improving discrimination among different objects with the similar appearance. As a complement to color images, a stereo image pair is utilized to obtain rough depth and edge correspondences for the two images. Some algorithms [3, 6, 4, 14, 1] adopt depth cue which is generated by stereo image pairs to deal with the complex scenarios. In [3], Cheng et al. compute salient stimuli in both color and depth spaces. In [6], Ju et al. propose a saliency method that works on depth images based on the anisotropic center-surround difference. In [4], Guo et al. propose a salient object detection method for RGB-D images based on evolution strategy. Their results show that stereo saliency is a useful consideration compared to previous visual saliency analysis. In [1], an active vision system is presented for visual scene segmentation based on the integration of several cues. These methods combine a set of foveal and peripheral cameras. Object hypotheses are generated through a stereo based fixation process. In [14], depth cues are used to enhance the salient object detection results. All of them demonstrate the effectivity of depth cue in the improvement of salient object detection.

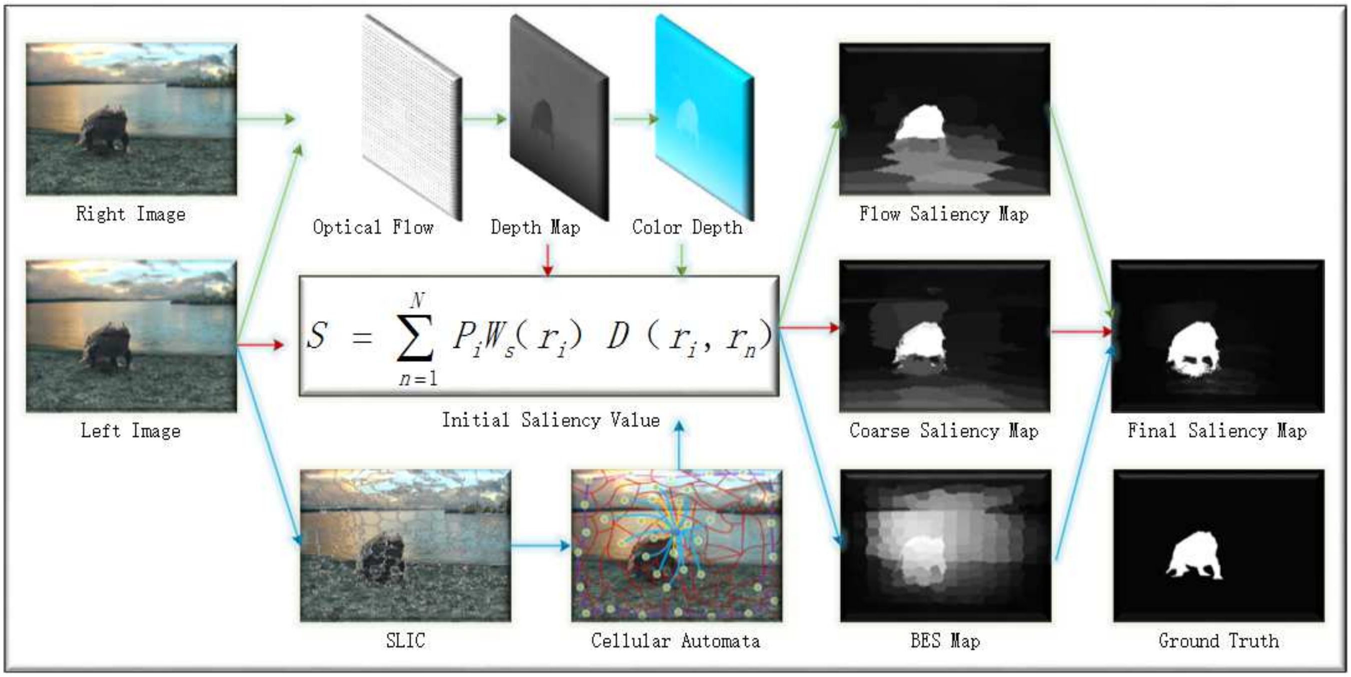


Figure 1. The proposed biologically motivated framework for automatically producing saliency map from stereo image pairs.

Although, those approaches can enhance salient object region. It is very difficult to produce good results when a salient object has low depth contrast compared to the background.

In this work, we propose a biologically motivated framework for automatically producing saliency map from stereo image pairs. We imitate the processing of the human visual perception. When the human see a scene, first, the eyes will focus on the center and ignore the background. Then the eyes will focus on the object in front of the scene. Finally, the eyes will focus on the salient objects by distinguishing some features different. Inspired from the intuition, we construct a new saliency detection framework comprising of three parts, namely background elimination component, optical flow saliency component assisted with the stereopsis and features difference component, to imitate this visual perception processing. The imitation process is shown in Fig. 1.

To evaluate the proposed method, we have constructed a new dataset of 100 stereo image pairs. Extensive experimental evaluations show that the proposed method achieves superior performance than the existing methods tested on this new dataset.

The rest of this paper is organized as follows. Section 2 elaborates the related works. Section 3 describes details of the biologically motivated stereo saliency detection framework. The following section presents the experimental results of our algorithm on three datasets. Finally, Section 5 concludes this paper.

2. Proposed Algorithm

The proposed stereo-based approach has four main parts: Coarse Saliency Map Generation, Flow Saliency Map Generation, Background Elimination Map Generation and Final Saliency Map Generation.

2.1. Color Depth Map Calculation

Given a stereo image pair, we can easily obtain a rough depth map using image correspondences. Since most of the stereo matching algorithms may not be reliable in complex scenes, due to large depth range or flat regions, we first use Flow [13] to generate the flow map, which is robustness in both indoor and outdoor scenes.

Then we get the depth map according to the flow map. We get the depth map by utilizing the RGB information. Specifically, SLIC [8] is applied to over-segment the color image. As the color image and the rough depth map are aligned, we calculate the significant peaks of the depth histogram within each superpixel on the rough depth map. Suppose that a peak contains n_p pixels, the bins next to the peak contain n_l and n_r pixels respectively, and the superpixel contains N pixels. A peak is considered significant if the following conditions are satisfied:

$$\frac{n_p}{N} \geq \delta_1, min\frac{n_p}{n_l}, \frac{n_p}{n_r} \geq \delta_2, \tag{1}$$

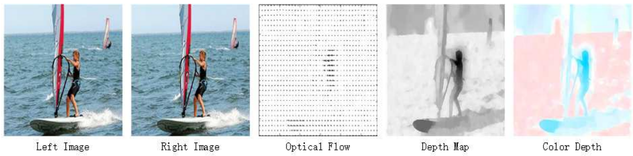


Figure 2. Visual process of color depth map calculation.

where δ_1 and δ_2 are the threshold values.

Pixels inside the superpixel are assigned with the average depth value of the nearest peak. This process can help smooth the rough depth map, remove tiny noisy areas, fill blank holes, and refine object boundary errors.

Finally, we calculate the color depth map via referencing the depth map (Fig. 2).

2.2. Coarse Saliency Map Generation

Human percept or recognize an object by the features difference between the salient object and its surroundings.

In this part, we use the color feature, depth feature which is obtained by the optical flow estimation and spatial feature to exploit the difference.

First, the left image is segmented into K regions via the K-means algorithm. The color feature saliency S_c is calculated via the following equation:

$$S_c(r_i) = \sum_{i=1, i \neq k}^{K} P_i W_d(r_k) D_c(r_k, r_i), \tag{2}$$

where K is the number of the background seeds, r_i and r_k represent the background region i and foreground region k respectively, $D_c(r_k, r_i)$ is the Euclidean distance between region i and region k in L*a*b color space, P_i represents the area ratio of region i compared with the whole image, $W_s(r_k)$ is the spatial weighted term of the region k, set as:

$$W_s(r_i) = e^{\frac{-D(r_i, r_m)}{\sigma^2}}, \tag{3}$$

where $D(r_i, r_m)$ is the Euclidean spatial distance between the region i and m, σ^2 is the parameter controlling the equation strength.

And the depth features saliency S_d is calculated as the same as the color features saliency calculation, denoted as:

$$S_d(r_k) = \sum_{i=1, i \neq k}^{K} P_i W_d(r_k) D_d(r_k, r_i), \tag{4}$$

where $S_d(r_k)$ is the depth saliency of I_d, $D_d(r_k, r_i)$ is the Euclidean distance between region k and region i in depth space.

Then, by visual physiological mechanism, we know that human use the fovea locate the salient objects, and the salient object in a picture always located at the center as well. Therefore the spatial features S_s is calculated as following:

$$S_s(r_k) = \frac{G(\|P_k - P_o\|)}{N_k} W_d, \tag{5}$$

where $G(\cdot)$ represents the Gaussian normalization, $\| \cdot \|$ is Euclidean distance, P_k is the position of the region k, P_o is the center position of this map, N_k is the number of pixels in region k, W_d is the depth weight, which is set as:

$$W_d = (max\{d\} - d_k)^{\mu}, \tag{6}$$

where $max\{d\}$ represents the maximum depth of the image, and d_k is the depth value of region k, μ is a fixed value for a depth map, set as:

$$\mu = \frac{1}{max\{d\} - min\{d\}}, \tag{7}$$

where $min\{d\}$ represents the minimum depth of the image.

Finally, the coarse saliency map is calculated as:

$$S_1(r_k) = S_c(r_k) S_s(r_k) + S_d(r_k) S_s(r_k), \tag{8}$$

2.3. Flow Saliency Map Generation

Human will always notice the objects in front of a scene. The pictures are taken by the same mechanism. Therefore, we get the stereo image pairs to represent the left view and right view of the human eyes.

In this part, we use the color depth map to replace the left image and use the coarse saliency map generation process to generate the flow saliency map. Then, we can get the flow saliency map which denotes as S_f.

2.4. Background Elimination Map Generation

Indicated by the visual physiological mechanism, human will always ignore the edge of a scene and use eyes fovea to focus on their interested objects. The pictures are taken by the same mechanism.

First, we use the efficient Simple Linear Iterative Clustering (SLIC) algorithm [8] to segment the image into smaller superpixels in order to capture the essential structural information of the left image. Based on the above mechanism that superpixels on the image boundary tend to have a higher probability of being the background, we assign a low saliency value to the boundary superpixels. For others, we assign a uniform value as their initial saliency values.

Then we use cellular automata to eliminate the background edge. We denote a superpixel generated by the SLIC algorithm as a cell. We assume that superpixels on the image boundaries are all connected to each other, because all of them serve as background seeds. We denote the neighborhood of the cell is 2-layer which is similar to the graph theory. We use the color feature to measure the similarity among cells. If a neighbor has more similar color with the cell, it will have a bigger impact on the cell at next moment. We define the influence factor between the cell i and the cell j as:

$$F_{ij} = \begin{cases} W_s(r_i) & ,j \in N_b(i) \\ 0 & ,j \notin N_b(i) \end{cases}, \tag{9}$$

where $N_b(i)$ is the neighbor set of i.

The next moment of each cells' status is controlled by its current status and its neighbors' status, so, we need to balance the influence of these two aspects. At the same time, if there are more similar characteristics among current cellular and its neighbors, the probability which they all belong to the foreground objects or background area is bigger, therefore, the neighbors should have an enormous effect on the current cells. So, we can constantly reinforce the effect of the foreground seed and assimilate the neighbor cells which have the similar color with current cell. In this way, we can calculate more salient objects. To measure the influence of a cell by its neighbors, we use confidence matrix C to represent the influencing process, which is denoted as:

$$C_i = a \times \frac{C_i - min(C_j)}{max(C_j) - min(C_j)} + b, \tag{10}$$

where i is the current cell, j its neighbors. The initial value of confidence matrix C_i equals to the reciprocal of the maximum of influence factor F_{ij}. The parameters of a and b control the range of confidence matrix.It can be seen from the definition of confidence matrix that the more similar between the current cell i and its is neighbors j, the smaller value the confidence matrix C_i will have at index (ij). Parameter b represents the prescribed influencing minimum. And the parameters of a and b should satisfy the following formula:

$$a + b < 1, a \geq 0, b \leq 0, \tag{11}$$

We consider that the case where the parts of salient objects will have the enormous difference in the color feature, so, the parameter a can not be set too large. In this paper, the parameters of a and b are set to 0.6 and 0.2 respectively. By setting the range, we can ensure that each cell is automatically updated to a more stable and accurate state.

Finally, we use the synchronous update rule shown in Eq.(12) to get the background elimination saliency map (BES). We denote it as S_b which is calculated by the following:

$$S_{t+1} = CS_t + (I - C)FS_t, \tag{12}$$

where C is the confidence matrix, I is the identity matrix, and F is impact factor matrix. S_t is the saliency value of the current state which is calculated by Eq.(2).

2.5. Final Saliency Map Generation

After obtaining the coarse saliency map S_1, the flow saliency map S_f and the background elimination saliency map S_b. We can get the final saliency map S of the left image via the fowling equation:

$$S = S_b S_f S_1. \tag{13}$$

The main steps of the proposed salient object detection algorithm are summarized in Algorithm 1.

Algorithm 1 The proposed saliency detection algorithm

```
Input: stereo image pairs $I$;
Output: final saliency map $S$;
generate the depth maps $I_d$ use the Optical Flow;
for each region $k = 1, K$ do:
compute color saliency values $S_c(r_k)$ via Eq.(2) and depth saliency values $S_d(r_k)$ via Eq.(4);
calculate the center-bias and depth weights $S_s(r_k)$ via Eq.(5);
get the coarse saliency map $S_1$ via Eq.(8);
end for
obtain flow saliency map $S_f$ and background elimination map $S_b$;
figure out the final saliency map $S$ via Eq.(13) ;
return  final saliency map $S$;
```

3. EXPERIMENTS

We evaluate our method on our SSD100 dataset, NJU2000 dataset [6], RGBD1000 [10] dataset and RGBD135 [3] dataset, and compare the results with the state-of-the-arts. More experimental analyses on the effectiveness of our method are given as follows.

3.1. Datasets

Our SSD100 dataset is built on three stereo movies. The movies contain both the indoors and outdoors scenes. We pick up one stereo image pair at each hundred frames. It totally has tens of thousands of stereo image pairs. We make the image acquisition and image annotation independent to each other, we can avoid dataset design bias, namely a specific type of bias that is caused by experimenters unnatural selection of dataset images. The chosen stereo image pairs are based on one principle: choose the one which the computer detect the salient objects within the complex scenes where even the human cannot tell the salient objects at once. After picking up the stereo image pairs, we divide the image pairs into left images and right images both in 960 × 1080 size. When we build the ground truth of salient objects, we adhere to the following rules: 1) we mark the salient objects, taking the advice of most people; 2) disconnected

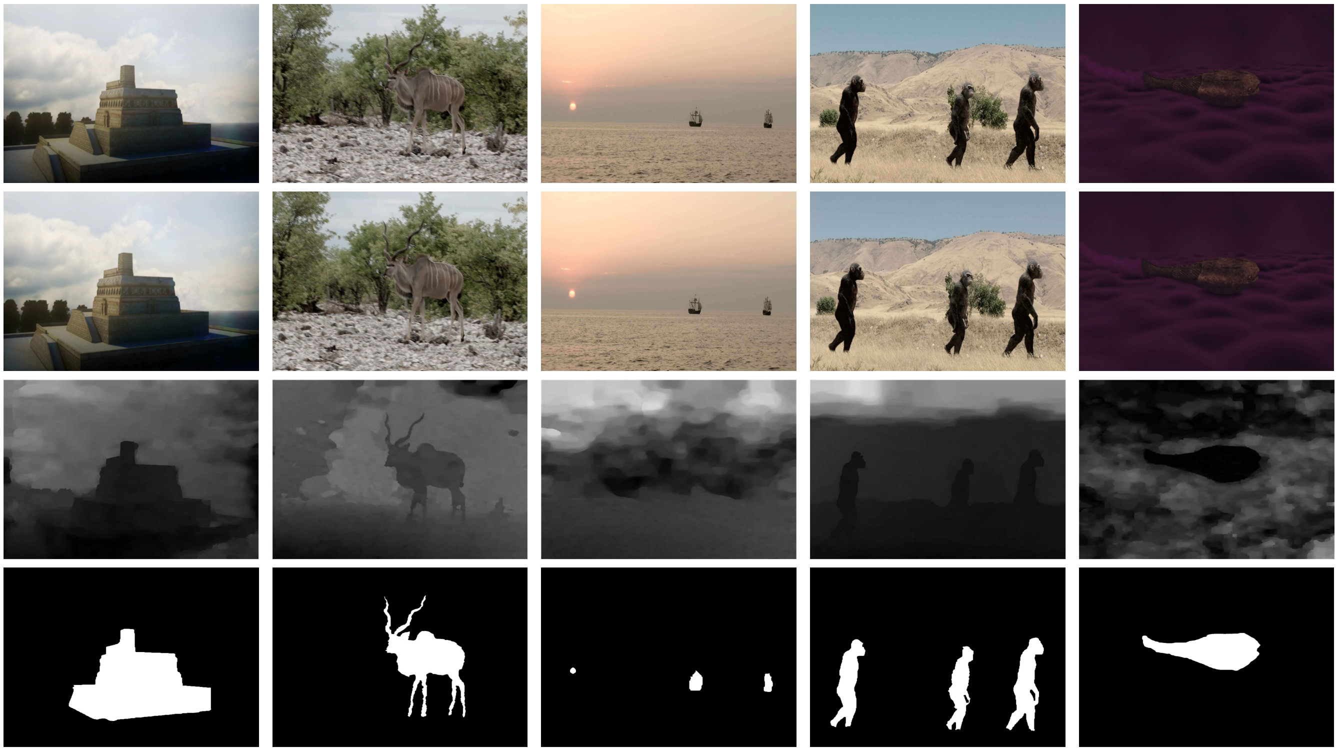


Figure 3. Examples of the proposed SSD100 dataset. From top to bottom: left-view images, right-view images, depth maps, ground truth.

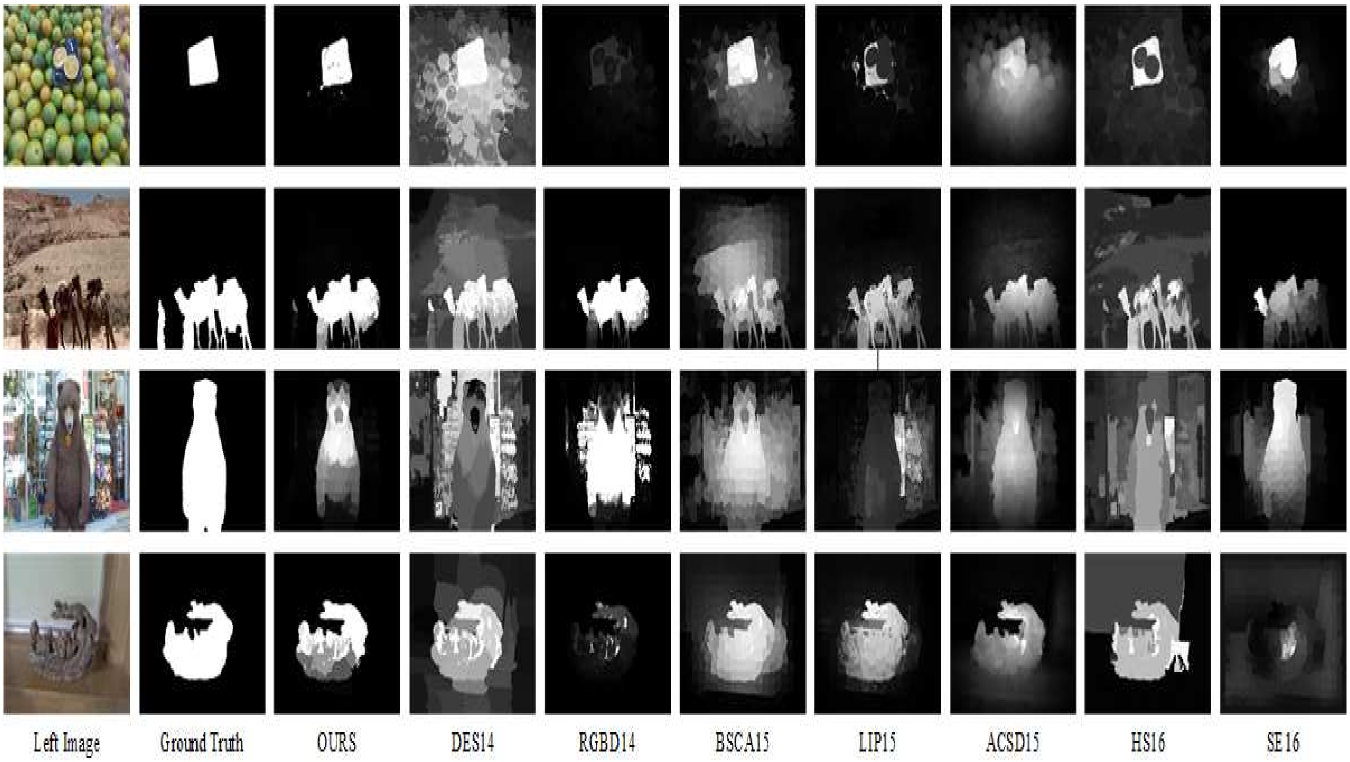
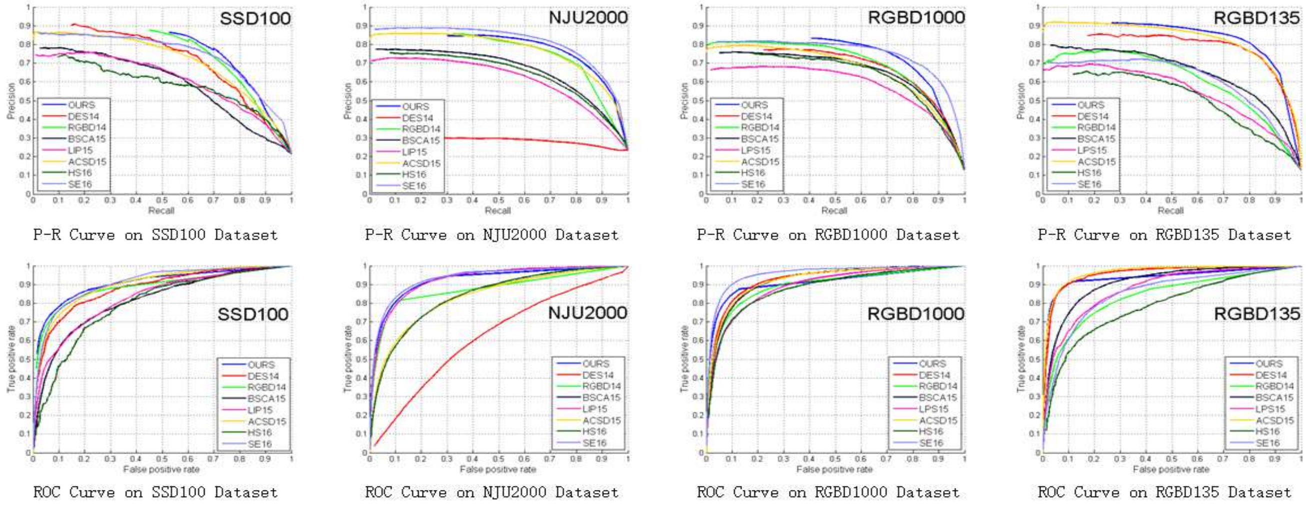


Figure 4. Visual comparison of saliency maps on four datasets.



P-R Curve on SSD100 Dataset

P-R Curve on NJU2000 Dataset

P-R Curve on RGBD1000 Dataset

P-R Curve on RGBD135 Dataset

ROC Curve on SSD100 Dataset

ROC Curve on NJU2000 Dataset

ROC Curve on RGBD1000 Dataset

ROC Curve on RGBD135 Dataset

Figure 5. The PR curve and ROC curve evaluation results on four datasets.

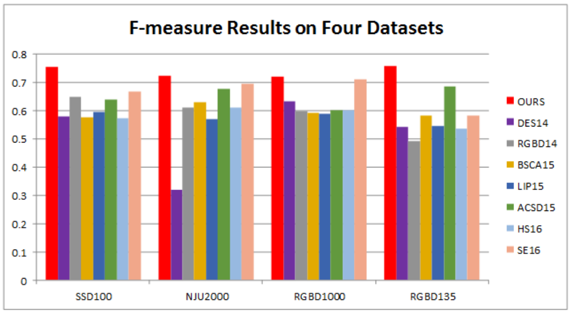


Figure 6. The F-measure results on four datasets. The higher values, the better performances.

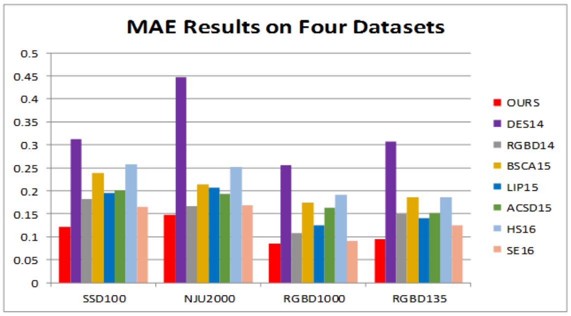


Figure 7. The MAE results on four datasets. The lower values, the better performances.

regions of the same object are labeled separately; 3) we use solid regions to approximate hollow objects, such as bike wheels. Besides, we will expand this dataset continually in future. Fig. 3 shows some examples of the SSD100 dataset.

The NJU2000 dataset contains 2000 images collected from more than 20000 stereo images. The stereo images are collected from Internet, 3D movies and photos taken by the Fuji W3 stereo camera. It has two revolutions of 492×600 and 587×440.

The RGBD1000 dataset includes 1000 RGB-D based images captured by Kinect, along with manually labeled groundtruth masks.

The RGBD135 dataset has 135 indoor images taken by Kinect with the resolution 640×480.

3.2. Evaluation indicators

Experimental evaluations are based on standard measurements including precision-recall curve, ROC curve, MAE (Mean Absolute Error), F-measure. The precision is defined as:

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

where $p_i \mid d(p_i) \geq d_t$ 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 recall define as:

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

The precision-recall curve is plotted by connecting the P-R scores for all thresholds.

The ROC curve is as the same as the P-R curve with different parameters.

The MAE is formulated as:

$$MAE = \frac{\sum_{i=1}^{N} \|GT_i - S_i\|}{N}. \tag{16}$$

where N is the number of the testing images, GT_i is the area of the ground truth of image i, S_i is the area of detection result of image i.

The F-measure is formulated as:

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{17}$$

3.3. Comparison

To illustrate the effectiveness of our algorithm, we compare our proposed methods with DES14 [3], RGBD14 [10], BSCA15 [11], LPS15 [7], ACSD15 [6], HS16 [12] and SE16 [4]. We use the codes provided by the authors to reproduce their experiments. For all the compared methods, we use the default settings suggested by the authors. And for the Eq. (3), we take $\sigma^2 = 0.4$ which has the best contribution to the results.

Fig.5 shows PR curve and ROC curve comparison results on four datasets.

Fig.6 and Fig.7 show F-measure comparison results and MAE comparison results on four datasets, respectively.

From the comparison results, we can see that our saliency detection results have a better robustness results on four datasets. Besides, we also show the visual comparisons as shown in Fig. 4, which clearly demonstrate the advantages of the proposed method. We can see that our method can detect salient objects more precisely. In contrast, the compared methods may fail in some situations.

4. Conclusion

In this paper, we proposed a biologically motivated saliency detection framework for automatically producing saliency map from stereo image pairs. We imitate the processing of the human visual perception to generate a framework which composes of background elimination part, color flow detection part and features difference saliency part. Besides, we build a new high-quality dataset with complex scenes based on dataset design principles. The experimental results on four datasets demonstrate our algorithm improves the quality of saliency detection and is more robustness. To encourage future works, we make the SSD100 dataset and related materials open. All of these can be found on our project website[1].

5. Acknowledge

We would like to thank anonymous reviewers for their helpful comments on the paper. This work was supported by the grant of National Natural Science Foundation of China (No.U1611461), the grant of Science and Technology Planning Project of Guangdong Province, China (No.2014B090910001) and the grant of Shenzhen Peacock Plan (No.20130408-183003656).

[1] https://chunbiaozhu.github.io/prof.cbzhu/

References

[1] M. Bjrkman and D. Kragic. Active 3d segmentation through fixation of previously unseen objects. In *British Machine Vision Conference, BMVC 2010, Aberystwyth, UK, August 31 - September 3, 2010. Proceedings*, pages 1–11, 2011.

[2] M. M. Cheng, N. J. Mitra, X. Huang, and S. M. Hu. Salientshape: group saliency in image collections. *Visual Computer*, 30(4):443–453, 2014.

[3] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao. Depth enhanced saliency detection method. 55(1):23–27, 2014.

[4] J. Guo, T. Ren, and J. Bei. Salient object detection for rgb-d image via saliency evolution. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2016.

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

[6] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu. Depth-aware salient object detection using anisotropic center-surround difference. *Signal Processing Image Communication*, 38(C):115–126, 2015.

[7] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price. Inner and inter label propagation: salient object detection in the wild. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 24(10):3176–3186, 2015.

[8] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Transactions on Medical Imaging*, 31(2):474–86, 2012.

[9] N. Ouerhani and H. Hgli. Computing visual attention from scene depth. In *International Conference on Pattern Recognition*, page 1375, 2000.

[10] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. *RGBD Salient Object Detection: A Benchmark and Algorithms*. Springer International Publishing, 2014.

[11] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *Computer Vision and Pattern Recognition*, pages 110–119, 2015.

[12] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016.

[13] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition*, pages 2432–2439, 2010.

[14] C. Zhu, G. Li, X. Guo, W. Wang, and R. Wang. *A Multilayer Backpropagation Saliency Detection Algorithm Based on Depth Mining*, pages 14–23. Springer International Publishing, Cham, 2017.

[15] C. Zhu, G. Li, N. Li, X. Guo, W. Wang, and R. Wang. An innovative saliency detection framework with an example of image montage. In *ACM MultiMedia Workshop 2017 Submission Proposal for South African Academic Participation Proceedings*, 2017.

[16] C. Zhu, G. Li, W. Wang, and R. Wang. Salient object detection with complex scene based on cognitive neuroscience. In *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*, pages 33–37. IEEE, 2017.